Institut zur Qualitätsentwicklung im Bildungswesen
Forschungsdatenzentrum
WISSENSCHAFTLICHE EINRICHTUNG DER LÄNDER AN DER HUMBOLDT-UNIVERSITÄT ZU BERLIN E.V.

Concept for the long-term availability of digital data sets at the FDZ at IQB

POSTANSCHRIFT Humboldt-Universität zu Berlin | Institut zur Qualitätsentwicklung im Bildungswesen (FDZ)
Unter den Linden 6 | 10099 Berlin
Tel: +49 [30] 2093 46552 | Fax: +49 [30] 2093 46598
fdz@iqb.hu-berlin.de | www.iqb.hu-berlin.de

Bibliographical information / Please quote as

Forschungsdatenzentrum am Institut zur Qualitätsentwicklung im Bildungswesen (FDZ am IQB) [Research Data Centre at the Institute for Educational Quality Improvement (FDZ at IQB)] (2022). *Concept for the long-term availability of digital data sets at the FDZ at IQB* [Konzept zur Langzeitverfügbarkeit digitaler Datensätze des FDZ am IQB]. Berlin: IQB – Institut zur Qualitätsentwicklung im Bildungswesen. Retrieved from: https://www.iqb.hu-berlin.de/fdz/Grundlagen/Langzeitverfuegb_1.pdf

With the collaboration of: Lisa Pegelow, Claudia Neuendorf and Malte Jansen

Content

1 Terms and definition of long-term archiving 4
2 Technical Infrastructure 5
3 Open Archival Information System (OAIS) 5
3.1 Ingest 6
3.2 Archival Storage 7
3.2.1 Which digital data are made available for the long term? 7
3.2.2 When are archival packages created? 7
3.2.3 Who is responsible for the creation of the archival packages? 8
3.2.4 How are the archival packages created? 8
3.2.5 Where is the long-term available data stored? Who has access? 10
3.2.6 How is the readability of the AIP ensured? 10
3.3 Data preparation 11
3.3.1 Documentation 11
3.4 Access 12

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 data, 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 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."

Digital data consists of a fixed sequence of bits (=bitstream) that are stored on data carriers. In order to keep digital data 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 data, 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 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 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 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)[1], backup, network). It takes care of backups, media monitoring and refreshing. To protect the data, the following backup and access control procedures[2] 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)[3]
- 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 Open Archival Information System (OAIS)

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).

[1] 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

[2] see also here: https://www.cms.hu-berlin.de/de/publikationen/ordnungen/ZutrittCMS and https://www.cms.hu-berlin.de/de/publikationen/ordnungen/cbo_html

[3] see also here: www.cms.hu-berlin.de/de/dl#/svc15 and www.cms.hu-berlin.de/de/dl/systemservice/fileservice/tsm, 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 taken off once a day (01:00)
- all backups are kept for 60 days
- since all backups are additionally stored on tape at the CMS central backup service, all retention times mentioned are increased by another 60 days.

Following the functional model of the Open Archival Information System (OAIS)[4], the FDZ at the IQB is the archive/repository or the link (connector) between data providers and data users.

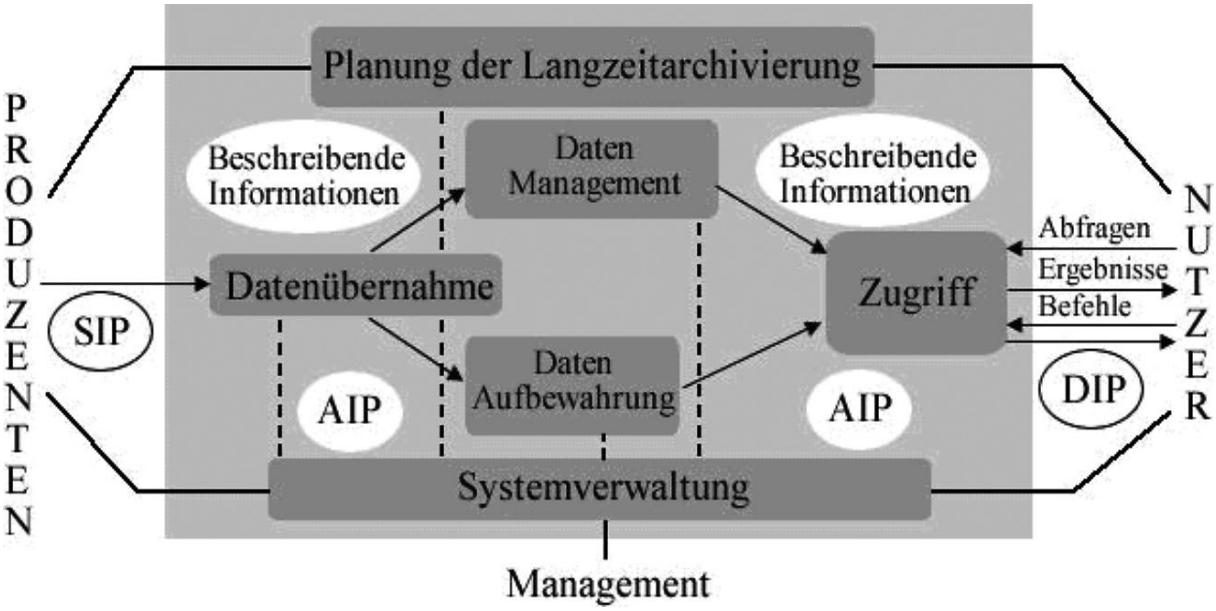

Fig.: OAIS reference model[5]

The data that the FDZ at IQB receives from data providers are considered original data at the FDZ at IQB and, according to OAIS terminology, denote the so-called Submission Information Packages (=SIP). The research data and documentation materials that are finally made available to the data users are referred to as Dissemination Information Package (=DIP). Those data that are archived long-term at the FDZ at IQB are called Archive Information Package (=AIP). The AIP contains all the materials received from the data providers (i.e. the SIP), the materials provided (i.e. the DIP), but also all the documentation and processing steps that were carried out at the FDZ at IQB in order to get from the SIP to the DIP. The AIP of a study thus consists of the SIPs, the DIPs and additionally of the accompanying materials provided; all available data are also available in long-term formats and a checksum file has been created for each AIP.

3.1 Ingest

Ingest is about the transfer of digital data.

When the data are transferred to the FDZ am IQB, a data provision contract is concluded between the FDZ am 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 am IQB and the archiving of the data there, and that all parties involved comply with the DFG regulations on "safeguarding good scientific practice".

[4] see for this: The Consultative Committee for Space Data Systems (SSCDS) (2012). Recommendation for Space Data System Practices. Reference Model for an Open Archival Information System (OAIS). Available online at: https://public.ccsds.org/Pubs/650x0m2.pdf. Last accessed: 24.3.2020.
[5] see: urn:nbn:en:0008-2010061762, p. 11

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

To ensure data quality, the FDZ at IQB has defined minimum requirements that the submitted data must meet. For all incoming data, the FDZ at 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 data comply with the scope of the FDZ at 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 data for secondary analysis, is assessed.

All work steps are documented in a 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 LTA storage facility of the Computer and Media Service (CMS) of the Humboldt University (HU) in 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 data is maintained.

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

The digital data of the studies published at the FDZ at IQB are made digitally available for the long term. The FDZ at the IQB makes data sets 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 data sets are mainly quantitative data.

In concrete terms, this means that the FDZ at IQB archives all digital data from the studies that are made available for the long term. In addition, the original data including the accompanying material are archived in long-term archived formats as well as all documentation and processing steps that were carried out at the FDZ at IQB in order to get from the SIP to the DIP. This includes, for example, processing syntaxes and documents for the evaluation of the data.

3.2.2 When are archival packages created?

An archival package (=AIP) for a study is created 1) after data transfer and the decision that the study should be made available and 2) with a new data version and contains the original data (=SIP) as well as these in long-term available format incl. the accompanying materials. After data

preparation, the archival package is filled with the other corresponding objects (=DIP, accompanying materials to be released, all 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 data provided by the FDZ at 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 data is stored in unencrypted, non-compressed, non-proprietary formats using open, documented standards.

Redundancy in data is useful and sensible in terms of data security, i.e., the FDZ at IQB stores text files in both PDF/A[6] 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.[7]

3.2.4.2 Procedure

At the FDZ at 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) and consists of the original data (=SIP), the converted original data, the provided and made available (research) data (=DIP) as well as the accompanying materials available in long-term available formats and the checksum file per AIP.

In concrete terms, this means that after receipt of the original data and the decision to make them available, they are immediately converted into long-term archivable formats, i.e. files 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 as well as 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 data versions.

After the data preparation the provided data (=DIP), the accompanying materials to be issued (all in long-term available formats) are stored in the AIP.

For each AIP, a checksum file is created to ensure data integrity.

[6] 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.

[7] https://www.loc.gov/preservation/digital/formats/fdd/fdd000468.shtml; last access: 03.07.2020

For each study and version, the associated administrative (e.g.: DOI, citation proposal, data provider, access regulations, rights) as well as descriptive (e.g.: study period, survey period, selection procedure) metadata[8] (in .xml and .csv format) are stored in the AIP, because the metadata are also made long-term available. The FDZ at IQB does NOT record variable or value labels as metadata. For each study as well as for a new version of the data package, the metadata is included in the archival package as a corresponding .xml file[9].

After data preparation, the archival package is filled with further objects: in addition to the provided data sets incl. the syntaxes[10] and the accompanying materials, these are central documents such as checklists, evaluations, contracts, the correspondence with the data providers for the preparation of the data, an access concept as well as relevant documentation - in each case in long-term available formats.

3.2.4.2.1 Archival package

An archival package always includes:

- Accompanying materials
  - Accompanying materials supplied by the data provider (only: declarations of consent/letters of approval)
    - as PDF/A
    - as .txt
  - Accompanying materials provided
    - as PDF/A
    - as .txt
- data sets
  - provided data set(s)
    - as .csv
    - as SPSS
    - if applicable as R
    - if applicable as Stata
  - Original data set(s)
    - as .csv
    - as SPSS
- Metadata
  - as .xml
  - as .csv
- Syntax/s of the provided data set(s)
- Central documents

[8] 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://www.forschungsdaten.info/themen/aufbereiten-und-veroeffentlichen/metadaten-und-metadatenstandards/)

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

[10] 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 one PDF/A per document
- ➢ Checksum file

3.2.5 Where is the long-term available data stored? Who has access?

The archival packages of the long-term available data 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 IQB.

3.2.6 How is the readability of the AIP ensured?

3.2.6.1 Migration and preservation measures

An archival package (AIP) contains the original data of the data depositor(s) (SIP), the data which is provided to data user(s) (DIP), and the data in long-term archived formats. The data provided to data user(s) is delivered as .sav file(s). We are aware that SPSS is a proprietary software, but it is the one that is predominantly used in the scientific community. But this problem is also the reason why the FDZ at IQB is switching to R and is making increased efforts to adapt/change the processes accordingly.

It is only possible to speak of the long-term availability of digital data at all if the digital data (in this case archival packages) 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 object that must be preserved at all costs. Migration is chosen as the preservation strategy at the FDZ at IQB.

3.2.6.1.1 Media migration

The archival packages 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 IQB does not need to migrate the data files 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 data or storage infrastructure take place.

3.2.6.1.2 Format migration

The FDZ at IQB converts the data provided to data user(s) (=.sav files) into long-term archived formats (.csv, .pdf/A, .txt). 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 IQB are not dependent on the version of the origin creation software (currently SPSS).

In case there are (major) software version jumps, the corresponding digital 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 long-term available digital documents will no longer be usable in the future due to their obsolete 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 (Over)checking / Control

A likely scenario for data loss is human error. A mechanism is therefore needed to check whether all data is still available unchanged. For this reason, the data librarian checks the AIPs and creates a checksum file (.sha256) for each AIP for the objects/files it contains. Before an AIP is transferred to the long-term archiving of the CMS, IQB-IT verifies the checksum file to ensure data 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 IQB is responsible for data preparation. The FDZ at IQB checks whether the data documentation is sufficient for researchers who were not involved in the data collection to analyse the data sets. In addition, the FDZ at IQB randomly checks whether the reported results can be replicated with the submitted data and whether the data are consistent with technical reports and scale documentation. Further checks are performed for plausibility, consistency, and data protection. Metadata are enriched. Finally, the FDZ at IQB, in close cooperation with the data providers, applies data cleaning measures (e.g., correcting typos, adding variable labels, value labels, and missing categories), inconsistencies and errors are reported, necessary corrections are made, and missing information is added, all to increase data quality as well as to preserve maximum information from the original data. This procedure ensures that the data is complete, usable and interpretable.

All steps in the process are documented in a database. Every single digital resource published in the FDZ at IQB is subject to quality control.

As part of the quality management process, the FDZ at IQB also sends an email to data providers after data preparation is complete to inform them that the data have been made available and to ask them to review the metadata published on the FDZ at IQB and VerbundFDB websites.

3.3.1 Documentation

For reasons of traceability, the steps taken to make digital data 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 data that the FDZ at 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 and documentation are archived in clearly defined, standardized file formats. In addition, syntax and

setup files are kept to document changes between different data versions. The existing internal documentation is available to all employees. All transformations of the data are documented.

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

For each study, the ID of the archival package is documented in the database and when the archival package was transferred to the CMS long-term archiving.

If variables or data sets are added to the data package that result in a new external version, a new version number is created and this data version is also long-term archived. Each data version has its own DOI. Each data package with a DOI receives its own AIP, i.e., new data versions have their own AIP.

3.4 Access

The FDZ at IQB assigns a Digital Object Identifier (=DOI) to each data package provided, thus ensuring the permanent availability of the research data provided to the scientific community. Likewise, the assignment of persistent identifiers enables the citeability of research data and thus also increases the visibility and transparency of the data providers.

Each data version has its own DOI, which allows researchers to replicate published results based on older data versions.

The data sets (=DIP) are made available upon request to researchers who have an academic affiliation, in various formats (SUF, ICA) depending on the confidentiality of the data sets as well as 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 IQB reviews applications for compliance with formal criteria for approval. These criteria essentially cover four areas:

- Are the 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 data? (=Are embargoes affected?)

If the formal criteria are met, a data use agreement is concluded, and only then is access to the 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 data to third parties by the data user 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 IQB must be notified.
- No data from individuals or groups of fewer than five individuals may be reported.
- The 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 the data user must pay a contractual penalty of €10,000.

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

The workflow for processing and sending the data sets in different formats to data users is documented in a database, which is only available to authorized employees of the FDZ at IQB.