IQB
Forschungsdatenzentrum

Concept for the long-term availability of research data

Humboldt-Universität zu Berlin
Institut zur Qualitätsentwicklung im Bildungswesen
Unter den Linden 6, 10099 Berlin

Telefon +49 30 2093 46500
iqboffice@iqb.hu-berlin.de
www.iqb.hu-berlin.de

KULTUSMINISTER KONFERENZ

HUMBOLDT-UNIVERSITÄT ZU BERLIN

Revised version of the previous document

Forschungsdatenzentrum am Institut zur Qualitätsentwicklung im Bildungswesen (FDZ at the IQB) [Research Data Centre at the Institute for Educational Quality Improvement (FDZ at IQB)] (2022). *Concept for the long-term availability of digital objects sets at the FDZ at the IQB* [Konzept zur Langzeitverfügbarkeit digitaler Datensätze des FDZ at the IQB]. Berlin: IQB – Institut zur Qualitätsentwicklung im Bildungswesen.

Bibliographical information / Please quote as

Forschungsdatenzentrum am Institut zur Qualitätsentwicklung im Bildungswesen (FDZ at the IQB) [Research Data Centre at the Institute for Educational Quality Improvement (FDZ at IQB)] (2025). *Concept for the long-term availability of research data.* [Konzept zur Langzeitverfügbarkeit von Forschungsdaten]. Berlin: IQB - Institut zur Qualitätsentwicklung im Bildungswesen. http://doi.org/10.5159/IQB_long-term-concept_v1

Inhalt

1 Terms and definition of long-term archiving 4

2 Technical infrastructure 6

3 Long-term availability at the FDZ at the IQB 7

3.1 Ingest 8

3.2 Archival Storage 8

3.2.1 Which digital objects are made available for the long term? 9

3.2.2 When are archival packages created? 9

3.2.3 Who is responsible for the creation of the archival packages? 9

3.2.4 How are the archival packages created? 9

3.2.5 Where are the long-term available digital objects stored? Who has access? 11

3.2.6 3.2.6 How is the readability of the AIP ensured? 11

3.3 Data preparation 13

3.3.1 Documentation 13

3.4 Access 14

1 Terms and definition of long-term archiving

Long-term archiving is understood to mean the long-term storage and preservation of the permanent availability of information. Especially in the digital domain, this is of not insignificant relevance, because in analogue archiving the focus is on preserving the physical medium; in long-term archiving of digital objects, on the other hand, it is not necessary to preserve the medium itself, but to prevent data loss through timely copying. Long-term archiving requires strategies to deal with changes in the types of storage media, file formats, etc. The archiving network nestor[1] defines long-term archiving as follows:

> *"Long-term means for the preservation of digital resources [...] the responsible development of strategies that can cope with the constant change caused by the information market. [...] Rather, it includes the preservation of the permanent availability and thus re-use [...] of digital resources."*[2]

Digital objects consists of a fixed sequence of bits (=bitstream) that are stored on data carriers. In order to keep digital objects permanently available, preservation measures must be taken (= **bitstream preservation**). However, the lifespan and reliability of data carriers is limited and individual bits can no longer be read over time or can be read incorrectly. Bitstream preservation uses technical measures such as checksums and redundancy to ensure that the bitstream remains unchanged over longer periods of time, even after technology changes. This type of data protection is an elementary measure for data security; it is about the pure physical preservation of the data and its readability. Bitstream preservation is a basic prerequisite for long-term digital archiving.

Backups are part of bitstream preservation. A sensible backup strategy is to archive several copies of the backup redundantly at different locations, distributed on different storage media. Backup is therefore purely a matter of data protection, i.e. the copying of data in order to be able to copy the data back in the event of data loss. This does not say anything about the readability or usability of the data.

In order to be able to reuse the permanently preserved digital objects, it must remain interpretable, and this is the core of long-term archiving, which is also often referred to as digital curation. **Digital curation** means maintaining, preserving and enriching digital research data throughout its life cycle. These measures go beyond the mere technical preservation of the bitstream and also require subject-matter expertise. For example, it is important that the files

[1] = German network for long-term archiving and long-term availability of digital resources

[2] Neuroth, H., Oßwald, A., Scheffel, R., Strathmann, S. & Huth, K. (2010). nestor Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung; im Rahmen des Projektes: Nestor - Kompetenznetzwerk Langzeitarchivierung und Langzeitverfügbarkeit digitaler Ressourcen für Deutschland. Version 2.3. Verfügbar unter: urn:nbn:de:0008-2010071949, Kap. 1:3. Last access: 5.6.2025

remain readable for new software used in the field. For research data management, this then also means that the long-term research value is preserved and the risk of digital obsolescence is reduced.

The data must not only be stored and preserved (curated), but in order to develop its added value, it must be able to be used by the scientific community. This can be enabled by trusted data centres. The Research Data Centre at the IQB is such a trusted data centre. This means that the FDZ at the IQB takes measures to enable the curation and preservation of digital research data throughout its lifecycle, as well as to securely archive and make available the digital research data at the FDZ at the IQB in perpetuity.

Long-term archiving is thus to be distinguished from pure data backup.

2 Technical infrastructure

The IQB and thus also the FDZ at the IQB make use of the infrastructure of the computer centre of the Humboldt-Universität zu Berlin, the Computer and Media Service (CMS).

In the case of storage services, the CMS is responsible for the IQB services, its servers and the underlying infrastructure (Storage Area Network (SAN)[3] , backup, network). It takes care of backups, media monitoring and refreshing. To protect the data, the following backup and access control procedures[4] are used to ensure the (physical) security of the digital archive holdings:

- Data centre and server rooms are secured against unauthorised access.
- Smoke and water detectors are installed
- Temperatures in the server rooms are monitored
- Redundant data storage at different locations (Adlershof, Mitte)
- Frequent incremental and full backups (the contents of the IQB network drives are backed up every night, 60-day retention period, 2 weeks in access)[5]
- Variety of storage media and frequent media refreshing

The CMS also has mirrors of the servers in case of technical failure, malicious action or human error. These store the last 2 weeks.

[3] Redundant network at Severn to provide fail-safe hard disk storage. A network that connects servers and storage systems via dedicated lines. Structurally, a SAN is set up in the same way as a local area network (LAN): there are hubs, switches and routers; see also: www.cms.hu-berlin.de/de/dl#/svc14

[4] see also here: https://www.cms.hu-berlin.de/de/publikationen/ordnungen/ZutrittCMS. Last access: 5.6.2025

[5] see also here: https://www.cms.hu-berlin.de/de/dl#/svc15 and https://www.cms.hu-berlin.de/de/dl/systemservice/fileservice/tsm. Last access: 5.6.2025; as well as on the subject of backups: creation and retention period.

PostgreSQL

- all databases are backed up twice a day (01:00 and 13:00)
- the midday backups are kept for 7 days
- the nightly backups are kept for 60 days
- as all backups are additionally stored on tape at the CMS central backup service, all mentioned retention times are increased by another 60 days

MySQL

- all databases are backed up twice a day (01:00 and 13:00)
- the midday backups are kept for 7 days
- the nightly backups are kept for 60 days
- as all backups are additionally stored on tape at the CMS central backup service, all mentioned retention times are increased by another 60 days

3 Long-term availability at the FDZ at the IQB

The FDZ at the IQB works along defined workflows - from data selection (according to the Collection Policy of the FDZ at the IQB) and ingest (incl. verification and validation) to contract conclusion and data preparation (incl. fine checking, in-depth checking, metadata enrichment and communication with data providers) to documentation and data availability (incl. publication and release).

Following the functional model of the Open Archival Information System (OAIS)[6], the FDZ at the IQB is the archive/repository or the link (connector) between data providers and data users.

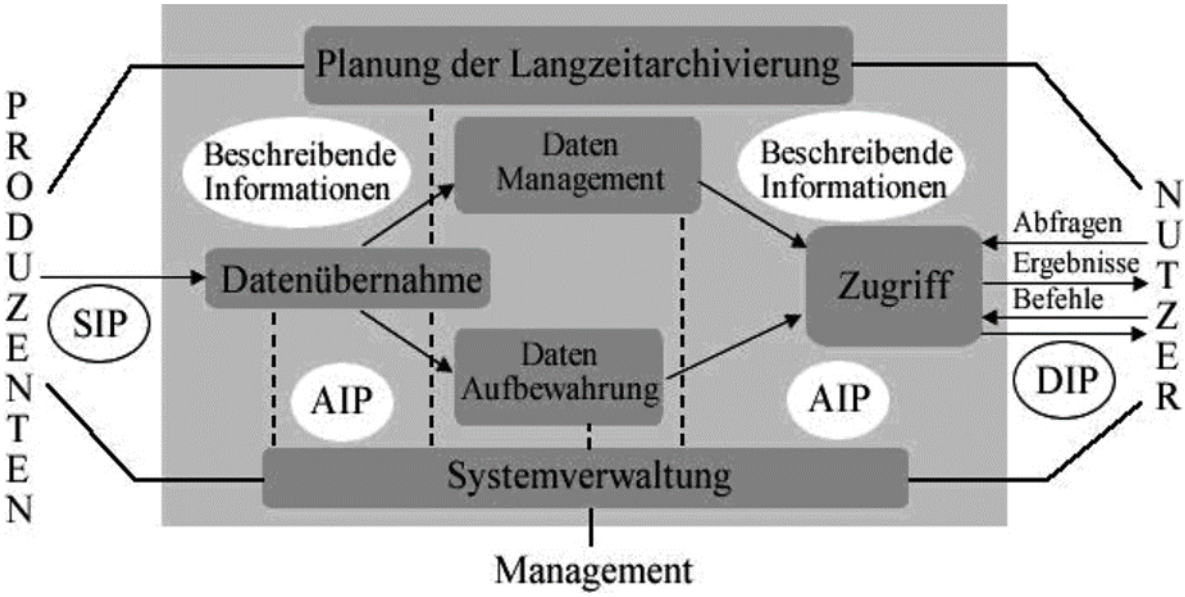

Fig.: OAIS reference model[7]

All digital objects (i.e: data in the form of datasets and further contextual information in the form of accompanying documentation material and metadata) that the FDZ receives from the data providers constitute the so-called Submission Information Package (=SIP) according to OAIS terminology.

The Archival Information Package (=AIP) comprises the digital objects - supplemented with metadata by the FDZ at the IQB – which are stored long-term at the FDZ at the IQB, i.e. archived for the long term.

The research data and documentation materials that are ultimately made available to data users are referred to as the Dissemination Information Package (=DIP).

[6] see for this: CCSDS (2024). Recommendation for Space Data System Practices. Reference Model for an Open Archival Information System (OAIS). Recommended Practice. Washington D.C.: CCSDS. https://public.ccsds.org/Pubs/650x0m3.pdf. Last access: 5.6.2025

[7] see: urn:nbn:de:0008-2010061762, S. 11

The AIP contains 1) all digital objects received from the data providers (=SIP), 2) the digital objects to be prepared for the data users (=DIP) and 3) all documentation and processing steps (incl. syntax/s) that were carried out at the FDZ at the IQB in order to provide one or more data products for the data users; but also contracts, access regulations, communication with the data providers and other accompanying documentation materials. In other words, the FDZ at the IQB generates the DIP from the SIP via the AIP creation. All digital objects are available in formats with long-term availability, and a checksum file is also created for each AIP.

3.1 Ingest

Ingest is about the transfer of digital objects.

When the data are transferred to the FDZ at the IQB, a data provision contract is concluded between the FDZ at the IQB and the data providers; this contract deals with and regulates all legal issues. The data provider assures that the rights of third parties are not affected by the transfer to the FDZ at the IQB and the archiving of the research data there, and that all parties involved comply with the DFG regulations on "safeguarding good scientific practice".

The FDZ at the IQB team and the data providers agree on an access concept which regulates which research data and accompanying materials are passed on to third parties and how (a. publicly on the website of the FDZ at the IQB or by default, b. only on request together with the data or c. never).

To ensure data quality, the FDZ at the IQB has defined minimum requirements that the submitted research data must meet. For all incoming digital objects, the FDZ at the IQB checks whether the material supplied is complete, correct and in a suitable condition (readable, virus-free, etc.) (=technical-formal check). In the subsequent content assessment, it is checked whether the research data comply with the scope of the FDZ at the IQB, the collection policy and the minimum requirements (=check whether all data sets meet both the technical and documentary requirements (=clarity of data set and documentation) and the legal requirements (=compliance with data protection, copyright(s), (law)). In addition, the re-use potential, i.e. the potential of the research data for secondary analysis, is assessed.

All work steps are documented in an inhouse database (=metadata management and documentation tool), which is only available to authorised employees. To standardise procedures, the entire ingest process is controlled by FDZ-internal documents, checklists and templates that are regularly updated.

3.2 Archival Storage

Archival storage includes the digital archive storage, its organisation as well as its structure in the narrower sense. This is about the procedure for creating archival packages. In the archive storage of the FDZ at the IQB, the AIPs are compiled in such a way that their retrievability as well as their readability and interpretability are ensured. The AIPs are then transferred to the LZA storage facility of the Computer and Media Service (CMS) of the Humboldt-Universität (HU) zu Berlin and stored there. Further rules apply there, including that there the checksums of the AIPs are regularly checked (bitstream) so that the integrity and authenticity of the digital objects is maintained.

3.2.1 Which digital objects are made available for the long term?

The FDZ at the IQB makes research data from national and international school performance studies as well as from national studies with competence measurements in the field of education available to the scientific community for re- and secondary analyses. So far, the available research data are mainly quantitative data. The digital objects of the studies published at the FDZ at the IQB are made digitally available for the long term.

In concrete terms, this means that the FDZ at the IQB archives all digital objects from the studies that are made available for the long term. In addition, the original data sets including the accompanying material are archived in long-term archived formats as well as all documentation and processing steps which were carried out at the FDZ at the IQB (e.g. processing syntaxes and documents for the evaluation of research data) in order to generate the DIP per study for data users.

3.2.2 When are archival packages created?

An archive package (=AIP) for a study is created 1) after data transfer and the decision that the study should be made available, and 2) for a new version and contains the SIP and the digital objects it contains in a format that is available in the long term. After data preparation, the result of which is the DIP for the study to be issued to data users, the AIP is filled with it - here, too, everything is in long-term available formats.

3.2.3 Who is responsible for the creation of the archival packages?

The data librarian is responsible for creating the archival packages.

3.2.4 How are the archival packages created?

The aim is that the digital objects provided by the FDZ at the IQB are and remain permanently available and readable, i.e. functional and usable, and that their content is and remains interpretable.

3.2.4.1 Data formats

As part of long-term availability, digital objects are stored in unencrypted, non-compressed, non-proprietary formats using open, documented standards.

Redundancy in digital objects is useful and sensible in terms of data security, i.e., the FDZ at the IQB stores text files in both PDF/A[8] and .txt formats, and table data in .csv format. The .txt files are only stored in case the primary PDF/A format should no longer be functional. Therefore, text files are additionally converted to .txt formats, but nothing more is done with them.

We do not save in SPSS portable format, since it is not recommended as an archive format by the Library of Congress, for example[9].

[8] PDF/A is a file format for long-term archiving of digital documents that has been standardised by the International Organization for Standardization (ISO) as a subset of the Portable Document Format (PDF). Standardised metadata is embedded in the document. PDF/A documents support full-text search and are self-contained and independent: Elements (fonts, colour profiles, etc.) needed for proper reproduction are included in the document. A PDF/A document must not contain references to external sources. Simple informative references such as links to web pages are permitted. PDF/A saves storage space. PDF/A documents remain valid without notice.

[9] https://www.loc.gov/preservation/digital/formats/fdd/fdd000468.shtml. Last access: 5.6.2025

3.2.4.2 Procedure

At the FDZ at the IQB, a so-called archival package is created for each study in order to create long-term available data. An archival package is the so-called Archival Information Package (AIP).

In concrete terms, this means that after submission of the digital objects by data providers and the decision that the research data of the study should be made available, these are immediately converted into long-term archivable formats, i.e. data sets in SPSS and/or .xls format are converted into .csv format, files that are available in R are saved identically in the AIP of the study, if necessary with the corresponding R packages and accompanying documentation (text) in PDF/A and additionally as .txt format and saved in the AIP of the study. The same applies to the procedure for new versions.

After data preparation, the digital objects to be provided (=DIP) are stored in the AIP; also all in formats with long-term availability.

For each AIP, a checksum file is created for all the digital objects it contains to ensure data integrity.

For each study and version, the associated administrative (e.g.: DOI information, citation proposal, data provider, access regulations, rights) as well as descriptive (e.g.: study period, survey period, selection procedure) metadata[10] (in .xml and .csv format) are stored in the AIP, because the metadata are also made long-term available. The FDZ at the IQB does NOT record variable or value labels as metadata, i.e. these are contained in the data sets, but are not recorded separately as metadata. For each study as well as for a new version, the metadata is included in the AIP as a corresponding .xml file [11].

After data preparation, the AIP is filled with further objects: in addition to the digital objects to be provided - i.e. including the syntax/s [12].

3.2.4.3 Archival package

An AIP always includes the following digital objects:

accompanying materials

- accompanying materials supplied by the data provider (only: declarations of consent/letters of approval)

---

[10] Two types of metadata can be distinguished from each other:

- bibliographic or administrative metadata = information on the administration of the data, information on the origin of the entirety of the data, of a more general nature, much less community-specific, as well as
- content-describing or subject-specific metadata = description of the data sets as well as additional information, discipline/subject-specific.

(see also https://forschungsdaten.info/themen/beschreiben-und-dokumentieren/metadaten-und-metadatenstandards. Last access: 5.6.2025)

[11] This is the xml file that is uploaded to the DOI registration service – daIra – in order to obtain the DOI for the study.

[12] The syntaxes themselves are versioned via git, both all intermediate steps and the respective final version of a syntax (= the one that forms the basis for which data is released).

- as PDF/A
- as .txt

- accompanying materials to be provided
  - as PDF/A
  - as .txt

data sets

- data set(s) to be provided
  - as .csv
  - as SPSS
  - if applicable as R
  - if applicable as Stata
- original data set(s) from data providers
  - as .csv
  - as SPSS

metadata

- as .xml
- as .csv

syntax/s of the data set(s) to be provided

central documents

- as one PDF/A per document

checksum file

3.2.5 Where are the long-term available digital objects stored? Who has access?

The AIP of the long-term available digital objects are stored in a folder named after ID and the acronym of the study in the archive store. Read access to the archive repository is available to all FDZ at IQB staff, write access is via a separate account to which the data librarian and its deputy have access.

From there, the AIP are transferred to the long-term storage of the CMS and deleted from the direct access of the FDZ at the IQB.

3.2.6 3.2.6 How is the readability of the AIP ensured?

3.2.6.1 Migration and preservation measures[13]

An AIP contains the digital objects in their original and long-term archived formats. The data provided to data user(s) is delivered as .sav file(s). The FDZ at the IQB is aware that SPSS is a

[13] see also here: nestor Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung. Version 2.0; hg. v. H. Neuroth, A. Oßwald,

proprietary software, but it is the one that is predominantly used in the designated community of the FDZ at the IQB. But this problem is also the reason why the FDZ at the IQB prepares the research data in R and is making increased efforts to adapt/change processes.

It is only possible to speak of the long-term availability of digital objects at all if the digital objects (in this case AIP) can be permanently accessed, are permanently readable, and are permanently preserved.

The preservation strategy depends on the significant properties, i.e., the properties of an digital object that must be preserved at all costs. Migration is chosen as the preservation strategy at the FDZ at the IQB.

3.2.6.1.1 Media migration

The AIPs are regularly backed up by the CMS as part of its regular data backup routine, so the CMS ensures data backup (bitstream preservation). That means that the FDZ at the IQB does not need to migrate the digital objects by itself (i.e., no data drive migration). The CMS performs refreshment and replication types. These describe the replacement of single data drives (refreshing) or a change in the storage methods used (replication). No changes to digital data or storage infrastructure take place.

3.2.6.1.2 Format migration

The FDZ at the IQB converts the research data to be made available to the data users (=.sav files) into long-term archivable formats (.csv). This is called a transformation - a migration process that also changes the content data of the archival package. So the long-term archived formats used by the FDZ at the IQB are not dependent on the version of the origin creation software (currently SPSS).

In case there are (major) software version jumps, the corresponding research data are up-dated to the new, latest version.

In the case that the format of the DIPs available until then (so far mainly SPSS) is no longer used, the data is converted into an alternative, different format. This procedure is time-consuming and is only carried out if it can be assumed, by following and observing the technical/technological developments, that the existing research data will no longer be usable in the future due to the format.

Once a year, a check is made for any migration that may become necessary.

In the event of migration, this must be followed by quality assurance, i.e. random checks are carried out to ensure that the digital objects can still be read and interpreted.

3.2.6.2 Verification and monitoring

A likely scenario for data loss is human error. A mechanism is therefore needed to check whether all digital objects are still available unchanged. For this reason, the data librarian checks the AIPs and creates a checksum file (.sha) for each AIP for the digital objects it contains. Before an AIP is transferred to the long-term archiving of the CMS, IQB-IT verifies the checksum file to ensure data

R. Scheffel, S. Strathmann, M. Jehn; verfügbar unter: http://www.langzeitarchivierung.de; Chapter 8 „Digitale Erhaltungsstrategien“ (Version 2.0): urn:nbn:de:0008-20090811378 and https://de.wikipedia.org/wiki/Migration_(Informationstechnik)#Medienmigration

integrity. The long-term archiving of the CMS ensures that the data stored there remains readable (=bitstream).

3.3 Data preparation

The scientific staff of the FDZ at the IQB is responsible for data preparation. The FDZ at the IQB checks whether the data documentation is sufficient for researchers who were not involved in the data collection to analyse the research data. In addition, the FDZ at the IQB randomly checks whether the reported results can be replicated with the submitted research data and whether they are consistent with technical reports and scale documentation. Further checks are performed for plausibility, consistency, and data protection. Metadata is enriched. Finally, the FDZ at the IQB, in close cooperation with the data providers, applies data cleansing measures (e.g. correction of typing errors, addition of variable labels, value labelling, consistent coding of missing values), inconsistencies and errors are reported, necessary corrections are made and missing information is added - all in order to increase data quality and at the same time to obtain maximum information from the original digital objects. This procedure ensures that the digital objects are complete, usable and interpretable.

All work steps are documented in a database. Every single digital object published in the FDZ at the IQB is subject to quality checks.

As part of quality management, the FDZ at the IQB sends an email to the data providers once data preparation is complete to inform them that the research data has been made available and to ask them to check the metadata published on the websites of the FDZ at the IQB and the VerbundFDB.

3.3.1 Documentation

For reasons of traceability, the steps taken to make digital objects permanently available are documented. Metadata contribute to the long-term preservation of digital objects. Of great importance in this context is information about the existing format of the digital objects that the FDZ at the IQB a) received from the data providers, b) releases to the data users and c) archives. This is because a digital object must not only remain readable, but also correctly interpretable. Data sets and documentation are archived in clearly defined, standardized file formats. In addition, syntax and setup files are kept to document changes between different versions. The existing internal documentation is available to all employees of the FDZ at the IQB. All transformations of the digital objects are documented.

All work steps are documented in self-developed database applications, which are only available to authorized employees of the FDZ at the IQB.

The ID of the AIP is documented in the database for each study, as is the date on which the AIP was transferred to the CMS long-term archiving system.

If variables or data sets are added to a study, resulting in a new external version, a new version number is created and this version is also archived for the long term. Each version has its own DOI. A new version receives its own AIP.

3.4 Access

The result of a curation process is a data product, represented by a PID (here: DOI), which ensures the persistent findability and thus the permanent availability of the research data for the scientific community – a standard requirement of the LZA.

The allocation of persistent identifiers also enables research data to be cited, thus increasing the visibility and transparency of data producers.

In detail: Each data product corresponds to a DOI. A data product can contain several data sets - these are often several individual data sets that are related to each other in terms of content (e.g. originate from a study) and are provided together in one or more ZIP files (e.g. for different data formats, i.e. .dta, .sav) - but always have the same underlying anonymisation strategy. This means that for each study there can be both a data product with sensitive data sets (SUF Remote) and a data product without sensitive data sets (SUF Off-site, CUF). A data package now represents a combination of several data products from the same study due to the different granularity in the sensitivity of the data sets. An AIP can consist of one or more data products of a study, depending on how many data products the study has due to the granularity in the sensitivity of the data sets.

As each data version has its own DOI, it allows researchers to replicate published results based on older data versions.

The research data (=DIP) is made available on request to researchers with an academic affiliation via various access routes (off-site, remote), depending on the confidentiality of the data sets and the location of the researchers, for scientific, non-commercial purposes.

Applications for data access are submitted online and include a brief project outline summarizing the theoretical underpinnings, hypotheses, and planned analyses. Application guidelines, a sample project application, and the application form are available online. The FDZ at the IQB reviews applications for compliance with formal criteria for approval. These criteria essentially cover four areas:

- Are the research data to be used for purely non-commercial and scientific purposes?
- Will data protection be respected?
- Do the planned analyses comply with the contractual agreements with the data provider?
- Is it ensured that the planned analyses will not jeopardize ongoing qualification and publication work on the research data? (=Are embargoes affected?)

If the formal criteria are met, a data use agreement is concluded, and only then is access to the research data granted.

Data users agree to the following:

- Use is permitted only for scientific purposes in research and teaching.
- The transfer of rights to use the research data to third parties by data users is not permitted.
- No attempts may be made to re-identify individuals from the data set. In case of accidental re-identification, the FDZ at the IQB must be notified.

- No data from individuals or groups of fewer than five individuals may be reported.
- The research data must be deleted after completion of the project (or at the latest after expiration of the contract period of the data use agreement).

In the event of a breach of the terms of the contract, the right of use expires immediately and data users must pay a contractual penalty of €10,000.

The research data received from the FDZ at the IQB must be destroyed after completion of the analyses for which they were provided.

The conditions of data use include that data users may only store the research data received from the FDZ at the IQB in an access-protected manner as well as on password-protected storage media. The research data may only be transferred to countries that have an adequate level of data protection.

The workflow for preparing and sending the research data to data users via the various access ways is documented in a database that is only available to authorised employees of the FDZ at the IQB.

Data users can find further contextual information and accompanying documentation material on the study on the corresponding landing pages, which are publicly accessible via the DOI.