Protocol for Researchers

1 Version

Draft, last update by Eric Maris on October 14, 2016

2 Scope

This protocol is in effect as of 04-11-2016 and replaces all previous versions of the DI-RDM protocol for researchers

3 Audience

The audience for this protocol consists of all people conducting research at the Donders Institute (DI). This includes research assistants, internship students, PhD students, guest/visiting researchers, postdocs, senior researchers, and principal investigators. These people will be collectively denoted as *researchers* in the subsequent text. This group is to be distinguished from the *research administrators*, who have their own protocol.

4 Context

The context of this protocol is the *Research Data Management* (RDM) at the Donders Institute (DI), which involves interaction with a digital data repository. This protocol describes how researchers must interact with this repository. Future changes to the organization and user interface of the repository may have consequences for this protocol, and these changes will be integrated in future versions of this protocol. The researcher should therefore always ensure to use the latest version of this protocol. This protocol serves three purposes:

1. *Data preservation* for internal reuse
2. *Research documentation* for increasing reproducibility

3. *Data sharing* with the external scientific community

The data repository allows a researcher to archive digital data into collections. The repository contains separate collections for preservation, for research documentation and for sharing. Repository users can have different predefined roles in a collection. These roles differ in their rights with respect to the collection.

1. A *manager* can add and remove users to the collection, and can change the role of other users. A manager also has the rights of a contributor.
2. A *contributor* can add, modify and delete files. A contributor also has the rights of a viewer.
3. A *viewer* can view the content of files.

When a collection is initiated, a research administrator assigns a manager to the collection (see the *DI RDM Protocol for Research Administrators*). Typically, the collection manager is also the supervisor of the research project to which this collection belongs. A collection manager can add other users to the collection in the role of manager, and these then obtain the same rights.

5 Data, Metadata, and Attributes

There exist several definitions of data. For the purpose of this protocol, we focus on *research data*, which we define as follows:

All information that is (1) generated as a part of the research process and (2) on which a scientific report is/will be based.

This definition does not only include empirical data, but also simulated data, computer scripts for analysis and simulations, stimuli presented in experiments and the computer scripts for presenting them, etc. A good way to determine what is the research data on which a study is based, is asking oneself what information has contributed to the results on which you report in your publication. And a good way to delineate the empirical from the other research data, is by asking oneself whether one has used some device for obtaining these data: empirical data are always collected using a device (a button box, a keyboard, an MRI scanner, an EEG or MEG system, a video camera, a touch screen, a microscope, …). Not all studies and publications depend on research data, as defined above. This holds for theoretical, perspective and opinion papers, as well as for reviews. This protocol is only relevant for those studies and publications that do depend on research data.

Empirical research data can be acquired in digital or non-digital form. This distinction is relevant because only digital data are to be archived on the data repository. However, this protocol addresses both digital and non-digital data. Specifically, this protocol also describes how non-digital data have to be preserved and documented. Common non-digital data are psychometric and performance data on paper or film (e.g., questionnaires, psychological tests, handwritten text, photos), and samples or biochemical assessment of biological tissue (e.g., brain tissue, blood, saliva).

Metadata is data about data. Most familiar are the metadata at the level of individual files: the filename, its type (.docx, .pdf, .dcm, .mat, .xlsx, …), its creation date, the date it was last modified, etc. In this protocol, we will mainly consider metadata at the collection level, because this allows for metadata that are very useful for research data management. Some examples of such collection-level metadata are the persons that have contributed to the study, the budget from which it is financed, the date and time it was created and closed, the type of data that was collected (behavior, fMRI, EEG, MEG, genomics, transcriptomics, video, …), the species (human, non-human primate, rat, mouse, …), and the topic of the study.

The function of metadata is to describe the data in a way that increases their value for potential users ("Are these data interesting for me?","With which software can I process them?", ...). For that reason, the concept of metadata is closely linked to *data publication*. Importantly, because published data should not change (just like a published paper also should not change), the current view is that also metadata should not change. However, there are a few situations in which it is very useful to have attributes that could change while the data themselves do not. For example, over time, a given published data set may be used in an increasing number of publications. All these publications are highly relevant attributes of that data set, even though technically they cannot be called "metadata".

In this protocol, a distinction is made between *metadata* and *attributes*: metadata are a special class of attributes that have the distinguishing feature that they are communicated to potential users of published data. Because the current view is that metadata should not change if the published data don't change, the metadata of a closed collection (see further down, for a description of "collection closure") cannot be modified. However, this collection can have other attributes (technically, not metadata) that *can* be modified after closure.

6 Accessing the Repository

6.1 Authentication

When visiting the website of the data repository, one can see the attributes of all closed *Data Sharing Collections* (DSCs; see further). However, one cannot see any files, nor the attributes of the other collections. To access the files in the repository, and to see the metadata of the other (non-DSC) collections, one must first log in. This is also called *authentication*. There are three ways to authenticate:

1. Using a local identity provider (RU or RadboudUMC) and logging in with your RU or RadboudUMC credentials. This applies to RU and RadboudUMC employees (U-, resp., Z-number), for RU students (s-number), and for affiliated non-employees (E-number). It is possible that a person has multiple credentials (e.g., a U- and a Z-number). If a physical person logs in using multiple credentials, the system will treat that physical person as multiple users, mapped one-to-one on the different credentials.
2. Using a trusted federated authentication infrastructure (e.g., the Dutch SURFconext and the international eduGAIN) and logging in with the credentials provided by the researcher's employer. This authentication option can be used by employees of the organizations that participate in the federated authentication infrastructure – so not necessarily RU or RadboudUMC.
3. Using a non-trusted identity provider service such as Linkedin, Google, Facebook, or Twitter.

One can only authenticate after *signing up*. By signing up, an internal user profile is generated, via which one can access files on the repository (after being authorized for access; see further). One needs this internal user profile to up- and download files using a client for file transfer (see further).

6.2 Authorizations and authorization levels

After a successful authentication, a repository user still cannot read or write a collection's files. To do this, a user must first be added to that collection in a particular role. The research administrator and collection manager can add a user to a collection. The collections to which a user is added in a particular role determines what that user can do in the repository. This is also called the user's *authorizations*. Not every user can be added to an arbitrary collection in an arbitrary role. In fact, a user's possible authorizations are determined by the way he/she has authenticated. To describe a user's possible authorizations, a distinction is made between four different user types, ordered according to their authorization level. These four user types form a hierarchy in which every next user type has the authorizations of the previous type plus some more:

1. An *anonymous user* (also called, non-registered user or visitor) visits the website of the repository without authenticating him/herself. An anonymous user can view the attributes of all closed DSCs. However, he can be added to none of the collections, and therefore cannot read a single file.
2. A *non-trusted registered user* authenticates him/herself against a non-trusted identity provider (e.g., Google, Facebook). This user type can be added as a viewer to a closed DSC.
3. A *trusted registered user* authenticates him/herself against a trusted authentication service (e.g., SURFconext, EDUgain). This user type can be added as a contributor or viewer to all collection types (including *Data Acquisition* and *Research Documentation Collections*, DACs and SICs; see further). A contributor cannot only view but also modify a collection's files.
4. A *DCX employee* authenticates him/herself against the RU or RUMCN identity provider *and* is registered by a center's research administrator as an employee of that center. In the Donders Institute, we distinguish between four centers: DCC, DCCN, DCN*S (DCN Science Faculty), and DCN_M (DCN Medical Faculty), jointly denoted as DCX. This user type can be added as a _manager* to all collections of his center.

6.3 One physical person, multiple user profiles

In principle, the same physical person can be represented with multiple user profiles in the repository. This may happen if that physical person has a digital identity (account) on multiple identity providers and/or multiple digital identities on a single identity provider. This would for example happen if that person is employed by multiple institutions (e.g. Radboud University and the Max Planck Institute). If that person authenticates him/herself using the different digital identities, then he/she will also be represented with multiple user profiles in the repository, each of which is likely to have different authorizations.

6.4 Authorizations linked to employers

Some identity providers represent an organization that also acts as an employer (a university, a governmental organization, ...). The membership of these organizations changes with employment. Because employees in academia often change organization, they also change identity provider. Importantly, when a user signs up using the credentials (ID plus password) provided by an organization, an internal user profile is created that is specific for this organization. Therefore, when a user becomes authorized (in a particular role) for some collection, that authorization is linked to the organization where the user is employed, namely via that organization's identity provider.

When a user looses access to this identity provider, e.g. by switching jobs, the authorizations granted to the user profile linked to that identity provider are no longer accessible. Consequently, he/she loses access to the collections for which he/she used to be authorized via this identity provider. To regain access to these collections, the user must do the following:

1. Find another identity provider (e.g., another employer with an identity provider) with which he/she can sign up and create an associated user profile in the repository.
2. For the collections to which he/she wants to have access, ask the managers to be added with the appropriate authorizations linked to his/her new user profile.

When a user, after a period of unemployment, returns to his/her old organization and regains access to that organization's identity provider, that user will also regain the authorizations that he had prior to leaving the organization.

7 Up- and Downloading Files

A web interface does not allow for a convenient up- and download of large files and a large number of files. For that reason, the up- and download of files to and from the repository has to be performed by means of a specialized client, such as a *WebDAV* client (called after the protocol that is used to transfer files over the internet). A very good WebDAV client is Cyberduck, which can be downloaded for free.

Up- and downloading files is only possible after authenticating oneself at the website of the repository. Via the website, one can then obtain a *data access password* that can be used in combination with one's *data access username* to up- and download files. This data access username is *not* the username with which one authenticates oneself at the website of the repository (for RU and RUMCN employees, their U- or Z-number). The data access username and password are used in combination with the WebDAV client.

It is also possible to up- and download files using a command-line interface. However, this file transfer is only possible to and from a few dedicated computers, which are typically located in the labs. This file transfer is intended for large volumes of data, and the dedicated computers will only be operated by support staff and a limited number of experienced users.

8. Preservation – Data Acquisition Collection

8.1 Objectives

1. Preserving the (digital) research data in their original form.
2. Documenting conformity with the relevant laws and regulations pertaining to data acquisition.
3. Annotating the data to increase its scientific usefulness.

This protocol involves both required (8.2) and recommended (8.3) operating procedures.

8.2 Required

8.2.1 Initiation

A *Data Acquisition Collection* (DAC) is initiated by the research administrator upon request by a researcher, typically following formal approval by the center director, or on the basis of criteria put forward by the center director (e.g., the Project Proposal Meeting at the DCCN; see Center Specific). The protocol for initiating a DAC is described in the *DI RDM Protocol for Research Administrators*.

8.2.2 General

By *data*, we mean *all information that is (1) generated as a part of the research process and (2) on which a scientific report is/will be based*. This does not only include empirical data, but also simulated data, computer scripts for analysis and simulations, stimuli presented in experiments and the computer scripts for presenting them, etc.

Data must be archived in their original form. Here, original means the following: without any manipulations that limit future analyses of these data.

The DAC must be annotated in two ways: (1) by providing collection attributes (see further), and (2) by adding one or more human-readable documents in which the data are annotated.

A DAC has one or more managers, and these can add other users to the DAC, thereby giving them access. During data collection, only members of the research team that performs the data acquisition can be added to the DAC. After DAC closure (see 8.2.9), other members of the DI may be added to the DAC, allowing them to re-use the data. Users that are not a member of DI may not be added to a closed DAC. These non-DI users can only get access to the data via a closed *Data Sharing Collection* (DSC, see further).

8.2.3 DAC annotation

The DAC annotation is a set of one or more documents that is uploaded to the collection. It contains information that is necessary to properly interpret and (re-)analyze the data. This information involves three parts: (1) a description of the type of data that was acquired, (2) the organization of the files and the folders and its relation to the design of the study/experiment, and (3) experimental information that is written into the datasets by the data acquisition system.

Describing the type of data can usually be short, but might include details on the equipment used and the location where the data was recorded.

The description of the organization of the data is essential for a study in which participants participate in multiple recording sessions, as well as for an experiment in which different types of data are acquired. The DAC annotation must include a description of the different sessions and data types, and how these are mapped onto the experimental conditions.

The experimental information written by the data acquisition system refers for example to stimulus events that are relevant for the analysis of the data, or to behaviour (e.g., button presses, eye movements). This information is often present in the form of experimental log files or as trigger or annotation channels. As a part of the DAC annotation, researchers should include a description of the codes that are used for these stimulus and behavioral events (i.e., a table associating the events to their meaning).

The DAC annotation should be provided in a format that is easily accessible to present and future colleagues, for example, .docx, .txt, .pdf, .xlsx. The annotation should be written such that someone with domain specific skills can reuse the data. It is required to specify details that vary between studies, such as the mapping of trigger codes on experimental parameters. It is not required to explain details that can be considered commonly shared knowledge.

8.2.4 Conformity with laws and regulations

In a DAC, information will be collected that demonstrates conformity with valid local, national and international laws and regulations, as evaluated by an ethical review board (see further). As outlined below, the way to demonstrate this conformity is different for human and animal data. This protocol does not detail the relevant laws and regulations, nor does it include measures to enforce conformity with them.

8.2.4.1 Human data

1. The researcher must indicate whether approval was obtained from an accredited *Medical Ethical Reviewing Committee* (MERC) or non-accredited local ethical reviewing committee and, if this is the case, specify the name of this committee (e.g., CMO Regio Arnhem-Nijmegen, Ethics Committee Faculty of Social Sciences) and the registered identifier of the approved application. If approval was obtained from multiple ethical review boards (e.g., because the DAC contains data of multiple studies, which were evaluated separately), then all these ethical review boards must be indicated. Guidelines for deciding about the appropriate ethical review board can be found in a separate document in the Donders RDM Information Package.
2. The signed informed consent forms must allow for the de-identified data to be shared via a data sharing collection (DSC). If the researcher plans to share the data of this DAC together with data of the same participants in other DACs, then the signed informed consent forms must mention this linked sharing.
3. The researcher must upload a copy of the *Participant Information Brochures* (PIBs) that he/she has provided to the participants.
   There are three types of information brochures: general, method-specific (e.g., EEG, MEG, fMRI, audio/video), and study-specific. For the former two, template information brochures have been written specifically for use in combination with this protocol. Prior to being used, these template brochures, or modified versions thereof, must be submitted to the appropriate ethical review board.

8.2.4.2 Animal data

The researcher must specify whether approval was obtained from an accredited Animal Care and Use Committee. If this is the case, then the researcher must specify the name of this committee (e.g., Dier-experimentencommissie Radboud Universiteit Nijmegen) and the registered identifier of the approved application.

8.2.5 Digital data

Whenever possible, digital datasets must be added to the DAC immediately after acquisition. The digital datasets must contain all the scientifically relevant data that were obtained from the participants or samples under investigation. These datasets must also include files that are often considered auxiliary, such as log files containing behavioral or questionnaire data. Technical data that do not provide information about the correctness of the results, such as those related to some calibrations of the measurement device, should not be part of the DAC.

In case of human data, the name of the files/folders in which a participant's data are stored may not contain information that allows this participant to be identified in a direct way (e.g., by including the participant's name as a part of the file/folder name).

Instead, this file or folder name should contain a code that uniquely identifies the participant in the project.

The key that relates the participant identification code and the participant's personal information (name, address, telephone number,…) may not be kept in the DAC. The DI centers determine where and on which medium this so-called *pseudonimization key* is kept.

Digital datasets can be uploaded in two ways: (1) manually by the researcher (as described in 7), and (2) by an automatic upload procedure that is controlled by the lab manager or ICT group. The DI centers decide which upload process is to be followed for which lab and for which study type.

8.2.6 Non-digital data

Non-digital data (e.g., biological tissue samples, questionnaires on paper) cannot be stored in the repository. Their location has to be specified in the appropriate field, either by selection from a menu or by entering it as free text. The menu may contain labels (e.g., DCN Biological Tissue Bank) that do not specify a particular physical location but instead refer to a location that is commonly known in the respective center.

Typically, within a room, the non-digital data of multiple DACs are stored at a location that can be further specified by referring to a cabinet (e.g., nr. 2), a shelf (e.g., shelf A), a freezer (e.g., freezer B, rack 2), etc. These location indicators must be provided in one or more files whose format must be easily accessible, for example, .docx, .txt, .pdf, .xlsx. In case the non-digital data are relocated with the room, these files must be updated.

Typically, parameters of non-digital data will be converted to digital format. All digital data that are obtained in this way must be added to the DAC. Storing non-digital data in the archive after conversion to digital format does not automatically imply that the original non-digital data can be discarded. This is because the conversion process can be selective, in the sense that not all parameters are extracted that are possibly of scientific value. For some non-digital data (e.g., questionnaires, psychological tests, handwritten text) the conversion to digital format involves little or no loss of scientific value. For other non-digital data (e.g., blood and other tissue samples), by performing an analysis (e.g., biochemical), only a few parameters (e.g., hormone levels) are converted to digital format.

In cases where the center director decides that the non-digital data (e.g., tissue samples) do not have to be archived, it suffices that only the converted digital data is archived.

8.2.7 Data reduction

For some data, the scientific usefulness is not substantially affected if the data are first reduced prior to archiving. Different scientific disciplines use different types of data reduction. For instance, spatial and/or temporal downsampling is common for high-density broadband biological signals, and compression (e.g., jpeg) is common for images and video. For this type of data, the reduced data may be stored on the repository. In this case, the data reduction algorithm must be described as a part of the DAC annotation.

8.2.8 Pilot data

Pilot studies differ from regular studies in that it is not their intention to collect data systematically, or to provide evidence for generalizable conclusions (i.e., generalizable beyond the sample investigated). Pilot studies may differ from the main part of the study in several ways. For example, the pilot data could be collected prior to obtaining formal approval for the data acquisition, or they were collected using a preliminary protocol. For all these different pilot types, there is a single protocol, which involves the following:

1. In case the center regulations and the ethical approval allow for pilot data being added to the DAC, then the pilot data must be added as soon as possible after initiation of the DAC.
2. It must be specified in the DAC annotation (see further down) which part of the data are pilot data.

8.2.9 DAC closure

A collection may only be closed when its content is in accordance with this protocol. Only a collection manager can close a collection, and he/she is thus responsible for the decision whether or not the collection is in accordance with this protocol.

When a collection is closed, a permanent read-only copy of this collection is created and a persistent identifier is assigned to it. A collection manager may ask the system administrator to re-open a closed collection, but may only do so with a good reason, for example if the closed collection contains incorrect and/or insufficient information. Carefulness is required here, because closing a collection twice also requires twice the amount of storage (and thus costs twice as much). In fact, when a re-opened collection is modified and thereafter closed again, then a second permanent read-only copy of the collection is created, again with a unique identifier. All the read-only copies that are created at collection closure are thus snapshots of the collection at different time points.

A collection is closed in *two* steps:

1. The status of the collection is changed from *open* to *tobeclosed*. In the *tobeclosed* status, the collection has become read-only. In this state, the collection can be inspected by all contributors, allowing them to determine whether it is complete (as specified by this protocol). If this turns out not to be the case, then the manager can change the collection's status back to *open*.
2. The status of the collection is changed from *tobeclosed* to *closed*. Hereafter, a permanently read-only copy of the collection is created and a persistent identifier is assigned to it. If the collection would be re-opened, this read-only copy would not be modified. The actual creation of this permanently read-only copy is determined by the repository, and this may involve some delay relative to the change of the collection status to *closed*.

There can be situations in which it is useful to make multiple collection snapshots. For instance, this may be the case when the DAC belongs to an ongoing study, and one wants to write a paper using the data collected until present. The identifier that is assigned to this collection then serves as a reference in other collections that make use of this DAC snapshot. This workflow is a form of versioning.

8.2.10 Centre specific - DCC

The DCC does not have a dedicated location for non-digital data. For psychological tests and questionnaires that were administered on paper, the DCC protocol is that the paper booklets are scanned and the resulting files are uploaded to the RDM repository.

8.3 Recommended

8.3.1 General

8.3.2 Neuroscience-specific attributes

Neuroscience-specific attributes are useful to find DACs via a search query. Neuroscience terms can be selected from two controlled vocabularies: the topic list of the Society for Neuroscience (SFN), and the Medical Subject Headings (MESH; used by PubMed for indexing articles). When selecting terms from these controlled vocabularies, the following domains may guide the collection managers and contributors: technique, topic, species, disorder, and brain area.

8.3.3 Associating DACs

Some DACs are associated with existing DACs, for example, because they all belong to the same longitudinal study. Associations between DACs are documented by specifying the identifiers of the associated DACs.

8.3.4 DAC annotation

It is recommended to include stimulus presentation scripts, including software version number in the DAC.

8.3.5 Checking contributions

As a part of his/her final responsibility for a collection, it is recommended that a collection manager checks the contributions of all persons that have write access to this collection. For that purpose, a collection manager can make use of a monitoring tool (available to all users) that produces a log of all the changes in the collection, specifying who uploaded/modified which file at what time (source verification).

9 Research Documentation Collection

9.1 Objectives

1. Document the process via which data are converted into published results.
2. Provide a digital platform for collaborators to contribute to and review good scientific practices.

In a *Research Documentation Collection* (RDC), information will be collected that may be relevant for evaluating good scientific practices. However, this protocol does not specify any scientific integrity rules, nor does it include measures to enforce conformity with such rules.
This protocol involves both required (9.2) and recommended (9.3) operating procedures.

9.2 Required

9.2.1 Initiation

A RDC is initiated by the research administrator upon request by a researcher on the basis of criteria put forward by the center director. The protocol for initiating a RDC is described in the *DI RDM Protocol for Research Administrators*.

9.2.2 General

A RDC pertains to the process that has led to a scientific publication. It contains the documents that may be relevant for evaluating good scientific practices. The core content of a RDC are files that document the scientific process in which data are converted into results (statistical tests, summary measures, figures, tables, etc.). For simplicity, in the following, this part of the scientific process will be denoted as the *scientific process*, without any qualification as to the specific result it contributes to the publication.

An RDC must be initiated before the proofs of the accepted journal article are sent back to the journal's editorial office. An RDC may also be initiated in a very early stage of the scientific process, for instance, at the start of the data analysis. This has the advantage that the RDC can be used to share all kinds of documents with the future co-authors (analysis scripts, figures of preliminary results, the different versions of the manuscript, ...).

The RDC is *not* to be used for storing all the processed data that are generated as a part of the data analysis. For that purpose, typically, one should use the storage that is part of a computing environment (either a desktop or a server) to which most researchers have access. Processed data that can be added to the RDC should have the character of results (temporary or final).

9.2.3 Attributes

The researcher must add the following information:

1. Title of the manuscript
2. List of all co-authors

Following publication, the researcher is required to add the DOI (a persistent identifier) of the published manuscript. Besides the DOI, the researcher can also provide the Pubmed ID (PMID) or arXiv identifier (other persistent identifiers).

9.2.4 Contributors to the publication

All coauthors to a publication share responsibility for good scientific practices and therefore must be contributor to the RDC. Consequently, all coauthors must be registered users in the repository. It is the responsibility of the RDC manager to invite co-authors external to the DI to create a user profile. Following registration of the user profile, the RDC manager must add the co-author as contributor, thereby giving him/her access to the files in the RDC.

9.2.5 Documenting the origin of the data

When the published results depend on data, the RDC must refer to them. There are two ways to document the data on which the published results depend:

1. If the data is represented in the repository as one or more DACs, the RDC must be associated with the corresponding DACs by specifying the DAC identification numbers. Only closed DACs can be associated with the RDC. A single closed DAC may be associated to multiple RDCs.
2. If the data (as defined in *5 Data, Metadata, and Attributes*) is not represented in the repository as one or more closed DACs, the researcher must add the data to the RDC or document the source of the data by means of a persistent identifier. This situation applies, for instance, when the DAC is not yet closed, or when the data were collected at another institute. It also applies when the published results only depend on computer scripts (as in modeling work), in which case these scripts must be added to the RDC.

If the data is represented in the repository, then the collaborators on the RDC must obtain the right to access the relevant data in the DACs. If collaborators on the RDC do not have this right, there are two ways for these collaborators to get access:

1. The DAC manager adds the RDC collaborators to the collection in the role of viewer.
2. In agreement with the DAC managers, the RDC manager copies the relevant parts of the DACs to the RDC. The selection of the data from the DACs must be documented (in free text format) in the RDC.

9.2.6 Reproducing the results of the publication

By the time the scientific paper is accepted for publication, the RDC must contain all the information that a knowledgeable colleague needs to reproduce the results in this publication. This information could be used by an independent audit committee that investigates how the results of a publication were obtained. Very related information must be provided as a part of a Data Sharing Collection (DSC; see further), which is intended for external researchers that re-use the data.

Publications may differ substantially in the amount and detail of the information that is required for reproducing the results. Collaborators on a collection are expected to demonstrate scholarship in providing the required information.

9.2.7 Associating to the shared data

For all publications of which the results depend on data, the data must be shared. These data can be shared as a Data Sharing Collection (DSC) in this DI repository, or as a collection in a DI-external repository. In both cases, after collection closure, the shared

data can be identified by a persistent identifier. The RDC must be associated to the shared data by adding its persistent identifier as an attribute.

9.2.8 RDC closure

See 8.2.10.

9.2.9 Data/study type specific

9.3 Recommended

9.3.1 Analysis scripts

Often, the process via which data are converted into published results is partially or fully specified by analysis scripts that can be executed by software packages such as MATLAB, R, Python, SPSS, Bash+FSL, etc. The recommended way of documenting the scientific process is by providing these analysis scripts.

9.3.2 The editorial and peer-review process

It is recommended that the complete scientific publication process is documented in the RDC. The following files are part of this process: the files that are initially uploaded to a journal's manuscript submission system, the reviews, the reply to the reviewers, the proofs, etc.

9.3.3 Checking contributions

As part of his/her final responsibility for a collection, it is recommended that a collection manager verifies the contributions of all persons that have write access to this collection. For that purpose, a collection manager can make use of a reporting tool that produces a log of all the changes in the collection, specifying file uploads and modifications by the different collection contributors.

9.3.4 Registered reports

A registered report is a publication for which the experimental methods are pre-registered and reviewed before data are collected. The objective of registered reports is to neutralize a variety of inappropriate research practices, such as selective reporting of results, undisclosed analytic flexibility, and publication bias.

If the study was conducted as a part of a registered report, then specify (1) the authority (e.g., a journal, a website such as https://osf.io, https://clinicaltrials.gov, or https://aspredicted.org) where the methods and proposed analyses were pre-

registered and reviewed prior to the research being conducted, and (2) the identifier of this pre-registration.

10 Data Sharing Collection

10.1 Objectives

1. To allow external researchers to extend scientific findings by reanalyzing data with new methods, and/or by addressing new research questions using these data.
2. To allow external researchers to reproduce scientific findings.

This protocol involves both required (10.2) and recommended (10.3) operating procedures.

10.2 Required

10.2.1 Initiation

A *Data Sharing Collection* (DSC) is initiated by the research administrator upon request by a researcher on the basis of criteria put forward by the center director. Typically, a DSC is initiated after a paper has been accepted for publication. The protocol for initiating a DSC is described in the *DI RDM Protocol for Research Administrators*.

10.2.2 General

A DSC pertains to the reuse of the data that was used for a scientific publication. Every publication of which the results are based on data must have an associated DSC that contains all these data.

A DSC contains (part of) the data of one or more DACs as well as additional files. These additional files contain relevant information, both for external researchers that want to reanalyze the data to extend the published results and those that only want to reproduce these results.

A DSC must be initiated before the proofs of the accepted journal article are sent back to the journal's editorial office.

Contrary to DACs and RDCs, which are only for DI-internal use, DSCs are made available to the international scientific community. This does not imply that everyone

can read and download the DSCs; access to a DSC requires authorization, and this can only be granted by the DSC manager.

Finding a DSC on the web must be easy. Therefore it must contain so-called *discovery attributes*.

Sharing data of human participants requires that their privacy must be guaranteed. Therefore these data can only be shared in a way that conforms to the relevant laws and regulations.

The results of some publications do not depend on newly collected data, but on published (shared) data. The DSC for such a publication must not contain a copy of these published data. Instead, this DSC must only contain the data that is unique for the publication. This may involve computer scripts that implement novel analysis methods, or novel empirical data that is compared to the published data.

10.2.3 Discovery attributes

The researcher must add both general and neuroscience-specific attributes that allow the DSC to be found on the web. The neuroscience-specific attributes allow the DSC to be found independently from the original publication. The general attributes, which all pertain to the publication, are the following:

1. Title of the publication
2. List of all co-authors
3. Journal in which the publication has appeared (will appear)
4. Persisten identifier (e.g., a DOI) of the publication
5. PubMed identifier (PMID) of the publication

There are two types of neuroscience-specific discovery attributes:

1. Free text attributes, of which there are two subtypes:
    1. Keyword, like the keywords of a journal article.
    2. Description, like the abstract of a journal article.
2. Neuroscience terms that are selected from two controlled vocabularies: the topic list of the Society for Neuroscience (SFN), and the Medical Subject Headings (MESH; used by PubMed for indexing articles). When selecting terms from these controlled vocabularies, the following domains may serve as a guide: technique, topic, species, disorder, and brain area.

10.2.4 Conformity with laws and regulations

Sharing human data requires that the participants have signed an informed consent form that allows the data to be shared. In case the data were acquired at the DI, this must be documented as a part of the relevant DACs (i.e., the DACs that contain the data on which the results are based, as specified in the RDC).

Sharing human data also requires that these data do not contain elements on the basis of which the human participants can be identified in a direct or indirect way. Two types of information are especially relevant: personal background information and research-related information. (In the Dutch Data Protection Act, these are denoted as, respectively, "communicatie gegevens" and "onderzoeksgegevens".) Personal background information pertains to the subject's past and present status with respect to health, education, occupation, activities, etc. Personal background information on the basis of which a subject could be uniquely identified (e.g., name, bank account number, Burger Service Nummer) can never be shared. Some personal background information may be essential in order to reproduce the results on which the publication is based. Only under this condition, this personal background information may be shared. This requires however that the Data Use Agreement (DUA, see further) specifies that the collection contains sensitive data and that the agreement specifies how this sensitive data may or may not be used.

It is sometimes possible to identify human participants on the basis of research-related information such as anatomical MRIs, video-, and audio data. These are so-called identifiable human data. Tools are available that remove the information that allow for this identification, such as software for removing facial characteristics and ears from anatomical MRI, and transcriptions of audio data. If these tools are available, researchers must use them to de-identify their participants' data.

10.2.5 Preparing the data for sharing

One must distinguish between the situation in which all data in the DAC can be shared, and the situation in which some data of the DAC must not be shared (see 10.2.5). In the former situation, the data are prepared for sharing by copying them from the DAC(s). In the latter situation, the researcher must do the following:

1. Download the relevant data from the DAC(s).
2. Perform all the operations that are required in order for the data to be shared (see 10.2.5).
3. Upload the result of the previous step to the DSC.

10.2.6 DSC annotation

See 8.2.9.

10.2.7 Reproducing and extending on the published results

The DSC must contain all the information that a knowledgeable colleague needs to reproduced and extend on the published results. The first step of a reanalysis typically involves that the published results are reproduced. Publications may differ substantially in the amount and detail of the information that is required for reproducing the results. Collaborators on a collection are expected to demonstrate scholarship in providing the required information.

It is highly unlikely that an external colleague would only want to reproduce the published results using exactly the same computer scripts that were also used by the authors. Instead, it is much more likely that such a colleague would want to investigate related effects, or modulations of the published effects by variables that are present in the data. This typically constitutes the second step of a reanalysis. Collaborators on a collection are expected to provide the information that allows external colleagues to perform such analyses, amongst others by providing inline comments in their analysis scripts.

10.2.8 Specifying the data use agreement

Every DSC requires a *Data Use Agreement* (DUA) that specifies the conditions under which data is shared. The repository offers several standard DUAs, both for human and for non-human data. The DUA should adhere to the legal standards and the local policies. The DUA may also include details on specific intellectual property rights and limitations on the reuse of the data.

The RDM documentation includes a decision tree to guide the choice for a particular DUA.

When an authenticated external researcher agrees to the DUA corresponding to the DSC, he/she is automatically added as a viewer to the DSC, which gives him/her read access.

10.2.9 Associating to the RDC

All publications of which the results depend on data are also represented in the repository as a RDC, and therefore a DSC can be associated with one or multiple RDC. The DSC must be associated with the corresponding RDCs.

10.2.10 DSC closure

See 8.2.10.

There is one important difference between the closure of, on the one hand, DSCs, and on the other hand, DACs and RDCs: after closure of a DSC, this collection is assigned a persistent identifier that is exported to the web and (after clicking on it) is resolved to its so-called *landing page* (a special page on the repository's web interface); the persistent identifiers that are assigned to closed DACs and RDCs are not exported to the web and are not resolved to a special landing page. A closed DSC's persistent identifier has the same status as the persistent identifiers that are typically used to refer to published journal articles, such as the *digital object identifiers* (DOIs). A persistent identifier allows for direct access to a digital object, even if its URL has changed.

As prerequisite for publication, several journals now request a persistent identifier of the shared data, which they publish in the footnote of the article. DI researchers can use the PID of the closed DSC for that purpose.

If a closed DSC contains incorrect and/or insufficient information, a collection manager may re-open it, allowing for changes to be made. If this re-opened DSC is modified and thereafter closed again, then a second snapshot of the DSC is generated, with a unique persistent identifier.

10.2.12 Data/study type specific

10.2.12.1 Photo-, video-, and audio data

It is technically impossible to de-identify photo-, video-, and audio data without compromising their scientific value. At the same time, it is unrealistic to expect that most candidate-participants for these studies would agree with their data being shared without de-identification. For these reasons, the DSC for a published study that involves photo-, video-, or audio data must not contain the identifiable data of those participants that have indicated on their informed consent form that they did not agreed with sharing. However, the DSC that corresponds to this study must contain the identifiable data of all participants that *did* agree with sharing. In addition, if the published study also contained non-identifiable data (e.g., response times, electrophysiological data), then all those data must be shared (i.e., the non-identifiable data of all participants).

10.3 Recommended

10.3.1 Analysis scripts

Often, the process via which data are converted into published results is fully specified by analysis scripts that can be executed by software packages such as MATLAB, R, Python, SPSS, Bash+FSL, etc. Sharing analysis scripts is the recommended way of providing information to colleague allowing him/her to reproduce the results in the publication.

10.3.2 Checking contributions

As a part of his/her final responsibility for a collection, it is recommended that a collection manager check the contributions of all persons that contribute to this collection. For that purpose, a collection manager can make use of a reporting tool that produces a log of all the changes in the collection, specifying who uploaded/modified which file at what time.