ZIHUb

Mental Health Alliance

Guide to Research Data Management ZIHUb

Process phases

Research data management and data protection

Guide to Research Data Management

Process phases
Research data management and data protection

Content

1 Introduction

Large sums of public money are invested in research projects. The comprehensible correctness of research results is therefore an important element for the trust of society in science as well as the trust of scientists among themselves. The ***data underlying research results, as well as their management and quality, therefore play a prominent role and are part of good scientific practice (DFG, 2013).*** This applies in particular to projects involving human studies, since on the one hand statements are made about potentially health-relevant results and on the other hand the protection of sensitive health data of the subjects is of paramount importance, especially in studies that may touch on stigmatising diseases (Pommerening et al., 2014).

The availability and use of ever more modern measuring instruments and analytical methods is creating ever larger volumes of increasingly complex research data in different variants, most of which is available in digital form. In the research processes at the ZI, different types of research data are constantly being generated, including extremely sensitive human data, e.g. genome-wide data or imaging data from which the identity of the person can be reconstructed. The rules of good scientific practice as well as legal regulations on the protection of personal data contain indications and concrete requirements for the handling of research data. At the same time, explicit demands for the public availability of research data have emerged in the research environment from third-party funding bodies and international journals, which express the value of research data. Various guidelines for handling and re-using research data have been developed. For an introduction to the topic, an online tutorial has been created, for example, with ***"Mantra"*** from the University of Edinburgh[1]. Like other research organisations, ***Heidelberg University*** has also published ***a Research Data Policy***[2] with some basic guidelines for research data management. With ***"Scientific Data"***[3]***,*** a Nature journal has been created that is exclusively dedicated to the publication of descriptions of data sets. However, as far as human data is concerned, the publication of research data is strictly limited by German data protection legislation.

The question arises as to how a systematic handling of research data can be realised that takes into account the problem areas, especially in dealing with data protection, and at the same time pays attention to the developments in the research landscape. Which areas are relevant in detail for effective research data management that complies with the legal requirements is often difficult to decide due to the high level of complexity as well as the only generally formulated legal texts.

Research data management encompasses all aspects of collecting, storing, making available, preserving and archiving data and is to be regarded as an essential area of responsible and honest research.

[1] http://datalib.edina.ac.uk/mantra/
[2] http://www.uni-heidelberg.de/universitaet/profil/researchdata/
[3] http://www.nature.com/sdata/about

This ***guideline aims to*** support ***the targeted handling of research data*** against the backdrop of the ever-increasing and in part contradictory requirements currently developing in legislation and the research environment. It makes the ***level of the data workflow in the research process transparent with the associated process requirements in order*** to enable forward-looking and effective research data management that meets current developments. It shows rules and recommendations for individual process phases of the data workflow and, in particular, integrates the requirements for research data management with human data resulting from the data protection laws of the federal and state governments. The guide takes the perspective of the scientist in the research process.

The guidelines are based on the current state of the international discussion in the field of research data management as well as the Research Data Policy of Heidelberg University and were adapted to the subject-specific characteristics at the ZI in cooperation with the project group. In addition, this guideline can be used as a ***basis for creating a data management plan***, which is increasingly required in applications for third-party funding. It is to be regarded as a basis that is subject to constant further development, especially in cooperation with the group of ***decentralised data protection coordinators at the ZI***.

First, a precise definition of different variants of research data is made, the area of tension between the different requirements for research data and its management is shown and existing concepts for research data management are briefly presented (chapter 2). The requirements for the management of research data from a data protection perspective are discussed in chapter 3. Subsequently, based on existing concepts and regulations, a guideline is developed that provides rules, design guidelines and checklists for the individual process phases of research data management in accordance with the life cycle of research data and, in particular, integrates the requirements of data protection (Chapter 4).

2 Definition of research data and related terms

In empirical sciences, research data are an essential basis for scientific work, for obtaining results and for the resulting publications.

According to the DFG's "Guidelines on the Handling of Research Data" (2015), **research data** are among other things

- **Measurement data,**
- **Lab values,**
- **audiovisual information,**
- **Software,**
- **Simulations,**
- **Rehearsals,**
- **Cell cultures or**
- **Methodological test procedures such as questionnaires or survey data**

and thus the basis for generating or validating research results (DFG, 2015, DFG, 2013, Mantra, 2016).

Research data can either be collected and analysed for specific individual projects, transferred to databases for long-term studies or subsequently used by third parties against the background of other research questions and analytical goals. Which type of data is generated in which format in the respective research process depends on the discipline; the measurement instrument used is also decisive for the format of the research data generated.

2.1 Areas of influence on research data management

Research data are exposed to a field of tension of different, dynamically developing and partly diverging demands (see Figure 1), which are briefly discussed below.

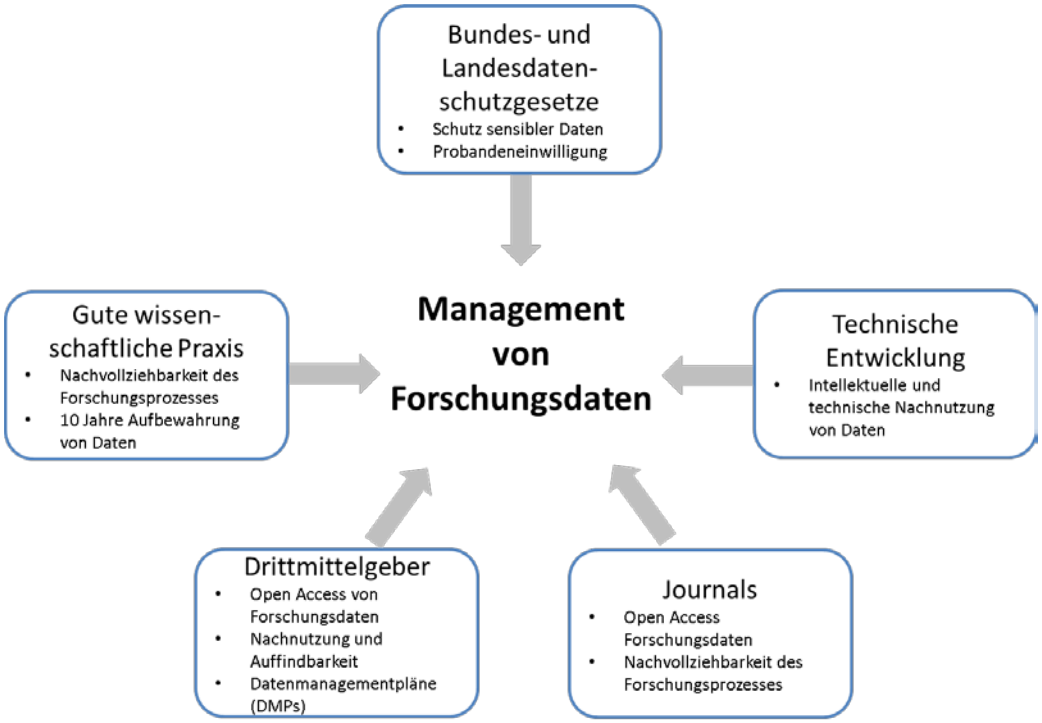

Fig. 1: Areas of tension in the requirements for research data management (own illustration)

Good Scientific Practice

Research data enable the traceability of research results. They are usually collected at great financial expense and with the help of great expertise.

To ensure that all analytical steps and published results are traceable, good scientific practice dictates that research data that form the basis for publications should be stored for at least ten years after publication on durable and secure media at the institution where they were generated (DFG, 2013).

The responsibility for the long-term storage of the data on which a publication is based lies with the scientist responsible for the project. Good practice is to store this data separately with the corresponding publication.

When the researcher leaves the institution, he or she can receive a copy of the data from his or her projects. Depending on the respective subject, it must also be specifically defined whether the raw data actually need to be archived in the long term, or at what level of aggregation of the data this appears to make sense (DFG, 2009). Subject-specific definitions are currently being developed; for psychology, please refer to Schönbrodt et al. (2016), where the level of aggregation of psychological research data to be archived is defined and the DFG guidelines on research data management for the field of psychology are concretised....

Third-party funder

In addition to this requirement of good scientific practice with regard to research data, further demands have arisen in recent years. Third-party funding bodies now demand the publication of research data wherever possible, e.g. in discipline-specific or institution-specific [4]repositories such as ***HEIDATA at Heidelberg University***. ***R3data.org***[5] is an online directory of discipline-specific research data repositories supported by various research organisations, including the DFG.

Research data with a DOI[6]or CC licence can become a citable research "product". The research policy objective of third-party funders is in particular the subsequent use of research data for further analysis purposes. For research data management within the framework of a project, separate funding can be applied for from the DFG, BMBF and the EU. The EU Commission has introduced the so-called Pilot Data Open Access using the FAIR principles - Findablity, Accessibility, Interoperability, Reusablity (see also Wilkinson et al., 2016) - and thereby enables all newly applied for or funded projects to establish a robust and systematic data management including a data management plan and to make research data - where possible - publicly available for subsequent use (European Commission, 2016). Data management plans have in some cases become an integral part of project applications, so that online systems have been developed to support the creation of such a plan. [7]

Journals

Journals also require the publication of research data on which papers published in them are based. In all Nature and PLOS journals, making all materials, data, codes and protocols available to interested readers

[4] https://heidata.uni-heidelberg.de/
[5] http://www.re3data.org/
[6] https://creativecommons.org/
[7] e.g. http://www.dcc.ac.uk/dmponline

is a prerequisite for accepting a paper for publication. Possible restrictions in this regard must be made clear when the paper is submitted.

Technical further development

Further developments in technology and in the scientific community can affect the technical and intellectual usability of research data and pose a major challenge to the management of research data. Examples of further developments include new data or file formats, new interfaces or new scientific working methods.

Data protection laws

The protection of e.g. highly sensitive health data of participating test persons is of outstanding importance in biomedical research projects. It is in the common interest of scientists and patients or test persons to minimise all risks and impairments to the participating patients or test persons that may result from inappropriate handling of their personal or personally identifiable data. Research must always strive to achieve a balance between the public interest in progress in the therapy of diseases and the individual interests of the participating subjects with regard to their informational self-determination. For the management of research data, this results in high demands on the design of subject consent, the pseudonymisation and anonymisation of data, as well as the protection and security of personal data and research data of the subjects. The ***principle in data protection "prevention is better than prohibition" must be met as well as possible.*** At the same time, a balance must be struck between implementing an appropriate and feasible level of protection that enables and does not hinder research on the one hand, and preventing data misuse on the other. Especially in projects dealing with potentially stigmatising diseases, the protection of personal data is of utmost importance. (Pommerening et al., 2014)

2.2 Current developments in the environment

In September 2016, the Helmholtz Association published[8] a position paper on the handling of research data, which addresses the changes in the research world brought about by the availability of digital research data and the associated open science movement. The main postulates are the development of an appropriate information technology infrastructure and the publication of research data generated by scientists of the Helmholtz Association. The individual centres of the Helmholtz Association will develop discipline-specific guidelines that can also take into account legitimate reasons for temporary withholding or permanent strictly controlled access, for example to protect personal data or the rights of third parties or the well-balanced interests of participating scientists.

The core of this initiative is the Helmholtz Data Federation (HDF), which aims to develop an internationally networked research data infrastructure. Coordinated by the Karlsruhe Institute of Technology, the HDF will establish a data infrastructure to improve research data management in the Big Data area. It is planned

[8] See https://www.helmholtz.de/fileadmin/user_upload/01_forschung/Open_Access/DE_AKOS_TG-Forschungsdatenleitlinie_Positionspapier.pdf

as the nucleus for a Germany-wide research data infrastructure that will be open to the entire German science system and compatible with the European Open Science Cloud currently under development. The aim is to make data permanently and securely storable and reusable.

3 Data protection in biomedical research: Legal regulations for handling human data

3.1 Fundamentals of data protection law

The explanations in this section discuss the declaration of consent as well as the anonymisation and pseudonymisation of personal data, i.e. data that allow the person to be re-identified, as central elements of the data protection principles that are a prerequisite for research work with human data in Germany and Europe. They are based on Pommerening et al. (2014, p. 36ff.). In ***general, research with human data must always pursue a balance between the societal interest in advances in treatment options, including future ones, with the individual interests of the subjects involved.*** The data used in research projects are in principle to be classified as special personal data according to § 3 para. 9 of the Federal Data Protection Act (BDSG) or [9](health data). The research clauses specifically formulated in the law for health data require that subjects be informed and consent obtained, and that anonymised and pseudonymised data be used with priority.

Declaration of consent

The legally permissible use of human data for research purposes, which must preserve the subjects' informational self-determination, requires, with few exceptions, the existence of a declaration of consent. This must be given voluntarily and without concern for possible disadvantages in the event of refusal[10]. In the process of information by medical staff for obtaining consent, it must be taken into account in the case of ***patients that,*** due to their illness, they may feel dependent on the treating staff asking for consent. Any conveyance of an attitude of expectation is to be avoided, especially in the case of patients, e.g. with regard to special treatment. In the consent carried out before the start of data collection, it must be clear to the patient or test person for which processing, which research question, which use of which data during which storage period the consent is given.

[9] see § 13 para. 2; § 28 para. 6 BDSG; § 33 LDSG BW

[10] For the requirements, cf. § 4a BDSG, § 4 LDSG BW

However, the more concretely the consent is formulated, the more restrictive it can be for the research: The more precisely the purpose of the data processing is stated, the more precisely the necessary data set, the group of persons required for the data processing as well as the project duration can be specified. In contrast, a more open description of the data collection and storage is usually accompanied by a lower restrictability of the data scope, a longer storage of the data and a larger group of persons entrusted with the data processing. It is widely recognised that in biomedical research it is often difficult to limit oneself to a specifically identifiable research question. For this reason, it is often accepted that only disease-related restrictions on data use are formulated in the consent. An example of this is the use of data for research on PTSD. The concreteness of the purpose-relatedness thus only serves the goal of informational self-determination of the subjects or patients as long as the restrictions of the research questions can also be understood and comprehended by the majority of the subjects and patients. The requirement of comprehensibility of the declaration of consent can thus already mean a softening of the principle of the narrowest possible definition of the purpose-relatedness of the data collection. For example, the formulation of a restriction of the research direction to a specific subtype of depression does not tend to provide any information gain for the patient or test person if he or she cannot sufficiently distinguish this subtype from the overall concept of "depression". However, it is possible for every test person or patient to distinguish between storage of the collected data for an unlimited period of time and storage limited to e.g. five years. Similarly, patients can distinguish between the use of data at only one research site or in national or European networks.

Anonymisation and pseudonymisation

Data protection laws define the anonymisation and pseudonymisation of data[11]. Both are used to exclude the assignment of data to a specific or identifiable person or at least to make it significantly more difficult. Both procedures are therefore fundamentally suitable for reducing the need for protection of the collected data or minimising the risks of identifying a person, which can be given by the storage and processing of the data.

According to § 3 para. 6 BDSG, § 3 para. 6 LDSG BW, data is anonymised if it can either "no longer" or "only with a disproportionate effort in terms of time, costs and labour be assigned to a specific or identifiable natural person". Anonymisation of the collected data can therefore generally be assumed if identifying and collected human data are separated, there is no longer an allocation rule to the associated persons and a re-identification of the person is no longer possible on the basis of the data alone. Anonymised data are no longer considered personal data, so that the regulations of data protection legislation, including the requirement for consent, i.e. the necessity of prior consent to the use of data, no longer apply to them; this data must therefore no longer be handled in accordance with the rules of data protection legislation.

[11] see § 3 para. 6 BDSG, § 3 para. 6 and 7 LDSG BW

A problem in dealing with anonymised biomedical data is that this status can change over time, e.g. if a user of the data can infer the identity of the person based on a combination of medical and social data. As an example, consider a borderline patient who moved from a larger town where she lived at the time of data collection to a small village with about 300 inhabitants, where she is the only person suffering from this disorder; the combination of place of residence and disorder thus makes re-identification possible, even if the personally identifiable information (name, etc.) has been deleted. In such situations, the data would again be personal data or data that could be related to a person, and would have to be handled in accordance with the data protection laws of the federal and state governments. ***The problem for researchers who are responsible for the secure handling of data is the fact that it is not always possible to assess in advance with sufficient accuracy whether and when such a scenario may occur.*** As a preventive measure, it is therefore recommended that even anonymised data should only be made available to defined user groups for specific purposes and, in particular, that the free availability of medical data on the Internet in data repositories should be avoided for the time being.

Pseudonymising ***data*** is, according to § 3 para. 6a BDSG, § 3 para. 7 LDSG BW, the "replacement of the name and other identification features by a mark for the purpose of excluding or significantly complicating the identification of the person concerned". In contrast to anonymised data, pseudonymised data still has an allocation rule to the identity data of the corresponding persons. However, the attribution rule is not known to all users of the data. The definition of a project-related data protection concept therefore requires a distinction between groups of persons who know the assignment rule and those who do not. ***Since the risk of re-identification of persons cannot be ruled out with sufficient certainty in the case of pseudonymised data, the handling of pseudonymised data must be aligned with data protection law, i.e. corresponding to the handling of personal data.***

A comparative analysis of anonymisation and pseudonymisation shows that the security advantage of anonymisation is strongly relativised due to the ever-increasing availability of extensive medical data sets and an inherent risk of re-identification of persons, even on the basis of actually anonymised data. ***When choosing the appropriate procedure to reduce the need for data protection, it is more decisive that many research scenarios, e.g. follow-ups, cannot be realised with anonymised data.***

Data protection solutions in research with human data usually require a separation of informational powers, i.e. independence of access to different components and parts of the data set. Central to this is the subject list, which stores the identifying patient data with the associated pseudonyms. The subject list must be stored in a secure steel cabinet to which only the project manager, for example, has access. The involvement of a trustee would mean that the administration and storage of this list is the responsibility of an organisation or person who is legally, spatially and personally independent. Since the very fact that a person's name and address or other identifying data is stored in such a list may possibly say something about the person's condition, this data should also be considered sensitive and worthy of protection. The long-term storage and processing of sensitive pseudonymised health data requires a legally sound regulation of responsibility for the data that is comprehensible for each participating patient or test person.

According to § 34 (1) BDSG, § 21 LDSG BW, every subject has the right to information about the personal data stored about him or her, including derived data, which is only possible as long as the data has not

been anonymised. "Derived data" in this context refers to e.g. analysis results. According to § 35 BDSG, §§ 22, 23 LDSG, probands have the right to have their personal data corrected, deleted or blocked. This right may conflict with the requirement of good scientific practice to keep data on which a publication is based for more than 10 years.

3.2 Basic principles of data protection-compliant solutions

Pommerening et al. (2014, p. 58 ff.) list the following ***elementary principles as a basis for data protection-compliant solutions in biomedical research projects*** that implement the elements required by law:

- Reduce the risk of unauthorised re-identification as much as possible;
- Realisation of an informational separation of powers through separate storage of identifying personal data from medical data, in the maximum case with an independent trustee;
- If necessary, double or multiple pseudonymisation, especially for biomaterial;
- Long-term secure cryptographically generated pseudonyms and re-identification only along intended processing and approval procedures;
- Careful consideration of anonymisation and pseudonymisation for the intended use of the data;
- Transparent and clearly regulated responsibilities;
- Combination of technical and organisational measures for the protection of data;
- Redundant protection for the protection of data;
- Solutions that are as simple and economical as possible, adapted to the need for protection and proportionality;
- Best possible use of the data collected at great expense and, if necessary, at the personal risk of the test persons, in accordance with the informed consent;
- Informational self-determination of the subjects with right to know and right not to know;
- Avoidance of role conflicts, in particular to avoid any circumvention of the principle of informational separation of powers.

4 Guidelines for Research Data Management at the ZI

In recent years, various guidelines have been established for the targeted management of research data during the research process and after the completion of a project. In particular, the models of the Digital

Curation Center (DCC) [12], WissGrid (Ludwig, Enke, 2013) and FAIR - Findability, Accessibility, Interoperability, and Reusability - Data Principles (Wilkinson, 2016) should be mentioned. Other discipline-specific examples include the guidelines on research data management in the geosciences (Bertelmann et al., 2014) and in the social sciences (Jensen, 2012). In Germany, [13]a project dedicated to research data management has emerged from the scientific community in the form of the "Digital Information" priority initiative of the Alliance of German Science Organisations.

What the models have in common is the ***perspective on the entire life cycle of research data and thus the planning and realisation of the data workflow during and after the implementation of the actual research project with the linking of recommendations and checklists for the individual process phases of the data workflow***. Since data protection issues in biomedical research projects that use human data are only listed in very general terms in these guidelines, the present guideline for research data management at the ZI integrates the explanations of Pommerening et al. (2014) and also Harnischmacher et al. (2006) on the topic of data protection.

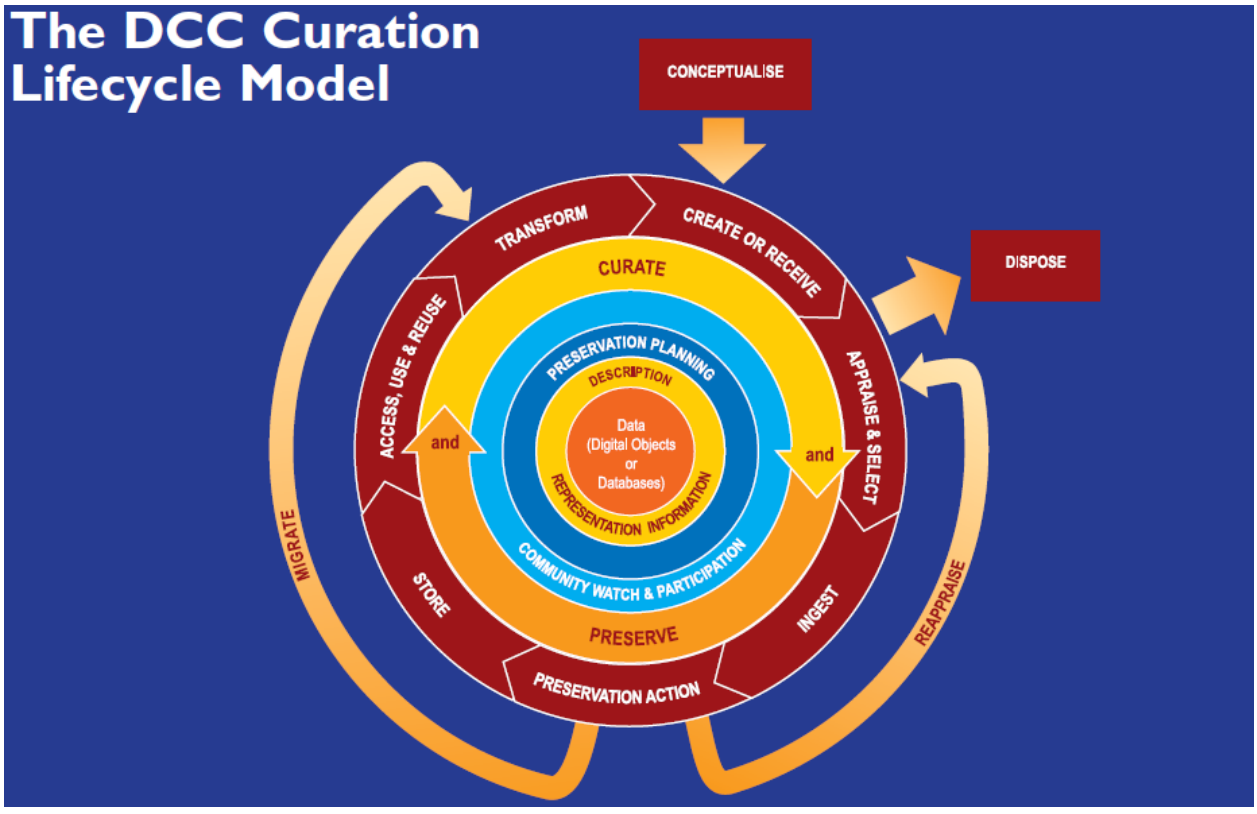

Fig. 2: The DCC Curation Lifecycle Model (Source: DCC)

[12] http://www.dcc.ac.uk/
[13] http://www.allianzinitiative.de/de/handlungsfelder/forschungsdaten/

Figure 2 provides an example of the DCC's very detailed model for the life cycle of research data during and after the lifetime of the actual project for information purposes.

Figure 3 shows the model of the life cycle of research data at the ZI developed within the framework of this guideline with elements important for the individual process phases, including data protection concerns. The individual process phases with associated tasks and checklists are explained in more detail in the course of this chapter.

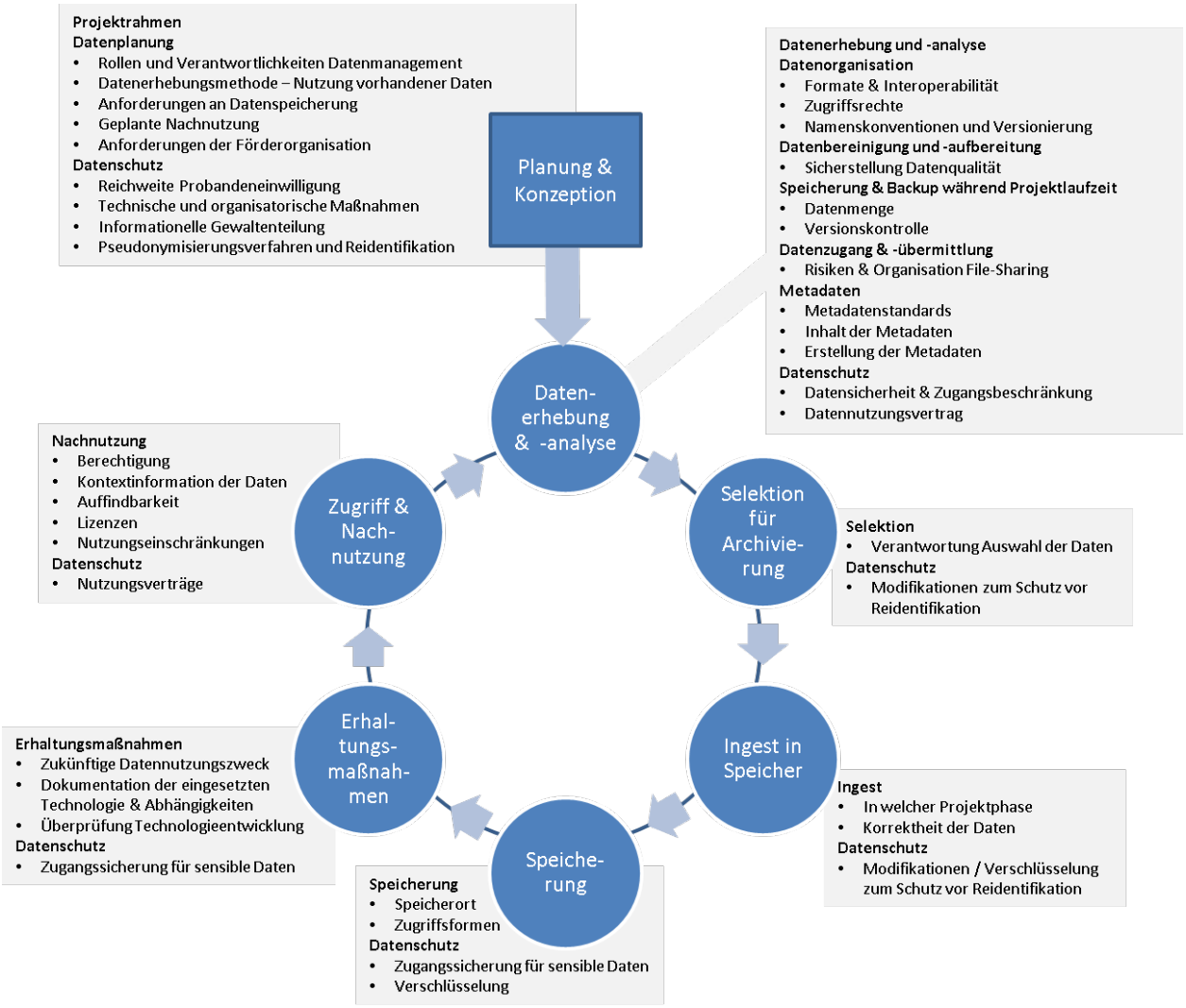

Fig. 3: Process life cycle of research data including data protection ZI (Own representation)

The ***benefits of systematic management of research data throughout the lifecycle of the data*** are multidimensional and include, in addition to the

- Ensure the integrity and reproducibility of results the

- In line with demands from research funding organisations, the
- Improving the quality of the data, a
- increasing the efficiency of the research process in the short, medium and long term, a
- Increasing data security and thus minimising the risk of data loss, as well as the
- Prevent duplication of data collection by re-using existing data.

In the following, the individual phases of the life cycle of research data (Fig. 3) are examined in more detail and provided with rules, recommendations and checklists in order to support the targeted and planned management of research data throughout the entire life cycle while at the same time taking data protection concerns into account. For this purpose, the guidelines from DCC, WissGrid, Research Data Management in the Social Sciences and Research Data Management in the Geosciences were used as a basis, expanded to include data protection and provided with examples from various departments and working groups of the ZI.

All checklists are the basis for the creation of Data Management Plans that may be required in project applications!

4.1 Planning and conception of the research data life cycle

The objective of a research project determines which data are collected and processed. Thus, at the beginning of the life cycle of research data is the planning of the project with an initial rough conception of the life cycle of the project's research data as well as basic questions about data protection in the case of a project with the use of human data.

Checklist external framework of the project

- Project name
- P.I.
- Funding organisation & potential specific data management requirements
- Internal and external project partners

Checklist data planning

- Who is generally responsible for data management in the project? Who is responsible for the individual phases in the research data life cycle? What human resources need to be brought in?
- Is training in research data management necessary for project staff? Who carries out this training?
- How is responsibility divided in the case of several project partners?
- What kind of data is generated or used in the project?

- Will new data be generated in the project? How will the data be generated (which devices, technologies, procedures, etc.)?
- Will existing data be used for the project? From within the institution or from outside? What are the access rights and copyrights?
- What are the roles and responsibilities in the data lifecycle?
- What kind of subsequent use of the data is planned? Is a publication of the data planned?
- What are the funding agency's specific requirements for handling or re-using the research data?
- What data storage requirements will arise in the project? What are the expected data volumes and production rates? On which media should the data be stored?
- In which formats should the data be recorded?
- Is the data genuinely digital or are analogue objects also digitised?

Example contents Data planning SniP:

For imaging data, larger data storage is needed, which is provided centrally, so that no special requirement analyses are needed in the planning of projects, except for a rough estimate of the number of data sets or the amount of storage to be expected. Currently, a FreeNAS system (zifnas), which enables file compression, snapshots and replication (instead of backup on zifnas2), is used for storage. The storage takes place redundantly on commercially available hard disks. A change of media for e.g. backup is not planned here due to the amount of data (currently about 35 TB compressed).

The project staff are responsible for *data management* within the projects. This includes the possible transfer of data from the tomographs to the project directories as well as the preparation of data that is initially available in paper form.

Local IT administration is provided for *organising storage and backup.* However, part of the responsibility also lies with central IT (e.g. replication of data and availability of systems).
For imaging data, there is also a predefined basic structure into which all projects fit (if possible).
Under a project root directory (e.g. SFB1158B09, under /zifnas/projects) there are then

- /ARCHIVE: DICOM, NIFTI raw data (unprocessed, at most converted) and log files (Presentation etc.)
- /data: Preprocessing and primary analysis (SPM: preprocessing / first level), e.g.
    - subj_id
        - anatomy [MPRAGE with segmentation]
        - fmri
            - {task} [functional tasks e.g. nback, faces, ...]
                - pre [Pre-processing]
                - {standard_indirect, *standard_direct}*
                - first [first-level Analyse]
                - {standard_indirect, *standard_direct}*
                - onsets [processed log files]

- /neuro_quest: Data tables (SPSS, Excel, etc.)
- /user: subdirectories with analyses defined by staff (ideosyncratic, second-level analyses)
- /documentation: Project documentation, questionnaires in PDF, flow charts etc.
- /batch: programmes/scripts for standard processing in all projects

- /QC: Quality control data/information, e.g. motion parameters, signal-to-noise ratios, spikes of imaging data.

Project staff undertake the conversion (DICOM to NIFTI) and pre-processing of the imaging data according to at least the defined standard procedures (depending on the sequence [e.g. 32- vs. 12-channel EPI] and software version [e.g. SPM8 vs. SPM12]).The pre-processing and primary analysis (first level) is the same across projects if possible. The project data remain in this structure for the entire duration of use in the project and possible subsequent use.
Genetic data (genotyping, usually now whole chips with 500k SNPs) or data from other biomaterials (saliva, hair, e.g. cortisol) are organised across projects and have their own storage location. For these, an abstraction layer exists for organisation in the form of a web application.

In the case of several project partners, data storage and management are divided accordingly. In the PeZ project, for example, the geographical data of the test persons are stored by the geographers and the imaging data in the ZI. The appropriate structures are already available there.

Example data planning addiction clinic:
-What kind of data is generated or used in the project?

Paper data, data in survey databases (online), digital data in the form of MS Office documents, SPSS files, imaging data, log files and other text files.

-How is the data generated (which devices, technologies, processes, etc.)?

Completion of CRFs, MRI measurements (MRI scanner), neuropsychological experiments (PC), online input into survey databases (e.g. EvaSys).

-On which media should the data be stored?

Central server, managed by the IT department ("zifnas")

-In which formats should the data be recorded?

Analogue and digital, MS Office formats, SPSS format, ASCII files, MRI: DICOM, Analyze, NIFTII

Example data planning KFO:

- Roles and responsibilities Data management
  - Dedicated data manager and responsible for data protection: Martin Jungkunz
  - Responsible for the project: Christian Schmahl
- Data collection method-Use of existing data
  - Questionnaire survey
  - Laboratory measurements
  - fMRI measurements

- Personal data
- Data storage requirements
  - Questionnaire and personal data:
    - Access via web-based database
  - Laboratory and fMRI data:
    - Network storage
- Planned subsequent use
    - Contact data of the test persons are stored for follow-up studies after the end of the project, provided that a declaration of consent is available.
- Requirements of the funding organisation: none

Example data planning psychopharmacology:
The individual project members are responsible for data management, including the preparation, digitisation and transfer of raw data to storage. Previously, storage/backup took place on standard external hard drives, but recently a larger central server has been made available at zifnas and the transfer of existing data is underway.

Example of data planning KliPs:
Types of data generated in the project:

Attitudes and feelings (questionnaires), reaction times and number of correct ones (behaviour in experimental paradigms), physiological reactions (e.g. MRI, EDA, EEG).

Reproducibility of the data:

The data of the projects is not reproducible. However, the results are reproducible by applying the saved scripts to the converted data.

Original data, adjusted data, analysis data:

Original data are available in paper form, in the form of text files (logfiles) and in the form of text or binary files in different generic formats depending on the physiological measurement (e.g. BDF for peripheral physiology, Brain Vision Exchange for EEG data) and in the form of DICOM data for MRI data. The DICOM data are converted to NIFTII format and pre-processed, which is equivalent to cleaning. Neither the questionnaire data nor the behavioural data (from the log files) are cleaned, but they are either digitised manually (questionnaires) or transferred to SPSS via data conversion. However, it may be necessary to exclude outliers when checking the data. For the analysis of physiological data, special programmes are used, so that the format of the analysis data differs between the measurement methods.

Documentation of the origin of the data:

The origin of the physiological as well as the behavioural data is stored by headers in the original data. The origin of the questionnaire data is documented in an SPSS file.

Checklist Planning & Concept Data Protection

- Are human data generated or used, so that the regulations of the data protection laws have to be applied?
- Subject consent: The prerequisite for collecting data or obtaining and storing biomaterial for research purposes is the written consent of the material donor after appropriate written and verbal clarification. Is the planned storage period of the data for a long time? Are later questions, not yet known at the time of informed consent, planned to be investigated on the basis of the data? If so, increased requirements apply to patient consent as well as the pseudonymisation procedure. In the case of an indefinite storage period of the data, the declaration of consent must be formulated in a far-reaching manner and the subject must be made aware of the openness of the research project by stating the disease being investigated. For detailed information on the design of subject consent, please refer to Harnischmacher et al. (2006) and Pommerening et al. (2014, p.36 ff.). Discussed examples of consent can also be found at the Working Group of Medical Ethics Committees of the Federal Republic of Germany [14].
- What technical and organisational measures are implemented for the special protection of personal patient or proband information under the premise of appropriateness (e.g. access restrictions, cryptographic transformations, dual control principle for important decisions)?
- How is the separation of informational powers, i.e. separate storage of personal identification data and collected data in different rooms under different responsibilities, separate storage within the institution or using an external data trustee who is not under authority?
- What procedure is used to pseudonymise the data (single coding, double or multiple coding)? Does this minimise the risk of unauthorised re-identification as far as possible? Are the pseudonyms to be regarded as secure?
- Under what conditions may a re-identification take place (e.g. renewed contact for the collection of supplementary information or transmission of health-related information)?
- How is it ensured that the key or the assignment list for re-identification is kept secure and as a secret?
- In the event of data being passed on: Does a double pseudonymisation of the data take place to ensure increased protection?
- Can data be anonymised, if necessary, and when in the research data life cycle?

Example planning data protection KFO:
Reach subject consent:

- Contact consent for further study
- Consent for blood collection with the option to pass on to cooperation project
- Consent for use of data within the framework of the studies of the KFO256

Technical and organisational measures

- Information on consent forms (available? Yes/No) is stored in central database
- Multiple account types with different permissions within the database
- Within an account type, access is possible to either personal (contact data) or diagnostic data. This information is not displayed together.

[14] http://www.ak-med-ethik-komm.de/index.php?option=com_content&view=category&id=15&Itemid=105&lang=de

Informational separation of powers
Pseudonymisation procedure and re-identification

- Assignment of numerical subject codes
- Allocation table is access restricted

Example Planning Data Protection KliPs:

Subject consent:

After verbal information about the project and potential side effects, the study participant is given written consent to read and sign. This contains the note that the subject can withdraw his or her consent at any time without disadvantages for him or her, even without giving reasons. In studies in which blood is drawn, the volunteer receives an additional consent form from the Genetic Epidemiology Department, which also includes the use of his/her genetic data for further projects.

Pseudonymisation and protection of personal data:

The respondent is assigned a random code. This code, together with his or her name, can only be found on a list that is stored in a password-protected file. Only the project staff know the password. The list is on a virtual drive, secured by mirroring, to which only department staff have access. All study data of the subject are only provided with the code and are also stored on servers of the Central Institute of Mental Health secured by means of sealing. The consent form is stored separately from all the subject's data in a locked cabinet to which only staff of the department have access. The consent form cannot be used to identify the respondent's code.

Re-identification:

Re-identification of the respondent is only possible during the project. This only takes place if a chance find is discovered and the respondent therefore needs to be contacted. In addition to the consent form for the study, the respondent can sign a consent form to be contacted again. However, this does not contain the respondent code, so that re-identification is not possible via this further consent.

Anonymisation:

In current projects, where it is not necessary to make contact again, the list that assigns names and codes is destroyed at the latest at the end of the project.

4.2 Data collection and analysis

The actual generation of data raises issues concerning data collection and analysis, data organisation, data cleansing and preparation, storage and backup during the project, data access and transmission, metadata design and data protection.

Checklist data collection and analysis

- What types of data are generated or used in the project (e.g. one-off observation and measurement data, simulation data)?
- To what extent is the data reproducible and at what cost? Is it data that can be reproduced automatically without significant manual interaction (e.g. systematically processed data from raw data) or is it data that requires significant time and/or intellectual effort to produce?
- Which data are original data, which are cleaned data (derived from original data after control and correction), which are analysis data (the data that are actively worked with)?
- How is the origin of the data documented?

Example data collection and analysis addiction clinic:
-What types of data are generated or used in the project (e.g. one-off observation and measurement data, simulation data)?

Measurement data (cross-sectional and longitudinal), socio-demographic data, psychometric data

-To what extent is the data reproducible and at what cost? Is it data that can be reproduced automatically without significant manual interaction (e.g. systematically processed data from raw data) or data that requires significant time and/or intellectual effort to produce?

MRI data can partly be reproduced automatically if corresponding scripts are available with which they were generated; however, this is sometimes very computationally intensive
Manual steps such as aggregation of data, data entry, individual analyses, quality control sometimes require high time and intellectual effort

-Which data are original data, which are cleaned data (derived from original data after control and correction), which are analysis data (the data that are actively worked with)?

Original data:
- Paper questionnaires
- MRI data in DICOM format
- Log files

Adjusted data:
- entered / read questionnaire data / psychometric data, socio-demographic data
- Processed MRI data
- Processed log files

-How is the origin of the data documented?
In log files and MRT files: in the header
Questionnaires on paper: Header
Also in SPSS and/or Excel databases

Example Data collection and analysis Psychopharmacology:

-What kind of data is generated or used in the project?

Data on paper, data in specific apparatus associated programmes (MedAssociates, iMedTronic, QuantStudio, Biobserve, etc.), video files, MS Office files, statistics and graphics files (SigmaPlot, SPSS, Statistika, GraphPad).

-How is the data generated (which devices, technologies, processes, etc.)?

Behavioural data is recorded with a variety of different devices (drinking data, locomotion data, self-administration, vocalisation, videos, e-phys recording, etc.).
Laboratory methods according to the evaluation method (PCR, Western blot, in situ hybridisation, immunoplotting, etc.) either via specific programmes based on the equipment or as a subsequent analysis of the generated images.

-On which media is the data stored?

Recently on a central server managed by the IT department ("zifnas"), previously on external hard drives or equipment PCs

-In what formats is the data collected?

Analogue and digital, MS Office formats, and countless specific evaluation programmes/statistics and graphics programmes.

Example Data collection and analysis Genetic epidemiology:
-What types of data are generated or used in the project (e.g. one-off observation and measurement data, simulation data)?

Measurement data (EMA), sociodemographic data, psychometric data, clinical data, genetic/epigenetic/gene expression data

-To what extent is the data reproducible and at what cost? Is it data that can be reproduced automatically without significant manual interaction (e.g. systematically processed data from raw data) or data that requires significant time and/or intellectual effort to produce?

Reproduction of the data requires considerable personnel and financial effort, partly excluded. Manual steps such as aggregation of data, data entry, individual analyses, quality control sometimes require high time and intellectual effort.

-Which data are original data, which are cleaned data (derived from original data after control and correction), which are analysis data (the data that are actively worked with)?

Original data:

- Paper questionnaires and interviews
- Colour intensity data from genotyping

- Machine-determined gene sequences/genotype data/methylation data/expression data
- Gel photos
- Log files

Adjusted data:

- entered / read questionnaire data / psychometric data, socio-demographic data
- Processed sequences and genotype methylation and expression data
- Processed log files

-How is the origin of the data documented?

- In log files: in the file header as well as in separate attached annotation files
- Questionnaires on paper: Header
- Also in SPSS and/or Excel databases

Checklist data organisation

- In which file formats will the data be available? How should the formats be assessed in terms of interoperability? Do the formats allow data exchange within the community and long-term use? Do the formats comply with the common standards?
- How is the integration of existing data and data to be generated organised?
- What existing infrastructure for managing the data (e.g. file system structure, database system, data backup measures) can be used for organising the files?
- How are the access rights to the different data technically realised? What options are there for this or do technical restrictions have to be observed?
- What naming conventions are used for the unique naming of data? How are file systems and files structured and named? What options are there for this or must technical restrictions be observed?
- How is versioning handled, e.g. to make different processing stages of data traceable? How are original data, cleaned data and analysis files named and versioned? Is a versioning system used for version control in extensive and collaborative work? Is the DDI standard used for versioning (three-part version numbering "Major.Minor.Revision", with major changes and achieved work goals indicated by the counter "Major", starting at "1" and subsequent minor steps in the counters "Minor" and "Revision", starting at "0" - e.g. "Version 1.0.1: first revision of major version 1)?

Example data organisation SniP:

Digital data are stored in the usual or typical formats so that their usability over long periods of time is ensured (4.6), a possible exchange with other institutions (possibly 4.7) is possible and a conversion into other required formats is feasible at any time:

- Imaging data: DICOM (international standard for medical data, i.e. not only imaging), NIFTI[.gz] (de-facto standard for the analysis of imaging data with the widespread programme packages, possibly compressed [zlib]), Matlab file format (SPM enforces this proprietary standard) and text formats common for the respective operating systems (e.g. FSL).

- Questionnaires, etc.: SPSS (statistics), Microsoft Office file formats, PDF, OpenDocument and CSV/Text.
- Genetic data: plink (de-facto standard)

Access to digital data is via authentication at the domain and file system access permissions, which allow different rights to be assigned to each user of the SNiP working group. At a minimum, however, persons need a valid user account of the ZI. Special rules and/or community access with dummy users and community passwords do not exist.

Paper records are located in staff offices, mostly in the office of the staff member responsible for the project. Consent forms (MR, genetics, project, etc.) are kept in locked steel cabinets separate from the "analysis user data".

Electronic documents for study organisation are stored separately from the analysis user data on the group drive of the working group. Sensitive data that may be necessary there is protected by access authorisations and/or password encryption.

Recontact information for other studies (i.e. the subject has not yet consented to this study) is made available to authenticated users (recontributors) via a web application. For this purpose, they are trained or instructed accordingly and must additionally document this with their signature. The withdrawal of consent to recontact is implemented by blocking and deleting the entry. In addition, the written consent is marked as invalid.

Example data organisation addiction clinic:
-Which existing infrastructure for managing the data (e.g. file system structure, database system, data backup measures) can be used for organising the files?

Server of IT ("zifnas") and data backup by IT

How are the access rights to the different data technically realised? What options are there for this or do technical restrictions have to be observed?

Authorisation via the Active Directory domain of the ZI

-What naming conventions are used for the unique naming of data? How are file systems and files structured and named? What options are there for this or must technical restrictions be observed?

Project-specific, e.g. file name contains eight-digit subject number; subject number is composed of study number/consecutive number/survey number.

How is versioning handled, e.g. to make different processing stages of data traceable? How are original data, cleaned data and analysis files named and versioned? Is a versioning system used for version control in extensive and collaborative work?

Project-specific, e.g. separate folders for raw data and processed data; for MRI data, specific prefixes for further processed data after specific analysis steps.

Example data organisation KFO:
Formats & Interoperability:
oDatabase is available in MySQL
oData can be exported in . csv, . xls or .sav format
oLaboratory and fMRI data are available in the respective common formats
Access rights
oDiagnostic and personal data: see data protection
oLab and fMRI data: Access authorisation management through Windows Active Directory
Naming conventions and versioning:
oDiagnostic and personal data: Versioning via daily SQL dump (JJMMTT-kfomain.sql)
oLaboratory and fMRI data: is the responsibility of the subprojects and is not managed centrally

Example Data organisation Genetic epidemiology:

-In which file formats will the data be available?

Depending on the type of data. Since many different types of data are generated, many different formats are used. formats are in use. The original output of the measuring instruments is kept as far as necessary and feasible. If a standard file format If a standard file format exists for certain types of data, it will be used if possible (e.g. SPSS). In addition, plain text formats and open standards are preferred to binary and proprietary file formats, unless this would involve disproportionate additional time, storage requirements or loss of information.

-Which existing infrastructure for managing the data (e.g. file system structure, database system, data backup measures) can be used for organising the files?

Versch. IT servers (e.g. "zifnas") and data backup by IT as well as additional backup routines

How are the access rights to the different data technically realised? What options are there for this or do technical restrictions have to be observed?

Authorisation via the Active Directory domain of the ZI; For database computers, password-protected access with encrypted network traffic.

-What naming conventions are used for the unique naming of data? How are file systems and files structured and named? What options are there for this or must technical restrictions be observed?

Depending on the individual projects

How is versioning handled, e.g. to make different processing stages of data traceable? How are original data, cleaned data and analysis files named and versioned? Is a versioning system used for version control in extensive and collaborative work?

Versioning and labelling is partly project-specific.

In certain cases, e.g. with local (epi-)genetic data repositories, also separate folders for raw data and processed data across projects; in addition, specific, partly numbered prefixes for processed data depending on the processing status.

Versioning of Office documents sometimes also includes the date and time as well as the abbreviation of the editor in the file name.
For programme code: Use of automatic versioning and logging of the respective development environment or integration with widespread version control systems.

Example Data organisation Formats & interoperability Psychopharmacology:

There are no specifications, resulting in data management that is self-organised by employee, more or less project-related after the analysis, most data sets are available as Excel files, which facilitates compatibility.

Checklist data cleansing and preparation

- What quality standards are observed in the selection and preparation of data (e.g. standardisation of data processing, peer review of data, data validation, plausibility checks, monitoring to check consistency with source data) to ensure the quality of the data? What effort is required to comply with possible quality standards?
- Is there an established data workflow or de facto standard in the field for handling data and its format?
- Which software should be used for the analysis? Does this limit e.g. the data format or which data formats arise during data processing? What possible limitations arise from the data formats or the software used? Are the data formats proprietary or independent of the tools used? Are proprietary formats also changed in the course of software updates and thus no longer readable for older versions?
- What are the limitations of the data format for the exchange?

Example data cleansing and preparation SniP:
Quality control of imaging data: Functional and structural imaging data are assessed for signal-to-noise ratio (relative to other images of the same sequence) and possible spikes (technical artefacts) using a predefined procedure (dataQuality Toolbox, v1.5). The quality of the preprocessing is checked by experienced staff through visual inspection of the data. For functional data, additional motion parameters (e.g. framewise displacement) are used as a quality feature.

Example data cleansing and preparation KFO:
Ensuring data quality
oDiagnostic and personal data: incumbent on the data manager
oLaboratory and fMRI data: is the responsibility of the subprojects and is not managed centrally

Example Data cleaning and processing Genetic epidemiology:

Data preparation is study-specific.
At the same time, however, standardisation of the processing procedures and, if reasonable/feasible, use of existing or creation of corresponding new automated data processing pipelines including adequate logging of the individual processing steps wherever possible, whereby the user only adjusts individual parameters accordingly within a reasonable framework.
E.G:

- Genome-wide genotype data:
  - Automated extraction of genotypes from the raw intensity data of genome-wide typing with Illumina arrays
  - Pipeline developed in-house for quality control and filtering of extracted genotypes
  - Imputation of untyped SNPs according to the standard protocol used by the Psychiatric Genomics Consortium (PGC).
- Genome-wide as well as local methylation data: In-house pipeline for extraction of methylation values from raw data, quality control and batch correction in preparation for analysis. This integrates various software packages established and widely used in the field
- Clinical data: Independent multiple entry with automatic comparison of the entered versions. Computer-assisted routines for cleaning up inconsistencies. Reading in questionnaires with EvaSys

Checklist for storage and backup during the project

- How large is the planned amount of data? What are the implications for storage, backup, access and archiving?
- How are the data and all versions backed up during the project?
- Who is responsible for backup and recovery? How often does backup take place (automatically or manually) and where (data storage on laptops or external hard drives is very risky!)? How many copies are made? What effort (in terms of time and money) is involved?
- How is data encrypted when it is stored on USB devices for a short time?
- Is there at least one current copy of the data that can be easily accessed in an emergency?
- How is version control regulated for backups?
- What are the risks in terms of possible data loss (e.g. hardware failure, software failure, viruses, power failure, human error)?
- When data is collected in the field: How is secure transfer to the main systems ensured?

Example storage and backup addiction clinic:
-How are the data and all versions backed up during the project?

Server of IT ("zifnas") and data backup by IT

-Who is responsible for backup and recovery? How often does a backup take place (automatically or manually) and where (data storage on laptops or external hard drives is very risky!)?

Folder "searches" on the IT server ("zifnas") is automatically mirrored by IT.

Example storage & backup during project duration KFO:
Data volume
oDiagnostic and personal data: approx. 10GB including versioning
oOrganisational data (recruitment films, pictures, flyers, administrative data): 20GB
oLaboratory and fMRI data: is the responsibility of the subprojects and is not managed centrally
Version control: see above.

Example storage and backup psychopharmacology:
So far, staff members have taken care of backup options themselves; the recently established zifnas server "psychopharma" is managed by IT.

Example Storage and Backup Genetic Epidemiology:
-How are the data and all versions backed up during the project?

Server of IT ("zifnas" as well as flfs01r2) and data backup by IT

-Who is responsible for backup and recovery? How often does a backup take place (automatically or manually) and where (data storage on laptops or external hard drives is very risky!)?

Backup: IT department. Images of the current state are saved every hour, which the user can restore himself if necessary. In addition, daily tape backup for data with longer-term storage requirements.

Checklist data access and transmission for data analysis

- What are the risks regarding data security? How is secure access to data organised for project staff and external project partners?
- How is file sharing with external project partners organised?
- Is additional software needed to support collaboration in the project (e.g. virtual research environments)?

Example Data access and transmission Addiction clinic:
-What are the risks in terms of data security? How is secure access to data organised for project staff and external project partners?

Authorisation via the Active Directory domain of the ZI

-How is file sharing with external project partners organised?

FTP server provided by IT

Example Data access and transmission Genetic epidemiology:
-What are the risks in terms of data security? How is secure access to data organised for project staff and external project partners?

Authorisation to access data internally is via the Active Directory domain of the ZI

-How is file sharing with external project partners organised?

Data is exchanged externally either via IT's FTP server or by sending encrypted archives or TrueCrypt/VeraCrypt containers with adequate password/key length either by mail or, in the case of large archives, by a commercial file delivery service located within the scope of the EU Data Protection Convention.

Example Metadata Genetic Epidemiology:
Pipelines and software packages used/created automatically generate corresponding protocols
Content of the metadata: Raw data sources, processing steps, storage locations, place and time of analysis, name or abbreviation of the processors.
Storage is in plain text format if possible

Checklist Creation of Metadata

- Data without metadata are worthless and do not comply with good scientific practice - how and according to which metadata standards (e.g. ISO19115, Dublin Core) are metadata (project, institution, PI, topic, time frame, rights to data use, method, variable definition, vocabulary, measurement units, assumptions, software, device, experiments, version information, data formats) created to enable traceability and future use of the data?
- What specific purposes does the metadata system serve (e.g. making data traceable, making data visible, maintaining data)?

- Does the metadata express who measured or modelled what, when and with what?
- How is timely creation of metadata enabled? Is metadata generated automatically during data collection?
- How are the individual analysis steps and intermediate results documented in the metadata?
- In which format is the metadata stored?
- What are the hardware and software requirements for processing the metadata?

Example metadata KFO:
Metadata standards: SPSS, R or SAS syntaxes with precise documentation
Content of the metadata: Raw data sources, processing steps, storage locations
Creation of metadata: see standards

Data protection checklist for data collection and analysis

- Which data are subject to data protection requirements?
- How are technical measures appropriate to the size of the project implemented to secure personal data in line with data protection (e.g. access passwords, different write and read rights for project staff)?
- How is the protection of the identity of test persons ensured in the course of the project (the pseudonymisation of data must not be removed from researchers)?
- How is the technical and organisational protection of sensitive human data ensured?
- Is there a data use contract in which the data users assure not to attempt to identify persons whose data they have received (how is this to be legally classified or does this concern the risk assessment)?

Example of data protection SniP:
Trial participants receive a pseudo code (in new projects usually a random 5-digit combination of letters and numbers, so-called SNiP code) upon inclusion in the trial, which represents them in all stored data as well as paper documents. Only the consent forms and the recruitment database contain clear names, dates of birth, etc. and are stored separately from other paper documents (see above) or accessible only to authenticated users. Electronic storage under the project data takes place only in pseudonymised form.

Example Data protection Genetic epidemiology:
Personal and clinical data are strictly separated. Similarly, clinical and genetic data are pseudonymised several times in stages: Upon inclusion in studies, upon inclusion in DB with clinical data, upon sample receipt in the laboratory, upon sending the data externally after signing a data transfer agreement. Subject questionnaires with personal data are stored in a locked steel cabinet.

4.3 Selection and storage of data after project completion

Research data are retained beyond the completion of the actual project for a variety of reasons. After securing a working copy and interim results during the active research process, the data on which a publication is based is stored for the purpose of conforming to good scientific practice, the subsequent use of the data for later research questions or the fulfilment of legal requirements in the case of clinical trials. The selection of data to be retained must be done in a transparent and comprehensible manner.

Checklist Selection of data for storage after project completion

- How long should the data be kept after the project is completed? Are there requirements from the funder or legal requirements, e.g. for clinical trials?
- Which data should be archived? Which data are needed to make the results comprehensible? Which data are the basis of a publication and thus, according to the rules of good scientific practice, to be archived for ten years at the place of their creation?
- For whom is the data stored (re-use in own working group, publication in repository)?
- Who decides on the selection of data to be archived?
- What criteria are used to select the data?
- At what point does the selection take place?
- Which tools (e.g. software) are used for the selection?
- Do data have to be deleted after a certain period of time? What is the procedure if data should no longer be kept?
- Do the data meet specified quality requirements for the planned subsequent use?
- Is the data comprehensibly prepared and usable for third parties?
- How is the technical usability of the data solved?
- At what level of aggregation does it make sense to archive the data?
- What are the possibilities and requirements for subsequent use? Is the data not efficiently reproducible?
- In which archive or (possibly public) repository can the data be stored in the long term and, if necessary, published? What are the costs involved?

Example SniP Handling data after publication:
Data analyses that have been incorporated into publications are stored separately together with the manuscript and the associated documentation (including scripts, depending on the procedure) in another storage location.

Example Handling of data after publication Genetic epidemiology:

Data analyses that have been incorporated into publications are stored separately together with the manuscript and the associated documentation (including scripts, depending on the procedure) in another storage location.

Under German law, genetic individual data may not be stored externally in public repositories. Therefore this is not done

Data protection checklist for data retention after project completion

- Where possible, have modifications been made to the data set to reduce the risk of identification (e.g. substitution of certain dates, obscuring identifiers in image data, etc.)?

Example of data protection for orthodontics:
Modifications I Encryption to protect against re-identification: Access restriction within the project database

Example Data protection Genetic epidemiology:

Additional pseudonymisation level for data transfer

4.4 Ingest: Feeding the data into the long-term archive

Ingest refers to the process of feeding data into an archive or other repository for long-term storage after the end of the project, e.g. a specific folder or drive. This includes all processes that have to be realised between the approval for ingest and the finished ingest into a storage location, i.e. the preparation of the data, its transport as well as the actual ingest. Ingest does not necessarily have to refer to the feeding into an archive, but can also mean the transfer of the data into a separate folder.

Checklist Ingest

- In which project phase is the data transferred to the long-term memory?
- Is all the necessary metadata fully available to enable subsequent use of the data?
- Has the data been subjected to technical validation (format validation)?
- Have the data been checked for factual accuracy?
- In which way should the transport to the archive take place (network, sending of data carriers)?

- Does the use of archiving standards (e.g. ISAD-G[15]) make sense?
- Have as many processes as possible been automated?
- Is specific preparation of the data necessary?

Checklist Data Protection Long-Term Archiving

- Is there sensitive data under data protection law that must be transferred in encrypted form?

4.5 Storage of data after the end of the project

The long-term storage of research data is one of the most fundamental tasks in research data management. The most important factors influencing storage are the size of the data sets, their number and the frequency of access to the data sets. Essential requirements for the storage of research data are the integrity and reliability of the storage, the confidentiality as well as the availability and usability of the data.

Checklist for saving data after the end of the project

- Who is responsible for storing the data after the project has ended?
- What technologies are used to store the data?
- Where is the data stored?
- Are additional backup copies made and checked regularly?
- What is the expected amount of data (per year, over the total duration of the project)?
- What forms of access are predictable? How frequently and intensively is the data accessed?
- Are there special requirements due to special services for data use (e.g. computing capacity for data extraction)?

Example storage KFO:
Storage location:
oDiagnostic and personal data: own database server, backups are encrypted on protected area of the IT network disks.
oLab and fMRI data pseudonymised on IT network disks

Forms of access:
oDiagnostic and personal data: personalised database accounts
oLab and fMRI data: Access authorisation management through Windows Active Directory

Checklist Data protection Storage of data after the end of the project

[15] http://www.icacds.org.uk/eng/standards.htm

- Is there sensitive data under data protection law whose access must be secured? How is this realised?
- Has the data been encrypted?

Example of data protection for orthodontics:
Access security for sensitive data: personalised database accounts

4.6 Conservation measures during long-term storage

In order to keep digital data usable in the long term, a number of measures are necessary. This applies in particular to technical and intellectual reusability. Examples are new data or file formats, new interfaces or new scientific standards or working methods. It is important to first establish that there are relevant changes in the technology landscape or the scientific community. Typical measures to ensure technical reusability include adapting the software environment (e.g. porting the software, supporting additional formats) or adapting the data (format migration, conversion) to new software environments. Examples of preserving intellectual reusability include updating the contextual information necessary to understand the data, e.g. by referring to new terminologies or procedures.

Checklist conservation measures Data

- What should be possible to realise with the stored data in the future?
- Have the technologies used and dependencies on other datasets or services been documented?
- Are the usage goals and requirements for the use of the data documented?
- Is it regularly checked whether the requirements and the available technologies or dependencies have changed?
- How are the conservation measures documented so that they are comprehensible to outsiders?

Example maintenance measures KFO:
Future purpose of data use: see declaration of contact (personal data)
Documentation of the technology used & dependencies:
oVirtual web server (Linux)
o MariaDB MySQL database
oOwn PHP-based web frontend

Example Conservation measures Genetic epidemiology:

Future data use purpose: see declaration of consent

Conservation measures: Automated log of all changes

Documentation of the technology used & dependencies:

- Department internal section of the ZI Wiki page
- Documentation folder on the projects and project processes
- SOPs

Data protection checklist for conservation measures during long-term storage

- Is there sensitive data under data protection law whose access is restricted?

Example of data protection for orthodontics:
Access protection for sensitive data: encrypted backup

4.7 Access and subsequent use of data after the end of the project

Besides the pure technical preservation of data, enabling the use of stored data is one of the most important tasks of archives. The most important goal here is to provide authorised users with access to the data and to prevent unauthorised persons from accessing it.

Checklist after-use of the data

- May the research data be used by others within or outside the project, working group or institution?
- Are there reasons for not releasing the data to other people (data protection, confidentiality)?
- Which target groups will be interested in the data, what should they be able to realise with the data?
- Has sufficient contextual information been provided to allow re-use of the data?
- How can the data be found?
- Are the data subject to licensing conditions? Can e.g. Creative Commons licences be used?
- Do future users of the data have to adhere to certain restrictions on use (e.g. no commercial research)?

Example of after-use KFO:

- Authorisation: Project manager
- Context information of the data: see metadata
- Findability: PSM data structure (project folder)

Example of after-use Genetic Epidemiology:

- Authorisation: Project manager
- Context information of the data: see metadata
- Findability: Data structure of the genetic epidemiology (project folder)

Data protection checklist for subsequent use of data

- How is the transfer of pseudonymised data regulated by contracts that specify the use in detail?

5 Research Data Management: Use Case Imaging Data at the ZI

5.1 Unidentification of Person-Identifying Features in Structural and Functional MR Images

Depending on the recording modality, a 3D reconstruction of a person's face can be produced from MR images (DICOM, NIFTI). Most imaging data analysis packages offer such functions (e.g. fslview from FSL), but a number of DICOM image viewers also allow this procedure. Before sharing with research partners and/or for availability via OpenAccess databases, it is therefore necessary to ensure (it is mandatory) that these no longer contain the information that these procedures allow. This usually means that the face must be removed. The following procedures are currently known for this purpose:

1. **SPM** contains a 'defacing' procedure since version 12.
2. **FSL**: in the analysis pipeline, 'bet' (brain extraction) is fixed and can be used to extract the brain completely.
3. **FreeSurfer** also contains a defacing procedure (mri_deface).

When using these tools, the following potential problems should be noted:

- None of the programmes works directly on DICOM data. These remain unchanged in any case. DICOM data must be converted beforehand (NIFTI) with the exception of FreeSurfer.
- In individual cases it has to be checked whether the output is still useful for further analyses. For example, segmentation in SPM12 with a BET image of FSL - at least with the default settings - fails.

At this stage, the use of SPM is recommended, but not mandated, for facial defacing in imaging data at ZI: The defacing procedure is robust and further analysis seems unproblematic. Moreover, SPM is the only common open source analysis software available for all commonly used computing platforms, while the others mentioned above are only available for posix operating systems (Linux, MacOS).

5.2 Removal of person-identifying features in MR data in DICOM format

As a rule, DICOM files created during functional imaging contain personal identifying features. Even if no plain text names are entered in the patient database of the tomographs during registration, at least the date of birth, body weight, height and gender must be entered (these are mandatory fields, which are also partly required during the measurement for correct SAR / pulse calculation).

Therefore, at least the following tags must be removed from the DICOM files or overwritten senselessly:

- Patient's Birth Date
- Patient's Age
- Patient's Size
- Patient's Weight
- Patient's Sex

It should be noted at this point, however, that these data must be made available for evaluation in some anonymised form in the case of OpenAccess data, especially also in the case of NIFTI data. Of course, other tags can also be removed (sample code can be requested [snip_ananon_dicom_fields.py]). The DICOM format contains, for example, the names of persons who performed the measurement, the measurement date, etc.

In any case, after removing the information, check whether the DICOM data can still be converted/read correctly. If this is not the case, they are de facto unusable and any further storage is unnecessary.

A defacing procedure cannot be performed directly on DICOM data. ***Structural data in DICOM format can therefore never be considered anonymised.*** Moreover, they contain information about the time of the survey, the institution and the persons performing the survey. Therefore, DICOM data should generally not be made available for Open Access. Exceptions to this may exist in the case of research collaborations that receive legal protection through other mechanisms (e.g. cooperation agreements, joint research funding).

Instead, it is recommended that the data be passed on in NIFTI format. This also contains no more information that allows conclusions to be drawn about the data collection of any kind, and defacing can take place.

DICOM data require special protection due to the aspects mentioned above. Storage should be done in such a way that the group of persons with access authorisation can be restricted or is known.

5.3 Storage of pseudonymised imaging data in the FI

When storing pseudonymised imaging data, particular care should be taken to ensure that the storage locations or datasets are subject to appropriate access control and that sustainable storage can take place.

On the one hand, this means that the group of persons with access to the storage media can be/is restricted and that these persons are known by name. This means that the storage media cannot be arbitrarily removed or simply lost. IT server rooms are ideal for this purpose, as they are protected against unauthorised access by special measures. This also means that the use of mobile, external storage media (e.g. USB sticks, external hard drives) or storage on desktop systems in staff offices is not recommended.

On the other hand, access to the data should also be controlled, i.e. it should be known at all times who has read and/or write access to data stocks. Here, too, centrally managed storage solutions offer the possibility to allow or deny access (e.g. read only) based on the access permissions of the file systems.

Centralised solutions also lend themselves to the long-term storage of data, as better backup strategies or data replication can be implemented in these systems and possible hardware defects can be detected more quickly. Here, too, external storage media are at a clear disadvantage, because a defect is often only detected when an attempt is made to access the data. Self-burned DVDs, for example, only have a shelf life of a few years, so they would have to be regularly renewed or checked, and are therefore not an economical or secure solution.

In any case, the following should still be considered when storing on portable media: If it is really unavoidable to do so, then the encryption of files or the entire media should be considered. In any case, it is important to realise that 'deleting' files in such a way that they cannot be recovered even with technically (somewhat) more complex means is not as easy as one might think. Storage media that have been used to store data must be disposed of in any case in accordance with data protection regulations and this is done by the IT department.

5.4 Exchange of pseudonymised data between researchalso outside the ZI

When exchanging data, especially from imaging, care should be taken to ensure that it cannot be lost or copied without authorisation. This means in particular that transmission methods are to be preferred whose administrative control does not leave the FI or always remains with one of the institutions involved, whereby users, on the other hand, must authenticate themselves and the communication is encrypted. The following solutions are possible (without claiming to be exhaustive):

1. Database server with systems specially designed for this purpose (e.g. xnat).

2. Encrypted transmission via sftp.

On the other hand, sending data media by post is useless. Although this can be insured, it does not protect against access by third parties in the event of loss. Only in exceptional cases should external storage media be used for transport. If it cannot be avoided, at least the medium should be encrypted. There are appropriate methods for this for every operating system (e.g. Bitlocker for Windows) and only USB mass storage devices, i.e. hard disks or sticks, come into question. Appropriate recommendations for the implementation of such encryption can be obtained from the IT department of the ZI.

6 Storage offered by the IT department at the ZI

Currently, the IT department offers a central storage solution for storing research data at the ZI. The storage solution is flexibly expandable, i.e. the maximum capacity of the system can be expanded as desired by adding further individual hard drives up to a certain upper limit. This offers departments or work groups at the ZI the opportunity to integrate their own hard disks into the central storage system if additional capacity is required. In this way, these can also be managed centrally by the IT department, but are only available to the respective department or working group. The maximum capacity is currently 512 TB (as of Q4 2017).

The stored data is backed up by creating snapshots every hour and mirroring the data to an identicalsecond system. This second system is located in another so-called fire section in the Central Institute, so that in the event of a possible fire disaster, not both systems would be affected. All IT server rooms are secured by personalised access control. User authentication takes place via the central authentication system (Active Directory), access authorisation via centrally assigned group authorisations.

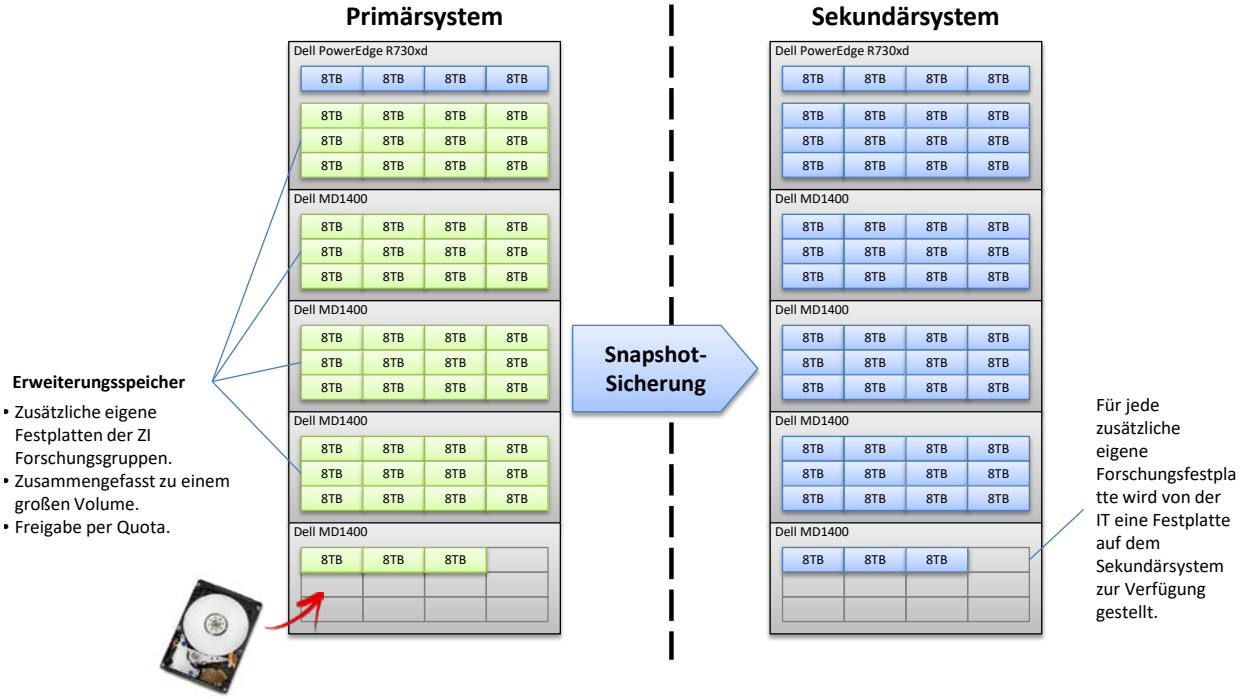

Fig. 3: IT infrastructure for research data storage at the ZI

7 Summary

Research data are an essential and valuable element of the research process and serve, among other things, the traceability of published results. Since research data is becoming increasingly extensive, especially due to the use of ever more modern research equipment, a number of initiatives, guidelines and demands for the handling of research data during and after the research process have emerged both nationally and internationally.

This guideline is intended as a "living document" which, on the basis of existing guidelines on research data management, presents a process description of the life cycle of research data that integrates the specific topic of data protection for research with human data. Experience-based further development will take place in cooperation with the group of decentralised data protection coordinators at the ZI.

Bibliography and recommended reading

Bertelmann, R., Gebauer, P., Hasler, P., Kirchner, I., Peters-Kottig, W., Razum, M, Recker, A., Ulbricht, D., Van Gasselt, S., (2014). Getting started with research data management in the geosciences. http://gfzpublic.gfz-potsdam.de/pubman/item/escidoc:749901:8/component/escidoc:749904/EWIG-Brosch%C3%BCre.pdf DOI: 10.2312/lis.14.01

DFG (2009). Recommendations for the secure storage and provision of digital primary research data. http://www.dfg.de/download/pdf/foerderung/programme/lis/ua_inf_empfehlungen_200901.pdf

DFG (2015). Guidelines on handling research data. http://www.dfg.de/download/pdf/foerderung/antragstellung/forschungsdaten/richtlinien_forschungsdaten.pdf

'Editorial: Data's Shameful Neglect' (10 September 2009) in Nature 461, p. 145, doi:10.1038/461145a. Published online 9 September 2009; corrected 23 September 2009

EU (2016). Guidelines on FAIR Data Management in Horizon 2020. http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

Harnischmacher, U., Ihle, P., Berger, B., Goebel, J., Scheller, J. (2006). Patient consent checklist and guide. Basics and guidance for clinical research. Schriftenreihe der Telematikplattform für Medizinische Forschungsnetze, Band 3, Berlin.

Helmholtz - Open Science Working Group (2016). Making the resource information more usable! Position paper on the handling of research data in the Helmholtz Association. https://www.helmholtz.de/fileadmin/user_upload/01_forschung/Open_Access/DE_AKOS_TG-Forschungsdatenleitlinie_Positionspapier.pdf

Jensen, U. (2012). Guidelines for research data management. Social science survey data. GESIS-Technical Reports 2012/07. http://www.gesis.org/fileadmin/upload/forschung/publikationen/gesis_reihen/gesis_methodenberichte/2012/TechnicalReport_2012-07.pdf

Ludwig, J., Enke, H. (Eds.) (2013). Guide to Research Data Management. Handouts from the WissGrid project. http://www.wissgrid.de/publikationen/Leitfaden_Data-Management-WissGrid.pdf

Pommerening, K., Drepper, J., Helbing, K., Ganslandt, T. (2014). Guide to data protection in medical research projects. Generic solutions of the TMF 2.0. Publication series of the TMF - Technology and Methods Platform for Networked Medical Research e.V., Vol. 11, Berlin

Schönbrodt, F., Gollwitzer, M., Abele-Brehm, A. (On behalf of the DGPs Board) (2016). Dealing with Research Data in Psychology: Concretisation of the DFG Guidelines https://www.dgps.de/fileadmin/documents/Empfehlungen/Richtlinien_zum_Umgang_mit_Forschungsdaten_20160929.pdf

University of Heidelberg (2014). Research Data Policy. http://www.uni-heidelberg.de/universitaet/profil/researchdata/

University of Edinburgh (2014). MANTRA Research Data Management Training. http://datalib.edina.ac.uk/mantra/researchstudent.html

Wilkinson, M. D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L.B., Bourne, P.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., Gonzalez-Beltran, A., Gray, A.J.G., Groth, P., Goble, C., Grethe, J.S., Heringa, J., 't Hoen, P.A.C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S.J., Martone, M.E., Mons, A., Packer, A.L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes E., Sengstag, T., Slater, T., Strawn, G., Swertz, M.A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data 3:160018 doi: 10.1038/sdata.2016.18 (2016). http://www.nature.com/articles/sdata201618