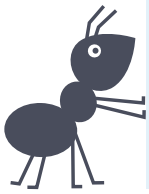
9a | Working Paper 2024

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

Long-term archiving of research data

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

Ute Hoffstätter, Anne Weber



September 2024

www.konsortswd.de

Long-term archiving of research data

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

Ute Hoffstätter[1], **Anne Weber**[1]

September 2024

https://doi.org/10.5281/zenodo.13692961

[1] German Centre for Higher Education Research and Science Studies (DZHW)

Note: This document has been translated from German (https://doi.org/10.5281/zenodo.10418834), therefore some of the references are only available in German. It focuses on research data infrastructure in Germany.

Abstract

This introductory guideline is intended to support (prospective) research data centres (RDCs) in the areas of data acquisition and long-term archiving. Based on the OAIS reference model, the minimum requirements for an archive are presented. These are applied to the terminology and processes in research data centres and enriched with references and examples from practice.

Keywords: Long-term archiving (LTA), OAIS (Open Archival Information System), research data centres (RDCs), data acquisition

Table of contents

1. Introduction[1]

The core tasks of a research data centre (RDC) are the archiving of data and data access for researchers. The German Data Forum (Rat für Sozial- und Wirtschaftsdaten [RatSWD] 2022, p. 7) states that in the RDCs, data is archived and made available to the scientific community via various modes of access in compliance with data protection regulations. Following the guideline on the core task of data access (Hoffstätter/Linne 2022), the focus here is on the area of archiving. Long-term archiving of this data and its metadata[2] is necessary for permanently secure data provision. As Altenhöner/Oellers (2012, p. 11) explain, more and more data of potential interest for secondary scientific use are available in digital and therefore easily storable and transferable form. Thus, sustainable long-term archiving concepts are required to ensure that these valuable data are not lost.

Long-term archiving is more than just backing up files over a certain period of time[3] (Liegmann/Neuroth 2010, Kap.1:3). This means that purely physical storage of the files is the basis, but not sufficient. Long-term archiving is the responsible development of strategies that can cope with the constant change of the digital world (Schwens/Liegmann 2011, p. 567). The aim is the long-term preservation and the technical and content-related interpretability of the digital resources, in this case the data and metadata (Liegmann/Neuroth 2010, Kap.1:3), for which important steps must be taken at the data acquisition stage.

This paper provides an overview of the particularly relevant aspects to be considered in the long-term archiving of research data, starting with data acquisition. In order to structure the existing requirements and use a common vocabulary, the Open Archival Information System (OAIS) reference model is used as a conceptual framework and its central elements and minimum requirements for an archive are presented (2 Open Archival Information System (OAIS) reference model). Based on this, assistance is then provided that can support RDCs in the area of social, behavioural, educational and economic data in fulfilling the central requirements set out in the model with regard to long-term archiving (3 Fulfilment of archiving requirements in RDCs for social, behavioural, educational and economic data). Finally, there is a brief summary and an outlook on other key elements that are relevant for functioning long-term archiving (4 Summary/Outlook).[4]

[1] Many thanks to Daniel Buck, Dr Jonas Recker and Dr designate Henrike Schmidtchen for their expert and constructive review comments.
All links contained in the document are as of 09.04.2024.

[2] Metadata is information that describes data and thus makes it understandable and interpretable.

[3] In the social sciences and other disciplines, a retention period of at least ten years for research data has become established as standard (see, for example, DFG Code, Guideline 17 https://wissenschaftliche-integritaet.de/kodex/archivierung/); however, the aim should be to retain data beyond this period.

[4] The explanations in this guideline focus on quantitative data but are kept as generic as possible so that they can largely also apply to other types of data. For qualitative data, see in particular QualidataNet (https://www.qualidatanet.com).

2. Open Archival Information System (OAIS) reference model

The Open Archival Information System (OAIS) reference model was developed jointly by various space organisations in the late 1990s (The Consultative Committee For Space Data Systems 2012). It is a very generic model that defines the elements, processes, tasks and responsibilities of long-term archives and makes important terminological specifications (OAIS-Übersetzung/Terminologie 2013, p. 1; Schumann 2012, p. 40). "An OAIS is an Archive, consisting of an organization, which may be part of a larger organization, of people and systems that has accepted the responsibility to preserve information and make it available for a Designated Community." (The Consultative Committee For Space Data Systems 2012, 1:1). The model itself is very extensive. Only those concepts that are centrally relevant to the focus of the guideline as an introduction to data acquisition and long-term archiving are presented here.

2.1. Central elements

The OAIS reference model (OAIS translation/terminology (2013) (see figure 1) contains:

- three **roles** (green on the left and right in the illustration and without colour coding at the bottom): Producers, management, consumers
- three **information packages** (blue in the illustration): SIP, AIP, DIP
- six **functional entities** (orange in the illustration): ingest, archival storage, data management, administration, preservation planning, access

Producers can transfer information packages to the archive in an OAIS. The management determines the general strategy of the archive as to how these information packages are processed in the functional entities. Consumers can then find and purchase information packages they are interested in.

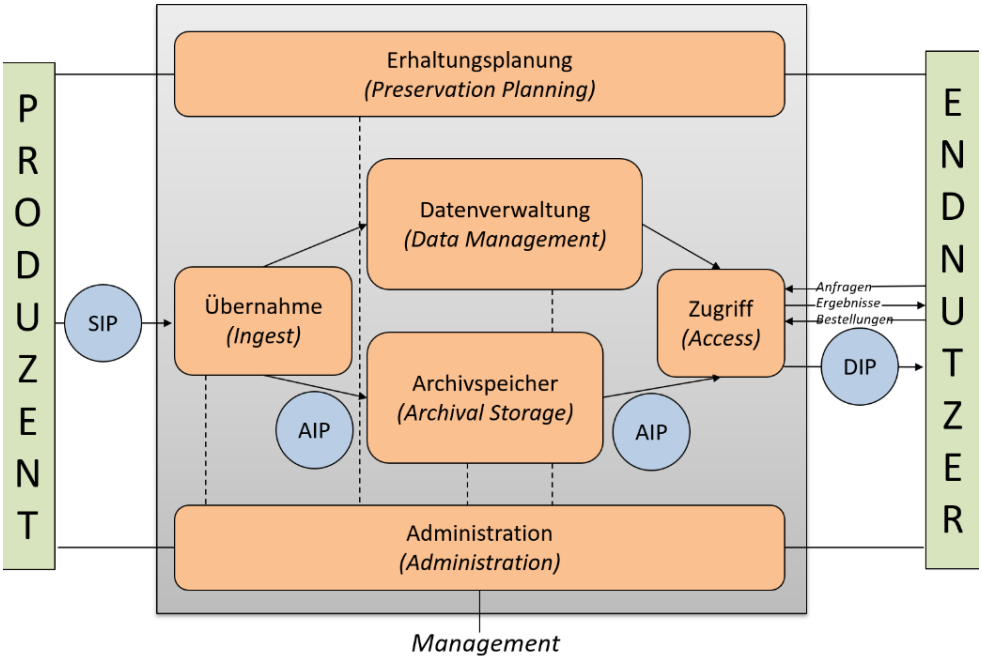


Figure 1: Environment model: Overview of the central elements of the OAIS reference model: Roles, information packages and functional entities (illustration based on OAIS-Übersetzung/Terminologie 2013).

These various elements of the OAIS reference model are described in more detail below.

2.1.1. Roles

The central roles in the OAIS reference model are:

- **Producers:** Groups of people who or systems that produce the information to be archived and transfer it to the OAIS for archiving
- **Management:** Group of people who are responsible for the OAIS in terms of its objectives (either as an independent archive or as part of a larger organisation). The group defines the general strategy of the archive as to how the transferred information is processed.
- **Consumers:** Groups of people who or systems that want to access and use archived information. With regard to the consumers, the OAIS determines a so-called designated community, which will use the archived information and to which a certain knowledge base (e.g. a certain specialist background) is attributed.

2.1.2. Information packages

In the OAIS reference model, three information packages[5] are defined, which are processed one after the other with different objectives.

- The **Submission Information Package (SIP)** is the version supplied by producers to the OAIS. The exact content is determined by the archive and is usually regulated by a submission agreement (e.g. contract).
- The **Archival Information Package (AIP)** is the version that is permanently stored and therefore the central object of interest of the archive. It is usually enriched in the archive by employees with information that is necessary for long-term archiving and preservation, such as metadata like the file format, creation date, persistent identifiers (e.g. DOI) (Schrimpf 2012, p. 55). The distinction to the SIP is made because not all of the producers' information is necessarily relevant to the archive and the files must be enriched with further information before archiving (Lavoie 2014, p. 14). At the same time, the content of the SIP can also be part of the AIP in unchanged form.
- The **Dissemination Information Package (DIP)** is the version that is provided to consumers. This package is generated from the AIP and may differ from it because the DIP only contains a subset of information (e.g. for data protection reasons) or has a different file format. The metadata provided for consumers also differs from the AIP, as the technical metadata, for example, is not relevant to them (Schrimpf 2012, p. 55). As a rule, this is the latest version that can be used with common systems and software.

2.1.3. Functional entities

The OAIS reference model defines six functional entities of an OAIS: The process of transferring the SIP to the OAIS is called **ingest**. In the archive, the transferred object is then enriched with further information to make it permanently comprehensible and technically available. This enriched object is transferred to the **archival storage** as an AIP. Long-term storage of the AIP is ensured in the archival storage. This includes the monitoring of inventories, integrity checks, error control and a process for the emergency recovery of files (OAIS-Übersetzung/Terminologie 2012, p. 32). When the AIP is submitted to the archival storage, the package descriptions (e.g. title, authorship,...) are transferred from the AIP to the **data management** functional entity. This is where employees manage the metadata of the archive holdings, i.e. the information that describes the archive holdings. The **preservation planning** functional entity develops holistic preservation strategies (such as migration and emulation) for the archive holdings. Among other things, technological progress and the needs of the intended designated community are monitored, and specific maintenance measures are planned. Preservation planning is comprehensive across the processes of transfer (ingest), data management and archive storage (Schrimpf 2012, p. 57). The **administration** functional entity provides services for managing the entire archive, such as a template for a Submission Agreement. This is where, for example, submission agreements are negotiated with the

[5] By information package, we mean an abstract package, i.e. the files themselves do not all have to be bundled in one place. For example, the metadata for the data, linked via an ID, can be stored in a separate database.

producers, archive operations are monitored, and standards, workflows and policies are defined and further developed (Schrimpf 2012, p. 57). A DIP is generated on the basis of the AIP for **access** to the objects by consumers. Consumers can research these DIPs and use the information stored in the data management system (e.g. title, keywords, etc.) to find and order or request them (e.g. via a data catalogue). These DIPs are then made available to the consumers.

2.2. OAIS mandatory responsibilities

Following on from the OAIS reference model described above, there are six mandatory responsibilities that must be met in order to operate an OAIS (OAIS-Übersetzung/Terminologie 2013, p. 28). The OAIS should:

1. **Negotiate for and accept appropriate information** from information Producers.
2. Obtain sufficient **control** of the information provided to the level needed to ensure Long Term Preservation.
3. Determine, either by itself or in conjunction with other parties, which communities should become the **Designated Community** and, therefore, should be able to understand the information provided, thereby defining its Knowledge Base.
4. Ensure that the **information** to be preserved is **Independently Understandable** to the Designated Community. In particular, the Designated Community should be able to understand the information without needing special resources such as the assistance of the experts who produced the information.
5. Follow **documented policies and procedures** which ensure that the information is preserved against all reasonable contingencies, including the demise of the Archive, ensuring that it is never deleted unless allowed as part of an approved strategy. There should be no ad-hoc deletions.
6. Make the preserved information available to the Designated Community and enable the information to be **disseminated** as copies of, or as traceable to, the original submitted Data Objects with evidence supporting its Authenticity.

3. Fulfilment of archiving requirements in RDCs for social, behavioural, educational and economic data

Research data centres for the areas of social, behavioural, educational and economic data can in principle be directly linked to the general considerations regarding an OAIS presented in the previous chapter: (Internal or external) data providers (producers) submit files (digital objects), e.g. research data and documentation material, to an RDC (OAIS). A data submission agreement is concluded between the data providers (producers) and the RDC (OAIS). The files and information provided (Submission Information Package – SIP) are received and processed in the data acquisition area (ingest). They are checked there, modified if necessary (e.g. anonymised) and enriched with information (in particular metadata). These enriched files and information (Archival Information Package – AIP) are sent for long-term storage (archival

storage). In the metadata catalogue (data management), the descriptive metadata for the data is made searchable and orderable for the community or communities of data users (designated community). The data that data users (consumers) can order (Dissemination Information Package – DIP) may be a modified version of this (e.g. because not all variables may be made available). In the data access area, the data users' requests are checked and the data (Dissemination Information Package – DIP) is then made available to the data users via the modes of access available in the RDC (OAIS) (e.g. download, secure remote access or guest research workstation). The RDC management (Management) designs the standards and policies of the RDC (OAIS) (e.g. contracts, monitoring of infrastructure, organisation of work processes) and their maintenance. Overall, the RDC (OAIS) must fulfil the six defined minimum requirements for an OAIS.

The following section provides guidance on how the generally formulated minimum requirements of RDCs in the areas of social, behavioural, educational and economic data can be met in concrete terms. The focus here is on the first five requirements, as these relate directly to the core archiving task, while the sixth requirement is aimed more at the core task of access, for which a guideline has already been published (Hoffstätter/Linne 2022).

The exact design of the fulfilment of the requirements always depends on the respective specifics (data types, number of digital objects, resources, mission, resources, etc.) of an RDC. Nevertheless, it is possible to identify particularly important and generally valid aspects and give examples that can be a helpful starting point for RDCs and, if necessary, help them to identify and fill their own gaps.

In the following, the terminology of the RDC community is used, including English terms also used in the RDC community, such as SIP, DIP and AIP or designated community.

3.1. Negotiate for and accept information

The first requirement relates directly to the process of data acquisition and its preparation, which in practice is often referred to as pre-ingest. This is usually an iterative process between RDCs and data providers, which is why the points explained in more detail below should not be seen as a clear-cut and strictly linear sequence in practice. A transparent presentation of services and costs helps RDCs and data providers to plan reliably.[6]

3.1.1. Scope and selection/prioritisation

If researchers wish to transfer their data to an RDC, they must first check whether the data falls within the scope of the RDC. In order to make the process as transparent as possible for data providers, an RDC should define its scope in advance. A collection policy is often used for this purpose. This document outlines the principles of the existence and (planned) development of the RDC. In this way, the inventory of an RDC can be built up and further developed in a coherent manner.

[6] See for example ZPID: https://rdc-psychology.org/de/service-katalog-fdm or GESIS: https://www.gesis.org/datenservices/daten-teilen

The documents of the following institutions can serve as an example of a collection policy:

- UKDS[7]: https://ukdataservice.ac.uk/app/uploads/cd234-collections-appraisal.pdf
- ICPSR[8]: https://www.icpsr.umich.edu/web/pages/datamanagement/policies/colldev.html
- GESIS[9]: https://www.gesis.org/fileadmin/upload/Datenservices/Collection_Policy/2023-05-25_Collection_policy_dt.pdf
- VerbundFDB[10]: https://www.forschungsdaten-bildung.de/collectionpolicy
- RDC Education[11]: https://www.fdz-bildung.de/collection-policy-fdz
- AUSSDA[12]: https://aussda.at/fileadmin/user_upload/p_aussda/Documents/AUSSDA_Data_Collection_Policy.pdf

It may also be necessary to weigh up the inclusion of different data that meet all the criteria of the scope and make an appraisal if, for example, not all data can be included at a certain point in time for resource reasons at the RDC. The appraisal of data in an RDC should be moral, consistent, transparent and comprehensible. The criteria and the process should therefore be clearly defined in advance. In principle, the objective of the institution (e.g. statutes, mission statement,...) and the designated community/communities as well as their expectations and needs should always be taken into account (Mauer 2016). Such a concept should therefore not be created by the RDC alone. The various stakeholders (data users, advisory board, scientific societies, research funders, research policy,...) should be involved in the development of the selection process. Whyte/Wilson (2010) provide guidance on what aspects should be considered when developing an approach to assessment and selection:

1. Relevance to mission
2. Scientific or historical value
3. Uniqueness
4. Potential for redistribution
5. Non-replicability
6. Economic case
7. Full documentation

[7] United Kingdom Data Service: https://ukdataservice.ac.uk/
[8] Inter-University Consortium for Political and Social Research: https://www.icpsr.umich.edu
[9] GESIS - Leibniz Institute for the Social Sciences: https://www.gesis.org/
[10] Research Data Education Network: https://www.forschungsdaten-bildung.de/
[11] RDC Education: https://www.fdz-bildung.de/home
[12] The Austrian Social Science Data Archive: https://aussda.at/

Weber/Piesche (2021) have also compiled a checklist as an aid to checking the archival value of research data (Weber/Piesche 2021, pp. 335-336, 346-347):

- Are there third-party requirements (funding organisations, data policies, guidelines of the research institution) that make it necessary to store the data for the long term?
- Do you have the necessary rights to use the data? Under what conditions do you own the data?
- Are the collected data unique and not reproducible or are the costs of reproduction higher than the costs of long-term storage?
- Is data collection unlikely to deliver better results as a result of technological progress?
- Is there a high level of interest in re-utilising the research data?
- Has the data not yet been fully scientifically analysed?
- Are the data characteristic or atypical for a research area or are they unique research results?
- Do the data possibly have a general or regional historical significance?
- Is the data quality good in terms of technology and content?
- Is descriptive metadata fully available or can it be generated?
- Can the necessary preservation metadata (reference, provenance, context and retention information as well as information on access rights) be provided?

In accordance with the minimum legal and documentary requirements, the assessment relates in particular to the uniqueness of the data and the potential benefit for research. Any time commitments for publication on the part of the data providers should also be taken into account in order to check and enable timely publication.

Examples[13]:

- ICPSR: https://www.icpsr.umich.edu/web/pages/datamanagement/lifecycle/selection.html
- UK Data Service: https://www.ukdataservice.ac.uk/media/455175/cd234-collections-appraisal.pdf
- RDC at ZPID: https://rdc-psychology.org/de/datenauswahl
- IQB: https://www.iqb.hu-berlin.de/fdz/Datenuebergabe/Template_Ingest_.pdf
  (This checklist also contains points aimed at checking fulfilment of the minimum requirements on receipt of data, see chapters 3.1.2 and 3.1.3)

If it turns out that a dataset cannot be included or cannot be included promptly, reference can be made to other archiving options, for example to other RDCs or networks (see chapter 4) or,

[13] Further resources: *CESSDA Data Archiving Guide*: https://dag.cessda.eu/Chapter-3/3-Data-review-and-appraisal and DCC How-To-Guide: https://www.dcc.ac.uk/guidance/how-guides/appraise-select-data#5

if necessary, to self-service research data repositories (e.g. repositories such as GESIS Archiving Basis[14] or RADAR[15]).

3.1.2. Minimum requirements for data and metadata

To ensure data quality and compliance with legal requirements, an RDC should clearly define certain minimum requirements for the data and metadata that must be met for an ingest. These should be communicated transparently to the data providers – ideally during the application or planning of a project, including a data management plan[16] – so that nothing stands in the way of efficient and legally compliant data acquisition and secondary use. Legal, technical and content-related minimum requirements should be considered.

Examples of minimum requirements:

- GESIS: https://www.gesis.org/datenservices/ueber-die-datenservices/standards-und-workflows-datenservices/vorbereitung-datenuebergabe-pre-ingest
- AUSSDA: https://aussda.at/fileadmin/user_upload/p_aussda/Documents/DataDeposit-Guideline_SUF_v2_0.pdf
- IQB: https://www.iqb.hu-berlin.de/fdz/Datenuebergabe/Template_Ingest_.pdf

a) Legal

The fulfilment of the minimum legal requirements of an RDC for data acquisition determines the basic archivability of the data and metadata. It is particularly important that the agreements with all project participants (funding body, client, research subjects, etc.) are designed in such a way that there is no legal obstacle to the transfer of data to an RDC. As the legal aspects are often a challenge for data providers, it can be worthwhile for RDCs to provide assistance in this regard.[17]

If the research data is based on information provided by individuals (e.g. online survey, interview,...), data protection issues[18] often need to be clarified, as personal data is also collected. As a rule, data is collected on the basis of the informed consent of the research subjects (declaration of consent), legitimate interest or a special legal provision in the case of official data. An informed consent must be checked to see whether it complies with the legal provisions that applied to the data collection (e.g. the GDPR, BDSG or LDSG for personal data). In addition, secondary use of the data must not be ruled out. Other aspects, in particular ethical considerations, can also play a role, which must be taken into account for each dataset.

[14] https://data.gesis.org/sharing as a social science repository
[15] https://www.radar-service.eu/ as a generic repository
[16] for example https://forschungsdaten.info/themen/informieren-und-planen/datenmanagementplan/ or specifically for educational research https://www.forschungsdaten-bildung.de/stamp-nutzen as well as Netscher/Jensen (2019b, pp. 38–40) on planning research data management and Uwe Jensen (2012)
[17] For specific legal questions on data acquisition, see also the explanatory document by Schallaböck/ Kreutzer, et al. (2023b). Further resources for support can be found in particular on educational research at https://www.forschungsdaten-bildung.de and at https://rdm-compas.org/.
[18] See also in particular Watteler and Ebel (2019) (sections 4.3.1 and 4.3.2).

Informed consent must be checked for these legal and ethical aspects in order to protect the rights of the research subjects or, in the context of data protection, the data subjects. Information on the requirements for informed consent can be found at Verbund Forschungsdaten Bildung (2018) and at Meyermann/Porzelt (2019). At the same time, the interests of research must be taken into account, also by considering the privileges granted to it by data protection laws. The RDC must take both perspectives into account appropriately during data acquisition.

At this point, examples of the different ways in which anonymised and personal data can be reused for research purposes are given. In the case of data collection based on informed consent, the transfer of data to third parties (personal or anonymised) should at least not be explicitly excluded. If the re-use of anonymised data is not explicitly excluded, its re-use is legally permitted unless there are other reasons to the contrary. In terms of research ethics, a planned re-use should be communicated to the persons studied in the *i*nformed consent, even if only the anonymised data is re-used. Statements excluding the use of anonymised data, on the other hand, can be interpreted under contractual law as part of an informed consent and thus exclude re-use (Meyermann/Porzelt 2019, p. 17). The re-use of personal data, on the other hand, is initially only legally possible with a corresponding declaration in the informed consent (Art. 7 GDPR; in terms of data protection law, the GDPR only applies to personal data). At the same time, its secondary utilisation should be critically examined from a research ethics perspective. In addition, Art. 89 GDPR and Section 27 BDSG, which regulate guarantees and exceptions regarding the processing of personal data for scientific, historical research and archiving purposes in the public interest as well as statistical purposes, must be mentioned in particular for the research context. However, its application requires a thorough balancing of the interests of the research and those of the persons concerned or the research subjects.

The legal review often shows that data can only be passed on for archiving and/or secondary use for anonymised[19] data, so that corresponding anonymisation work is required during data preparation (see also section c).

Furthermore, it should be ensured that the submission of data and metadata does not violate the rights of third parties and that good scientific practice is observed, for example with regard to the citation of reused material.[20] For example, the data providers must ensure that any question wording, scales, item sets, etc. used in the data, the survey instrument or other documents are correctly cited and that any existing terms of use (or licence conditions) are complied with. In an academic context, these will often be limited to citation obligations. In exceptional cases, however, the use may also be accompanied by payment obligations or prohibitions on public dissemination and reproduction. This can be the case, for example, with survey instruments that are also used for psychological diagnostics. Due to the fact that

[19] On the anonymity of research data, see also explanatory document version 2.0.0 (question 1.1) (Schallaböck/Hoffstätter, et al. 2023)

[20] This also includes reused and transferred public domain data without copyrights and licences Lauber-Rönsberg (2021, pp. 93–96).

awareness in the community is only now starting to grow, it is advisable for the RDC to point this out to the data providers and support them. The assessment of whether copyrights exist for research data and/or accompanying metadata (e.g. documentation materials such as survey instruments) may vary from case to case.[21] The key to legally compliant data acquisition is *that* – if copyright-relevant elements are present – the corresponding rights of use are transferred to the RDC via a submission agreement (see section 3.2).

Other aspects should also be clarified in the submission agreement (see chapter 3.2), for example the modes of access. If the RDC offers various modes of access for data users, it should also be clarified in this context which modes of access should be selected for the data and metadata. Data with sensitive information (e.g. geographical information on respondents) should be made available via more restrictive modes of access, for example. An introductory guideline on how data can be made accessible can be found at Hoffstätter/Linne (2022).

b) Technical-formal

In addition to basic technical and formal minimum requirements, such as the completeness, integrity and absence from viruses of all transmitted files and the machine-readability of the data, the authorised file formats are particularly important. For long-term archiving and to maintain the interpretability and usability of the files, open formats should be used wherever possible, as these are more stable over time. The research data can be transferred by the data providers in file formats that are widely used in the community. The RDC should publish a list of accepted file formats.[22] It is essential that the RDC is familiar with the relevant file formats and can also transfer them to other formats (migration) or technical environments (emulation) if necessary. In addition, specifications for file naming can also[23] be useful.

c) Content

In order to be able to utilise and interpret the data in a meaningful and long-term manner, minimum content requirements are also necessary with regard to data preparation and metadata. The data preparation of quantitative survey data includes, for example, the assignment of correct and comprehensible variable names as well as variable and value labels, the coding of open information and missing values (missings), data checking (e.g.

[21] "In summary, it can be stated that it is problematic for RDM that the protectability of individual research data can generally only be assessed in individual cases and even then not with sufficient legal certainty," write Lauber-Rönsberg et al. (2018, p. 3) in an expert opinion on the legal framework conditions of research data management. See in particular the explanatory document by Schallaböck/ Kreutzer, et al. (2023a), in which legal questions relating to research data centres are answered.

[22] Examples and assistance can be found in AUSSDA (Butzlaff 2022, p. 8) GESIS (https://www.gesis.org/datenservices/ueber-die-datenservices/standards-und-workflows-datenservices/vorbereitung-datenuebergabe-pre-ingest), UKDS (https://ukdataservice.ac.uk/learning-hub/research-data-management/format-your-data/recommended-formats/), Education Research Data Network (https://www.forschungsdaten-bildung.de/dateiformate) and Weber and Piesche (2021, pp. 347–349).

[23] See, for example, on file organisation Recker/Brislinger (2019) or also https://www.forschungsdaten-bildung.de/dateien-benennen#Benennung-von-Dateien

plausibility/consistency check) and cleansing, the generation of new variables and anonymisation (e.g. by aggregating values or deleting variables).

Key metadata for the comprehensibility of the data content are, for example, the survey/measurement instrument (e.g. questionnaire, interview guide), a data and methods report (with information on, for example content/design of the study, population, sampling, implementation of the survey, response rate, data preparation steps carried out),[24] codebook/variable report (for quantitative surveys) and, if applicable, various structured descriptive metadata[25] (depending on the systems and metadata standards used, see section 3.4).

3.1.3. Data transmission and ingest inspection

The data and metadata can be transferred to the RDC – either collectively or iteratively – in various ways, e.g. on a data carrier (such as a USB stick or external hard drive), via a collaborative work platform or via a specially designed digital tool.[26] An RDC should clearly communicate the possible transmission paths to the data providers. It is essential to ensure that data protection is guaranteed at all times, for example with the help of specially secured services and encryption technologies.[27]

Once the data and metadata have been received by the RDC, they should be checked whether they fulfil all the criteria for data acquisition – based on the previously defined, RDC-specific minimum requirements for the data and metadata (see section 3.1.2). It is recommended that a checklist be drawn up, on the basis of which all relevant points can be checked one by one in a structured manner and the test results can be documented and understood in writing. The exact procedure for the incoming data check and the granularity at which it is carried out must also be defined specifically for each RDC.

Examples:

- RDC at the IQB
    - technical-formal: https://www.iqb.hu-berlin.de/fdz/Datenuebergabe/00_Checkliste_te.pdf
    - content: https://www.iqb.hu-berlin.de/fdz/Datenuebergabe/Template_Ingest_.pdf

[24] see for example Watteler (2010).

[25] Structured/standardised metadata are data with a uniform vocabulary, for example information on sampling according to the metadata standard of the Data Documentation Initiative (DDI) or with measurement concepts, such as the International Standard Classification of Education (ISCED 2011) for school types and school systems, (Netscher/Jensen 2019a, p. 40; Wallace 2001, p. 255). In contrast, semi-structured or unstructured metadata is based on continuous text and provides the necessary context in detail, e.g. in the form of data and method reports or questionnaires. The use of both types of metadata is common in the social sciences and related disciplines.

[26] Transmission via unencrypted email is not a secure means of transmission. External services are not recommended – "especially if their use is associated with data processing outside the European Union" (e.g. Dropbox or Onedrive) – according to RatSWD (2020, p. 23)

[27] See also, for example https://rdm-compas.org/articles/datenubernahme-2.

(this checklist also includes an assessment of the fit with the content scope and an assessment of the general re-use potential (see section 3.1.1)).
    - initial check of the data: (IQB - Institut zur Qualitätsentwicklung im Bildungswesen 2021a): https://www.iqb.hu-berlin.de/fdz/Datenuebergabe/Vorlage_AbkStudi.pdf
- GESIS:
    - https://www.gesis.org/datenservices/ueber-die-datenservices/standards-und-workflows-datenservices/datenuebergabe-ingest resp.
    - https://www.gesis.org/fileadmin/upload/Datenservices/Eingangskontrolle/Dateneingangskontrolle_n__GESIS.xlsx
- RDC Education: https://www.fdz-bildung.de/datenkuratierung
- GESIS Archiving BASIS: https://rdm-compas.org/articles/tipps-checklisten-auswahl-bewertung#accordion-beispiel-sowidatanet-checkliste-fur-die-dateneingangskontrolle-von-forschungsdaten3-%F0%9F%93%9D

The data is often corrected, or the metadata clarified or expanded in an iterative process with the data providers. If necessary, it should also be agreed whether and which revisions will be carried out by the RDC (see chapter 3.1.4).

3.1.4. Acquisition and further processing

If the submitted research data and its metadata are suitable for acquisition in an RDC, the corresponding digital objects must be transferred to the infrastructure of the RDC as a read-only SIP and secured technically (e.g. via checksum verification[28]) against (even unintentional) processing.[29] The storage capacity must be provided by the RDC. It is also important that the RDC has a clear folder and file structure and file naming system, according to which the transferred data and metadata are stored, as well as a functioning version control system.[30]

To this end, relevant elements of the SIP are first copied and form the basis of the AIP. The RDC defines in advance which elements of the SIP are included in the AIP. The composition of the AIP varies depending on the processing, documentation and publication objective of a study (Mauer 2012, p. 209). Data and metadata may also be edited[31], e.g. by anonymising[32], aggregating or making other changes to the data and metadata (e.g. adding the persistent identifier for publication).[33] Whether the RDC or the data providers take on these tasks should

[28] The checksum of a file is comparable to a "digital fingerprint" of the file. Every change, no matter how small, leads to a different checksum. See for example https://www.dpconline.org/handbook/technical-solutions-and-tools/fixity-and-checksums

[29] Tips on this can also be found at https://rdm-compas.org/articles/datenubernahme-2

[30] Tips on file organisation can be found at https://rdm-compas.org/articles/fdm-hinweise-forschungsdaten-organisieren, https://www.forschungsdaten-bildung.de/dateien-benennen and Recker/Brislinger (2019) (section 5.3).

[31] For data preparation and documentation, see also Brislinger/Moschner (2019) and https://rdm-compas.org/articles/aufbereitung-qualitativer-und-quantitativer-daten

[32] For anonymisation, see for example Ebel and Meyermann (2015) for quantitative data or Meyermann and Porzelt (2014) for qualitative data.

[33] For initial test steps and data preparation, see e.g. IQB - Institute for Quality Development in Education (2021a). See also examples at 3.1.2 c.

be negotiated and clearly communicated in the course of the initial review and, in particular in the case of legally relevant aspects (such as responsibility for anonymisation), should also be set out in the submission agreement (see chapter 3.2). As a rule, the SIP files are also enriched in the AIP by RDC employees – with the submission agreement, other legal documents (e.g. informed consent and its evaluation), further information (e.g. documentation of the appraisal, documentation of communication) and structured metadata that is important for long-term archiving (cf. chapter 3.4). All changes to the data and metadata that take place in the AIP are documented in a history file and also added to the AIP.

For example, Mauer (2012, p. 209) for an AIP at GESIS:

- SIP: Original objects (e.g. data records, documents) or copies thereof transferred by data providers
- data record versions prepared in the archive
- Preparation syntax that establishes the relationship between the original and archive version, as well as further documentation of the preparation
- metadata generated in the archive to describe the data and other technical and administrative metadata

For example, the RDC at the IQB (IQB - Institut zur Qualitätsentwicklung im Bildungswesen 2021b) describes the AIP as follows:

- original data (=SIP), also in formats with long-term availability
- provided and made available (research) data (=DIP), also in long-term available formats
- available metadata, also in formats with long-term availability
- all documentation and preparation steps (preparation syntaxes...)
- further documents (evaluation of data, contracts, correspondence, contracts, access concept)
- checksums per digital object

An even more specific example of the folder structure of an AIP can also be found at the RDC at the IQB (IQB - Institut zur Qualitätsentwicklung im Bildungswesen 2021b, 11f).

3.2. Obtain sufficient control for long-term preservation

The RDC should ensure that a submission agreement is concluded with the data providers in which the legal, technical and organisational aspects of data acquisition are regulated[34] and it is determined how the SIPs are designed, i.e. which digital objects are handed over in which formats. This is usually done by means of a data acquisition agreement[35] but sometimes also in the form of general terms and conditions or terms of use. Before the legally binding submission of the data, the data providers must clarify who formally owns the copyright and rights of use. Only the party holding the rights of use can legally execute the submission and

[34] See also, for example https://rdm-compas.org/articles/rechtliches-datenubertragung-erstellung-und-empfang
[35] also known as a data arching contract, data archiving agreement or similar.

thus transfer these very rights.[36] A harmonised model contract for data acquisition in an RDC and an explanatory document on various legal issues have been developed and published as part of KonsortSWD (Schallaböck/Kreutzer, et al. 2023b).[37]

It is particularly important to note that when archiving research data, the RDC is generally granted so-called simple, i.e. not exclusive, rights of use [einfache Nutzungsrechte]. If copyrights exist in the material provided, these remain with the authors (exception only death of the authors and inheritance) (§29 of the Copyright Act). Depending on the details of the contractual conditions, the RDC is generally only granted the simple rights of use to store, reproduce and pass on data and metadata to third parties and, in particular, the right to transform the digital objects (data, metadata, etc.) into other formats to ensure long-term usability.[38] In the case of transferred simple rights of use, the data providers are also able to terminate the contract. In this case, the RDC may no longer hold and use the data. The fundamental power to decide on the use of the data ultimately remains with the data providers. It should be emphasised that it is not only the research data that may potentially be subject to rights of use. In the case of survey instruments (e.g. when using psychological scales in questionnaires), any existing third-party claims in terms of rights of use (or licences) must also be checked (for legal aspects, see in particular section 3.1.2 a and the references there).

If the data was generated at the same institution where the RDC is located, it is usually not legally possible to conclude a data acquisition contract. However, this must be checked on a case-by-case basis.[39] Nevertheless, it should also be ensured here that the disclosure and publication of data and metadata does not lead to legal problems for the institution and/or is not in the interests of the data providers. Data providers' obligations to third parties, for example with regard to publication obligations and deadlines, should also be recorded in order to take these into account as far as possible in the RDC's planning. For example, it is conceivable to create a document that is signed by the data providers and whose contents must be taken into account by the RDC. This allows the respective roles and areas of responsibility to be defined.

3.3. Determine designated community

It is necessary for an RDC to determine its intended designated community, i.e. the expected consumers, because only in this way can the basic knowledge of the designated community be

[36] see Lauber-Rönsberg (2021, pp. 90–93).

[37] This was based on the contracts of ten RDCs, which they made available on a voluntary basis for the design of a harmonised contract, and on the expertise of the law firm iRights.Law. In addition to explanations of the contract, the explanatory document also addresses specific questions from research data centres regarding data acquisition.

[38] See also https://www.gesis.org/datenservices/ueber-die-datenservices/standards-und-workflows-datenservices/vorbereitung-datenuebergabe-pre-ingest (Copyright)

[39] A case-by-case assessment is necessary here, as this can vary depending on the funding, employment relationship, etc. See also, for example, Wünsche et al. (2022), who therefore recommend a rights clearance before the start of the project.

defined and the comprehensibility of the information provided for them be ensured (Keitel/Mitcham 2023; OAIS-Übersetzung/Terminologie 2013, 16, 19, 30; cf. also chapter 3.4).

When defining the designated community, it is important to bear in mind that this can change over time (OAIS-Übersetzung/Terminologie 2013, p. 16). For example, it would be conceivable that researchers from other disciplines would also like to work with the data – for example in the case of interdisciplinary research questions and also with a view to linking different data. It is also possible that data initially prepared only for a national audience could also be of interest to an international audience. In addition, it could be that certain data, initially intended purely for research purposes, will also be explicitly opened up for university teaching in the future. There could also be an opening for areas outside of research and science, such as journalism, business or the general public. An attempt should be made to take such potential changes directly into account, i.e. ideally to define the designated community somewhat more broadly, as retroactively ensuring comprehensibility for a broader designated community can be difficult and entail additional effort (OAIS-Übersetzung/Terminologie 2013, pp. 30–31).

Examples of designated community definitions at RDCs from the fields of social, behavioural, educational and economic data can be found here:

- RDC at the IQB: "*The FDZ at IQB's Designated Community consists of researchers from various disciplines in which educational processes are studied such as educational science, psychology, sociology, economics, and political science. Data access is granted to researchers from either universities or non-university research institutions, students being equally welcome.*" (https://www.iqb.hu-berlin.de/fdz/Grundlagen/CTS_with_Appendi.pdf, p. 4)
- FDZ-DZHW: " *The primary target group for use of the data comprises the national and international scientific communities (researchers, teachers, and students) in the field of higher education research and science studies as well as related disciplines. But the data are also open to other groups, provided that their intentions of use are in accord with the intended purposes.*" (https://fdz.dzhw.eu/en/about-the-fdz)
- AUSSDA: "*The primary beneficiaries of AUSSDA's services are social science researchers. Secondary user communities include students, educators, media representatives and the general public.*" (https://aussda.at/fileadmin/user_upload/p_aussda/Documents/CTS_Report_AUSSDA.pdf, p. 3)
- GESIS - Leibniz Institute for the Social Sciences (2024 [in publication]): "*GESIS's services are 'targeted primarily at scientists who work with methods of empirical social research, especially at universities and non-university research institutions in Germany and throughout the world'" ("Target Groups", https://www.gesis.org/en/institute). Accordingly, the Data Services are aimed at national and international researchers, academic teachers, and students in empirical social research with an emphasis on the areas of sociology and political science as well as social science in its entirety. Other target groups include those working in related political, social, and commercial social*

science environments. The services are offered to researchers (both in universities and non-university research institutions) and students.

To be able to use the GESIS Data Services, users should bring the following competencies and skill sets:

- *Information literacy: ability to use online search tools (e.g. GESIS Search) to identify datasets that are suitable to meet the researcher's needs*
- *Data literacy: "ability to read, understand, create, and communicate [quantitative social science] data as information" (https://en.wikipedia.org/wiki/Data_literacy).*
- *Statistical literacy: "The abilities to understand and reason with data, or arguments that use data [...]. [S]cientists also need to develop statistical literacy so that they can both produce rigorous and reproducible research and consume it." (https://en.wikipedia.org/wiki/Statistical_literacy)*

This entails

- *the ability to read and understand documentation and metadata provided by GESIS Data Services to assess the usability and suitability of the data for the research problem;*
- *proficiency in standard quantitative methods for data collection and analysis in the social sciences at least on the level of a master's degree to understand and work with the data.*
- *proficiency in using standard statistical software to render, process, and analyse the data, in particular STATA, SPSS."*

3.4. Ensuring the immediate comprehensibility of information

In addition to the data, an AIP mainly contains metadata. These are necessary in order to have sufficient information about the AIP in the long term. In some cases, the metadata required for comprehensibility can already be requested from the data providers (cf. chapter 3.1.2), in other cases it must be supplemented by the RDC (in particular metadata for long-term archiving) or converted into a structured format (cf. chapter 3.1.4).

Preservation metadata are regularly categorised into the metadata classes shown in Table 1. These are not always clearly defined and there may therefore be overlaps in terms of content (Hein 2012, p. 91; Jensen et al. 2019, p. 154; Keitel/Mitcham 2023 based on Gartner 2008):

Table 1: Metadata classes for preservation[40]

descriptive	Bibliographic information, e.g. title, author and content information, e.g. abstract, survey method
structural	Information on the overall context of the archived information object, e.g. relationship of files to each other

[40] According to the OAIS reference model, the Preservation Description Information (PDI) could be used here: https://wiki.dpconline.org/index.php?title=4.2.1.4.2_Preservation_Description_Information

administrative	Information on the archiving process, e.g. legal (licences, contracts, access rights, copyright...), provenance (origin, file history, editor...)
technical	Most automated metadata, e.g. file formats, original paths, file size, versions, sizes, checksums

Descriptive metadata provide information about the content and creation of digital objects and thus also determine their findability, referencing and reusability. One of the most well-known and widely used metadata standards is the Dublin Core™ Metadata Element Set (DCMES)[41]. This describes a web resource using 15 optional core elements (e.g. originator, title, description...). However, other metadata schemas and vocabularies should also be used, e.g. to make the data[42] easy to find. In particular, newer metadata schemas such as *schema.org* or discipline-specific ones such as the Data Documentation Initiative (DDI) standard[43], an internationally recognised standard for the documentation of data generated by observational methods in the social, behavioural, economic and health sciences, should be used for this purpose (Limani et al. 2022).

Structural metadata describe the internal structure and relationships between the individual elements. In a simple case, this can be, for example, a specific folder structure (e.g. individual folders for different survey waves) or the relationship between individual documents. For more complex document structures, the METS (Metadata Encoding & Transmission Standard)[44] can be used.

Administrative metadata include all information that is useful for the actual archiving process, subsequent access and authentic reproduction of the archived resources. This includes copyrights and licence information (who may access the resources and under what conditions?). In addition, information on the origin and archiving, processes that the resource has undergone (format conversion, etc.) are documented here.

Technical metadata (sometimes as a subset of administrative metadata) can now often be determined automatically at file level (e.g. file type, coding, creation date, etc.). As a standard, textMD[45] or elements of it (e.g. coding, language, etc.) can be used here.

Other more overarching metadata standards (especially for administrative and technical metadata) are PREMIS ("PREservation Metadata: Implementation Strategies")[46] and LMER (long-term archiving metadata for electronic resources)[47]. A categorisation of the PREMIS,

[41] https://www.dublincore.org/specifications/dublin-core/dces/
[42] see also Hoffstätter and Linne (2022)
[43] https://ddialliance.org/
[44] http://www.loc.gov/standards/mets/mets-home.html
[45] https://www.loc.gov/standards/textMD/
[46] Introduction to PREMIS: http://www.loc.gov/standards/premis/; German: https://www.loc.gov/standards/premis/understanding_premis_german.pdf. PREMIS is a standard and not a tool or software solution. A list of tools can be viewed here: https://coptr.digipres.org/index.php/PREMIS_(Preservation_Metadata_Implementation_Strategies)
[47] The long-term archiving metadata for electronic resources (LMER) were developed by the German National Library

METS and LMER standards, among others, can be found at Dappert/Enders (2010). PREMIS is comparatively widespread.

The use of standards generally increases interoperability. For example, if two long-term archive systems use the same metadata standard, information such as descriptive metadata can be easily exchanged. Compliance with standards also enables the reusability of information in other contexts (e.g. in search engines) (Hein 2012, p. 92). The RDC community in Germany does not (yet) use a uniform metadata standard for long-term archiving, but some RDCs are planning to use PREMIS.[48]

3.5. Comply with proven maintenance policies and procedures

For the long-term preservation of an AIP, it is necessary for an RDC to define appropriate preservation procedures for itself, to formulate and document its own principles in preservation policies and, last but not least, to comply with them (OAIS-Übersetzung/Terminologie 2013, p. 32).

In addition to technical strategies (Technology Watch) for long-term preservation and conservation, institutional preservation policies should also include aspects such as the mission and task of the RDC or archive, the designated community (Community Watch) (cf. chapter 3.3), the legal framework (cf. section 3.1.2), the access concepts (e.g. changing technology and more additional knowledge regarding anonymity of data)[49] and the collection profile (cf. chapter 3.1.1), as well as an emergency plan[50] and ideally a succession plan or provisions in the event of the RDC's dissolution (nestor-Arbeitsgruppe Policy 2014; VerbundFDB - AG Cessation 2024 [in Veröffentlichung]). It should be coordinated in feedback with the various stakeholders. This creates self-assurance and self-commitment for the organisation. In addition, long-term effective strategic and organisational basic elements of a digital long-term archive are disclosed and can thus contribute internally and externally to a general increase in trust (nestor-Arbeitsgruppe Policy 2014, p. 2).[51] Like the long-term archiving process as a whole, a preservation policy is a living document and is constantly being adapted and further developed.

The concepts of the following institutions can be cited as examples:[52]

based on the work of the National Library of New Zealand. The object model is based on the "Preservation Metadata: Metadata Implementation Schema" of the National Library of New Zealand (2003).

[48] The metadata schema of the RDC at the IQB, for example, is transparently documented here: https://mdr.iqb.hu-berlin.de/#/catalog/c09f72c7-36cc-1580-b32f-605401c3c830

[49] In a guideline from the RatSWD (2020, p. 19) on data protection, it is pointed out that anonymisation should be re-evaluated at certain intervals.

[50] For example, the disaster planning of the ICPSR: https://www.icpsr.umich.edu/web/pages/datamanagement/disaster/index.html

[51] A preservation policy can also be used internally for strategic decisions and reference can be made to the written obligations.

[52] A detailed step-by-step guide from the Digital Preservation Coalition can also be consulted: https://www.dpconline.org/digipres/implement-digipres/policy-toolkit/policy-step

- RDC at the IQB: (IQB - Institut zur Qualitätsentwicklung im Bildungswesen 2021b): https://www.iqb.hu-berlin.de/fdz/Grundlagen/Langzeitverfuegb.pdf
- GESIS: https://www.gesis.org/fileadmin/upload/Datenservices/Preservation_Policy/Digital_Preservation_Policy_1.4.8.pdf
  Preservation policy supplemented by a preservation plan with concrete measures (https://www.gesis.org/fileadmin/upload/Datenservices/Preservation_Policy/Preservation_Plan_1-0-0.pdf)
- AUSSDA: (Bischof 2022): https://aussda.at/fileadmin/user_upload/p_aussda/Documents/Preservation_Plan_v1_0.pdf
- German National Library: (Schrimpf/Steinke 2018): https://d-nb.info/1159789827/34

For practical implementation, the schematic overview of the levels of action, objectives and measures for long-term archiving based on ideas from Thibodeau (2002), shown in Table 2, can be helpful. The concept is presented below and supplemented with practical tips, in particular from Weber/Piesche (2021).

Table 2: Levels of action, objectives and measures for long-term archiving

Action level	**Goal**	**Measure**
physical level (Bitstream Preservation)	Physical preservation and safeguarding of data integrity	Regular replacement of data carriers, redundant storage, checksum verification
logical level (Logical Preservation)	Maintaining the technical interpretability and authenticity of the data	Migration of file formats or emulation of the original system environment
semantic level (Semantic Preservation)	Maintaining the interpretability and retrievability of data content	Content description of digital resources using context information

The basis and prerequisite for long-term preservation is therefore the backup of digital objects on the physical level (bitstream preservation). The integrity of the digital object itself and its storage medium is ensured. The integrity of the digital objects can be checked by using so-called fixity checks in the form of checksum formation (Weber/Piesche 2021, p. 340).[53] An algorithm is used to generate a "fingerprint" from the data, which changes every time a change is made to the files. In addition, the physical data carriers (hardware and software components) on which the files are stored must be regularly checked and, if necessary, replaced. In addition, a redundant backup[54] of the files should be implemented. In the case of redundant backups

[53] See also https://rdm-compas.org/articles/datenerhalt-erhaltungsmasnahmen

[54] The extended 3-2-1-1-0 rule can be used as a redundant backup: 3 different copies of all the company's digital objects, 2 copies are stored on different storage media, 1 copy of the storage media is located far from the company's headquarters, 1 copy is offline or indestructible and 0 errors in the backup (to be verified by regular restores) (https://www.storage-insider.de/data-protection-day-2022-die-3-2-1-1-0-regel-fuer-backups-a-1091301/).

on different storage areas, the checksum also serves to ensure that the files are identical by regularly comparing checksums of the files and using a valid copy (backup) in the event of irregularities (Weber/Piesche 2021, p. 340). In addition to the redundant backup, a backup and disaster recovery strategy should be set up so that copies of files can be restored if necessary. For further aspects of data security, see also Weber/Piesche (2021, p. 333).

In order to ensure the logical level (logical preservation) in the sense of technical interpretability of the digital objects (e.g. software to open and read the files), the file formats and corresponding software must be tracked. In the event of imminent or acute obsolescence of file formats, measures must be taken – usually format migration or emulation. Standardisation of the formats to be migrated and knowledge of their structure are central to format migration with as little loss as possible. For this reason, care should be taken to use open, simple and standardised formats for the long-term storage of research data. In the social sciences and related disciplines, proprietary file formats are often used for archiving, for example formats from Stata and SPSS software. Here, particular attention must be paid to the integrity of the files and the preservation of the logical level[55], as an open, traceable storage format is not available or only available to a limited extent. Efforts such as an open (exchange) format[56], which may also be further developed as a storage format in the future, are to be welcomed. Another option is to emulate the original system environment in which the file formats can still be opened. This is particularly relevant for file formats that cannot be migrated or can only be migrated with great effort.[57] Preservation metadata (see section 3.4) can support the tracking of file formats, for example (Hein 2012, pp. 104).

If the files are preserved and technically interpretable, sufficient information in the form of descriptive metadata is still required to understand the data (e.g. individual values in a data table) (semantic preservation) (see chapter 3.4).

All levels must be taken into account in order to secure and preserve files in the long term. The process and structures must also be regularly reviewed, adapted and further developed to take account of technological and cultural change as well as changes in the company's own designated community (see section 3.3). If changes occur during the (regular) review of these aspects, a re-appraisal may be necessary, i.e. a new review of the data acquisition or review of the relevant aspects.

The specific technical implementation of the requirements for long-term maintenance can be handled very differently. From a pre-organised digital folder structure with checklists, process definitions and automation to comprehensive archiving software. An overview of so-called OAIS-compliant software can be found at Weber/Piesche (2021, pp. 349–352). Institutions often cooperate with other organisations or specialist departments, e.g. computer centres, in order to exploit synergies. If long-term archiving or parts of it are to be outsourced, various services can be considered. It is important here to precisely define and document the division

[55] The retention of the original files (SIP) including all migration steps is particularly important here.
[56] https://www.konsortswd.de/angebote/open-data-format/
[57] cf. also Weber and Piesche (2021, pp. 341–342).

of labour and areas of responsibility of the players. Each RDC should evaluate and determine the extent to which services are outsourced. The effects of outsourcing for possible certification, for example according to CoreTrustSeal (see also chapter 4), should be considered (Hoffstätter/Beck 2023, pp. 5–6).

3.6. Making available and provision

The sixth minimum requirement of an archive according to the OAIS model is the availability and provision of the files. This has already been described in the guideline by Hoffstätter/Linne (2022).

4. Summary/Outlook

In this guideline, RDCs in the field of social, behavioural, educational and economic data have been provided with assistance that can contribute to the fulfilment of the particularly central requirements for data archiving – from data acquisition to long-term archiving – in the sense of an OAIS.

Five requirements for an RDC were analysed in more detail:

- With regard to the requirement "Negotiating and accepting information", the steps to be taken by the RDC during pre-ingest and ingest and what needs to be taken into account were discussed.
- With regard to the requirement "Obtain sufficient control for long-term preservation", information was provided on the relevant content of a submission agreement between data providers and the RDC.
- With regard to the requirement "Determine designated community", it was explained what needs to be considered when defining the designated community of the RDC.
- With regard to the requirement "Ensuring the immediate comprehensibility of information", it was shown how long-term comprehensibility of the data can be created
- With regard to the requirement "Adhere to proven preservation policies and procedures", the levels of preservation that exist and how the preservation measures can be implemented and documented were shown

The explanations make it clear that many of the foundations for functioning long-term archiving are laid at the data acquisition stage. Overall, it is necessary to clearly define all requirements, while at the same time "actively evaluating and constantly redesigning work processes and structures, as all existing solutions can only be used for a limited period of time" (nestor-Arbeitsgruppe Policy 2014, p. 2).[58]

[58] The processes and mechanisms of long-term archiving can also vary in scope and evolve over time. The levels of digital preservation (https://ndsa.org/publications/levels-of-digital-preservation/) or the proposal for curation & preservation levels of the CoreTrustSeal Standards And Certification Board (2023) (https://zenodo.org/records/8083359) can be used here as a basis for long-term archiving and also for opportunities for further development.

It is recommended that this topic also be considered in connection with external certifications. While accreditation by the RatSWD[59] focuses on access to sensitive research data according to certain minimum criteria, the CoreTrustSeal[60] (CTS) is a certification option for demonstrating the trustworthiness of[61] long-term digital archives. The institution must describe its internal processes using a self-assessment procedure based on 16 criteria. These are reviewed by a board and awarded for three years.[62] These certifications can be used to demonstrate trustworthiness, legitimacy and quality to various stakeholders (data providers, data users, research funders), but also within the organisation itself, and to make these transparent and increase visibility (Recker et al. 2020, p. 101). Certification can also help to ensure that the organisation's own processes, internal communication and documentation are scrutinised, standardised and optimised. Ideally, the CTS requirements should be consulted directly as a checklist when designing and planning a repository. Documents from already certified institutions can serve as a valuable source of information.

These explanations here should be seen as a starting point. Closer examination of these topics should definitely be conducted in dialogue with other institutions in the community, neighbouring communities, and associations and networks in order to exploit synergies.[63] For RDCs in the field of social, behavioural, educational and economic data, it can be helpful to check whether they can join these existing associations and networks in order to benefit from existing solutions and structures and at the same time share and discuss their own developments in the group. In this context, a standardised understanding of terms and processes should also be promoted (Naumann 2023). In addition, it could also be worth examining whether certain processes, which are currently still very RDC-specific, could be further harmonised and standardised between different (similar) RDCs and also technically supported and partially automated, such as the definition of minimum requirements and the ingest check. Such overarching process optimisation should also be seen as a central task of an RDC in data archiving. At the same time, however, the need for a certain plurality of objectives and methods of different institutions should also be accepted (Naumann 2023).

[59] https://www.konsortswd.de/angebote/datenzentren/qualitaetssicherung-zertifizierung/akkreditierung/

[60] https://www.coretrustseal.org

[61] The TRUST principles (Lin et al. 2020b, 2020a) stand for transparency, responsibility, user focus, sustainability and technology as a set of overarching guidelines for trustworthy archives. They therefore have a certain amount of overlap with concepts such as the OAIS and the CoreTrustSeal, but these focus clearly on the aspect of long-term archiving. How exactly these and other concepts such as the FAIR and CARE principles are connected is currently being worked out in a working group (https://www.rd-alliance.org/group/rdawds-trust-principles-outreach-and-adoption-working-group/case-statement/rdawds-trust).

[62] In addition to this basic certification, there are other certifications such as the nestor seal, which works with an extended self-evaluation for digital long-term archives on the basis of DIN standard 31644. This seal is more comprehensive and detailed than the CoreTrustSeal. There is also certification in accordance with the ISO 16363:20128 standard, which ensures quality by means of an audit procedure involving an inspection.

[63] The focus, structure and tasks of the alliances can be different, but can also overlap. The Verbund Forschungsdaten Bildung (VerbundFDB), for example, is a thematically oriented cooperation project of research data centres from the field of educational research. QualiBi works closely with VerbundFDB as a platform for research data in qualitative educational research. QualidataNet, which is currently being established, relies on a methodological focus as a network of research data centres, archives and repositories that archive research materials from qualitative social research and make them available for re-use.

References

Altenhöner, R., & Oellers, C. (Eds.). (2012). *Langzeitarchivierung von Forschungsdaten: Standards und disziplinspezifische Lösungen*. Scivero.

Bischof, Christian (2022). *Preservation Plan* (No. 1.0). Vienna. The Austrian Social Science Data Archive. https://aussda.at/fileadmin/user_upload/p_aussda/Documents/Preservation_Plan_v1_0.pdf

Brislinger, Evelyn, & Moschner, Meinhard. (2019). Datenaufbereitung und Dokumentation. In U. Jensen, S. Netscher, & K. Weller (Eds.), *Forschungsdatenmanagement sozialwissenschaftlicher Umfragedaten: Grundlagen und praktische Lösungen für den Umgang mit quantitativen Forschungsdaten* (pp. 97–114). Verlag Barbara Budrich.

Butzlaff, Iris (2022). *Data Deposit Guideline (Public version)*. Vienna. The Austrian Social Science Data (AUSSDA). https://aussda.at/fileadmin/user_upload/p_aussda/Documents/Data-Deposit-Guideline_SUF_v2_0.pdf

The Consultative Committee For Space Data Systems (2012). *Recommendation for Space Data System Practices: Reference Model For An Open Archival Information System (OAIS)*. https://public.ccsds.org/Pubs/651x0m1.pdf

CoreTrustSeal Standards And Certification Board (2023). *Curation & Preservation Levels: Coretrustseal Discussion Paper*. https://doi.org/10.5281/ZENODO.8083359

Dappert, Angela, & Enders, Markus (2010). Digital Preservation Metadata Standards. *Information Standards Quarterly*, *22*(2), Article 13, 4. https://www.loc.gov/standards/premis/FE_Dappert_Enders_MetadataStds_isqv22no2.pdf

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

Gartner, Richard (2008). Metadata for digital libraries: state of the art and future directions. *JISC Technology and Standards Watch*. https://www.webarchive.org.uk/wayback/archive/20140613220103/http://www.jisc.ac.uk/media/documents/techwatch/tsw_0801pdf.pdf

(2024 [in Veröffentlichung]). *GESIS Data Services: CoreTrustSeal Requirements 2023-2025*. GESIS – Leibniz-Institut für Sozialwissenschaften.

Hein, Stefan. (2012). Metadaten für die Langzeitarchivierung. In R. Altenhöner & C. Oellers (Eds.), *Langzeitarchivierung von Forschungsdaten: Standards und disziplinspezifische Lösungen* (pp. 87–110). Scivero.

Hoffstätter, Ute, & Beck, Kerstin (2023). *Workshopdokumentation: CoreTrustSeal: Erstveranstaltung*. https://doi.org/10.5281/ZENODO.10213806

Hoffstätter, Ute, & Linne, Monika (2022). *Datenzugang. Einführung in das Thema Zugang zu Daten der Sozial-, Verhaltens-, Bildungs- und Wirtschaftswissenschaften in Forschungsdatenzentren.* https://doi.org/10.5281/ZENODO.7347063

(2021a). *Aufbereitung der Datensätze am FDZ am IQB / Steps of Data Preparation*. Berlin. Forschungsdatenzentrum am Institut zur Qualitätsentwicklung im Bildungswesen (FDZ am IQB). https://www.iqb.hu-berlin.de/fdz/Datenuebergabe/Vorlage_AbkStudi.pdf

(2021b). *Konzept zur Langzeitverfügbarkeit digitaler Datensätze des FDZ am IQB.* Berlin. Forschungsdatenzentrum am Institut zur Qualitätsentwicklung im Bildungswesen (FDZ am IQB).

Jensen, Uwe, Zenk-Möltgen, Wolfgang, & Wasner, Catharina. (2019). Metadatenstandards im Kontext sozialwissenschaftlicher Daten. In U. Jensen, S. Netscher, & K. Weller (Eds.), *Forschungsdatenmanagement sozialwissenschaftlicher Umfragedaten: Grundlagen und praktische Lösungen für den Umgang mit quantitativen Forschungsdaten* (pp. 151–178). Verlag Barbara Budrich.

Keitel, Christian, & Mitcham, Jenny (2023). *Defining the Designated Community*. https://doi.org/10.7207/twgn23-01

Lauber-Rönsberg, Anne. (2021). Rechtliche Aspekte des Forschungsdatenmanagements. In M. Putnings, H. Neuroth, & J. Neumann (Eds.), *Praxishandbuch Forschungsdatenmanagement* (pp. 89–114). De Gruyter Saur. https://doi.org/10.1515/9783110657807-005

Lauber-Rönsberg, Anne, Krahn, Philipp, & Baumann, Paul (2018). *Gutachten zu den rechtlichen Rahmenbedingungen des Forschungsdatenmanagements*. https://tu-dresden.de/gsw/phil/irget/jfbimd13/ressourcen/dateien/dateien/DataJus/DataJus_Kurzfassung_Gutachten_12-07-18.pdf?lang=de&set_language=de

Lavoie, Brian (2014). *The Open Archival Information System (OAIS) Reference Model: Introductory Guide (2nd Edition).* https://doi.org/10.7207/twr14-02

Liegmann, Hans, & Neuroth, Heike. (2010). Einführung. In H. Neuroth, A. Oßwald, R. Scheffel, S. Strathmann, & K. Huth (Eds.), *nestor Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung.* Hülsbusch / Univ.-Verl. Göttingen.

Limani, Fidan, Younes, Yousef, Bach, Janete Saldanha, Hiseni, Valentina, Mutschke, Peter, & Mathiak, Brigitte (2022). *Konsortswd Measure 5.2: Enhancing data findability Milestones 1, 2, and 3 report.* https://doi.org/10.5281/ZENODO.7224672

Lin, Dawei, Crabtree, Jonathan, Dillo, Ingrid, Downs, Robert R., Edmunds, Rorie, Giaretta, David, Giusti, Marisa de, L'Hours, Hervé, Hugo, Wim, Jenkyns, Reyna, Khodiyar, Varsha, Martone, Maryann E., Mokrane, Mustapha, Navale, Vivek, Petters, Jonathan, Sierman, Barbara, Sokolova, Dina V., Stockhause, Martina, & Westbrook, John (2020a). The TRUST Principles for digital repositories. *Scientific Data*, *7*(1), 144. https://doi.org/10.1038/s41597-020-0486-7

Lin, Dawei, Crabtree, Jonathan, Dillo, Ingrid, Downs, Robert R., Edmunds, Rorie, Giaretta, David, Giusti, Marisa de, L'Hours, Hervé, Hugo, Wim, Jenkyns, Reyna, Khodiyar, Varsha, Martone, Maryann E., Mokrane, Mustapha, Navale, Vivek,

Petters, Jonathan, Sierman, Barbara, Sokolova, Dina V., Stockhause, Martina, & Westbrook, John (2020b). Die TRUST-Prinzipien für digitale Repositorien. Advance online publication. https://doi.org/10.5281/ZENODO.6256222

Mauer, Reiner. (2012). Das GESIS Datenarchiv für Sozialwissenschaften. In R. Altenhöner & C. Oellers (Eds.), *Langzeitarchivierung von Forschungsdaten: Standards und disziplinspezifische Lösungen* (pp. 197–215). Scivero.

Mauer, Reiner (2016). *Auswahl und Bewertung von Forschungsdaten für die Archivierung und Nachnutzung*. GESIS – Leibniz-Institut für Sozialwissenschaften. http://www.nestor.sub.uni-goettingen.de/school_2016/slides/PERICLES_WP7_T7-3_UGOE_nestor_PERICLES_School_Presentation_03.pdf

Meyermann, Alexia, & Porzelt, Maike (2014). *Hinweise zur Anonymisierung von qualitativen Daten* (forschungsdaten bildung informiert No. 1). Frankfurt am Main. Forschungsdatenzentrum (FDZ) Bildung am DIPF; Deutsches Institut für Internationale Pädagogische Forschung.

Meyermann, Alexia, & Porzelt, Maike. (2019). *Datenschutzrechtliche Anforderungen in der empirischen Bildungsforschung. Eine Handreichung*. DIPF | Leibniz-Institut für Bildungsforschung und Bildungsinformation : Frankfurt am Main. https://doi.org/10.25656/01:21990

Naumann, Kai (2023). *Langzeitarchivierung im Landesarchiv Baden-Württemberg*. https://doi.org/10.5281/ZENODO.10276163

nestor-Arbeitsgruppe Policy (2014). *Leitfaden zur Erstellung einer institutionellen Policy zur digitalen Langzeitarchivierung* (nestor-materialien No. 18). https://d-nb.info/1051731216/34

Netscher, Sebastian, & Jensen, Uwe. (2019a). Forschungsdatenmanagement systematisch planen und umsetzen. In U. Jensen, S. Netscher, & K. Weller (Eds.), *Forschungsdatenmanagement sozialwissenschaftlicher Umfragedaten: Grundlagen und praktische Lösungen für den Umgang mit quantitativen Forschungsdaten* (pp. 37–55). Verlag Barbara Budrich. https://doi.org/10.3224/84742233.04

Netscher, Sebastian, & Jensen, Uwe. (2019b). Forschungsdatenmanagement systematisch planen und umsetzen. In U. Jensen, S. Netscher, & K. Weller (Eds.), *Forschungsdatenmanagement sozialwissenschaftlicher Umfragedaten: Grundlagen und praktische Lösungen für den Umgang mit quantitativen Forschungsdaten* (pp. 37–55). Verlag Barbara Budrich.

OAIS-Übersetzung/Terminologie, nestor Arbeitsgruppe. (2012). *Referenzmodell für ein Offenes Archiv-Informations-System: Deutsche Übersetzung*. nestor - Kompetenznetzwerk Langzeitarchivierung und Langzeitverfügbarkeit Digitaler Ressourcen für Deutschland.

OAIS-Übersetzung/Terminologie, nestor Arbeitsgruppe. (2013). *Referenzmodell für ein Offenes Archiv-Informations-System: Deutsche Übersetzung 2.0*. nestor - Kompetenznetzwerk Langzeitarchivierung und Langzeitverfügbarkeit Digitaler Ressourcen für Deutschland. https://doi.org/10.18452/1532

Rat für Sozial- und Wirtschaftsdaten (2022). *Tätigkeitsbericht 2020 der vom RatSWD akkreditierten Forschungsdatenzentren (FDZ).* https://doi.org/10.17620/02671.65

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

Recker, Jonas, & Brislinger, Evelyn. (2019). Dateiorganisation in empirischen Forschungsprojekten. In U. Jensen, S. Netscher, & K. Weller (Eds.), *Forschungsdatenmanagement sozialwissenschaftlicher Umfragedaten: Grundlagen und praktische Lösungen für den Umgang mit quantitativen Forschungsdaten* (pp. 81–95). Verlag Barbara Budrich.

Recker, Jonas, Helbig, Kerstin, & Neumann, Janna (2020). Zertifizierung von Forschungsdatenrepositorien: Wege, Praxiserfahrungen und Perspektiven. Advance online publication. https://doi.org/10.17192/BFDM.2020.2.8280

Schallaböck, Jan, Hoffstätter, Ute, Buck, Daniel, & Linne, Monika (2023). *Mustervertrag Datennutzung KonsortSWD.* https://doi.org/10.5281/ZENODO.8186162

Schallaböck, Jan, Kreutzer, Till, Hoffstätter, Ute, & Buck, Daniel (2023a). *Mustervertrag Datenaufnahme KonsortSWD.* https://doi.org/10.5281/zenodo.7648897

Schallaböck, Jan, Kreutzer, Till, Hoffstätter, Ute, & Buck, Daniel (2023b). *Mustervertrag Datenaufnahme KonsortSWD.* https://doi.org/10.5281/ZENODO.7648898

Schrimpf, Sabine. (2012). Überblick über das OAIS-Referenzmodell. In R. Altenhöner & C. Oellers (Eds.), *Langzeitarchivierung von Forschungsdaten: Standards und disziplinspezifische Lösungen* (pp. 51–68). Scivero.

Schrimpf, Sabine, & Steinke, Tobias. (2018). *Langzeitarchivierungs-Policy der Deutschen Nationalbibliothek* (Version 1.2 Stand: 4.5.2018). Deutsche Nationalbibliothek.

Schumann, Natascha. (2012). Einführung in die digitale Langzeitarchivierung. In R. Altenhöner & C. Oellers (Eds.), *Langzeitarchivierung von Forschungsdaten: Standards und disziplinspezifische Lösungen* (pp. 39–48). Scivero.

Schwens, Ute, & Liegmann, Hans. (2011). D 9 Langzeitarchivierung digitaler Ressourcen. In R. Kuhlen, T. Seeger, & D. Strauch (Eds.), *Grundlagen der praktischen Information und Dokumentation* (5th ed., pp. 567–570). Saur; De Gruyter. https://doi.org/10.1515/9783110964110.567

Thibodeau, Kenneth. (2002). Overview of Technological Approaches to Digital Preservation and Challenges in Coming Years. In Council on Library and Information Resources (Ed.), *The state of digital preservation: An international perspective; conference proceedings, Documentation Abstracts, Inc. Institutes for Information Science, Washington, D.C., April 24-25, 2002* (pp. 4–31). CLIR. https://www.clir.org/wp-content/uploads/sites/6/pub107.pdf

Uwe Jensen, GESIS (2012). *Metadaten für Forschungsdaten: Welche Standards gibt es?* GESIS – Leibniz-Institut für Sozialwissenschaften. 101. Deutscher Bibliothekartag, Hamburg.

Verbund Forschungsdaten Bildung. (2018). *Formulierungsbeispiele für „informierte Einwilligungen"* (Version 2.1). DIPF | Leibniz-Institut für Bildungsforschung und Bildungsinformation. https://doi.org/10.25656/01:22301

(2024 [in Veröffentlichung]). *Template zur Erfassung von relevanten Informationen für eine Nachfolgeregelung für den Fall der Auflösung eines Datenzentrums.* VerbundFDB - AG Cessation.

Wallace, David A. (2001). Archiving metadata forum: Report from the Recordkeeping Metadata Working Meeting, June 2000. *Archival Science*, *1*(3), 253–269. https://doi.org/10.1007/BF02437690

Watteler, Oliver (2010). *Erstellung von Methodenberichten für die Archivierung von Forschungsdaten*. Köln. GESIS – Leibniz-Institut für Sozialwissenschaften.

Watteler, Oliver, & Ebel, Thomas. (2019). Datenschutz im Forschungsdatenmanagement. In U. Jensen, S. Netscher, & K. Weller (Eds.), *Forschungsdatenmanagement sozialwissenschaftlicher Umfragedaten: Grundlagen und praktische Lösungen für den Umgang mit quantitativen Forschungsdaten* (pp. 57–80). Verlag Barbara Budrich.

Weber, Andreas, & Piesche, Claudia. (2021). 4.2 Datenspeicherung, -kuration und Langzeitverfügbarkeit. In M. Putnings, H. Neuroth, & J. Neumann (Eds.), *Praxishandbuch Forschungsdatenmanagement* (pp. 327–356). De Gruyter Saur. https://doi.org/10.1515/9783110657807-019

Whyte, Angus, & Wilson, Andrew (2010). *How to Appraise & Select Research Data for Curation.* Digital Curation Center (DCC); Australian National Data Service (ANDS).

Wünsche, Stephan, Soßna, Volker, Kreitlow, Vanessa, & Voigt, Pia (2022). Urheberrechte an Forschungsdaten – Typische Unsicherheiten und wie man sie vermindern könnte. Advance online publication. https://doi.org/10.17192/bfdm.2022.1.8369

Imprint

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

September 2024

KonsortSWD Working Paper:
As part of the National Research Data Infrastructure, KonsortSWD is expanding services to support research with data in the social, behavioural, educational and economic sciences. Our mission is to strengthen, expand and deepen the research data infrastructure for researching society. It should be user-orientated and take into account the needs of the research communities. The network of research data centres established by the German Data Forum (Rat für Sozial- und Wirtschaftsdaten – RatSWD) over the past two decades is an important cornerstone of this.
This series contains articles on research data management that are produced in the context of KonsortSWD. Articles that were externally and double-blind reviewed are labelled accordingly.

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

DOI: 10.5281/zenodo.13692961

Suggested citation:
Hoffstätter, U., & Weber, A. (2024). *Long-term archiving of research data. Introduction to the topic of data from the social, behavioural, educational and economic sciences in research data centres.* KonsortSWD Working Paper No. 9a/2024. Consortium for the Social, Behavioural, Educational and Economic Sciences (KonsortSWD). https://doi.org/10.5281/zenodo.13692961