IQB
Forschungsdatenzentrum

Versioning

As of: 10/2025

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

Forschungsdatenzentrum am Institut zur Qualitätsentwicklung im Bildungswesen (FDZ am IQB) [Research Data Centre at the Institute for Educational Quality Improvement (FDZ at the IQB)] (2025). *Versioning.* Berlin: IQB - Institut zur Qualitätsentwicklung im Bildungswesen. http://doi.org/10.5159/IQB_Versioning_v1

Content

1. Introdution

Access to study data may vary depending on the sensitivity of the data due to differences in granularity. The following access options are available at the FDZ at the IQB:

- SUF via Off-site (data product without sensitive data)
- SUF via Remote (datenprodukt with sensitive data)
- Campus Use File (CUF) via Download

As a result of data curation (= preparation), each study yields at least one and at most three data products. The FDZ at the IQB assigns a persistent identifier (= DOI) at the data product level. This means that There is one DOI per data product.

The resulting data products also depend on the curation level:

- At curation level *Standard Plus*, the only data product is a SUF Off-site.
- At curation level *Fokus*, the data product SUF Remote can be added.
- At curation level *Fokus Plus*, the data product SUF Remote is always added.
- The data product CUF can be created at any curation level.

In summary, this means that a data product is the result of a curation process. A data product consists of at least one data set, but can also contain several data sets. It has a persistent identifier and is passed on to users as a whole. The data products of a study differ in terms of the sensitivity/granularity of the data and, as a result, in terms of access restrictions. Multiple data products (in our case, always related to the same study) result in a data package.

The FDZ at the IQB distinguishes between two forms of change after ingest:

- Content change to an existing data product: Data (values of variables) in one or more datasets of the data product are substantially changed or data is added to the data product in one or more datasets → new external version of the data product → new DOI
- Administrative change to an existing data product: Change to metadata (including metadata in datasets, e.g. value labels), descriptive documents or supplementary files → internal version of the data product → no new DOI

We proceed on a data set basis: changes always only affect the individual data set and not the other data sets of the data product. Only the version information of the edited data sets changes. This means that in the case of a data product with several data sets, not all data sets may have the same version information.

For each change in an individual data record, the internal version variable is changed and the date of the change is specified. This is an internal data record process. The data record-related change is identified at the data record level (see naming convention).

2. New external version of a data product

When a new external version of a data product is created, xy recreates all descriptive and structural metadata, a new version number is generated, and this data version is also archived for the long term[1].

The new data product is assigned a new DOI. This means that each external data version has its own DOI. In this way, the existing DOI continues to refer uniquely to the previous version of the data product. All versions of a data product are cross-referenced in their respective descriptive metadata. New data versions of a study have their own AIP. The previous data version(s) and the associated previous Archival Information Package(s) (AIP) remain in long-term archiving. From the perspective of the FDZ at the IQB, the AIP represents a data package.

2.1 Criteria when a new DOI version is required

If analyzes and replications are different from the current / previous version, a new external DOI version of the data product is needed.

- Changes in the variables (e.g. changing the IDs to make it possible to link to other data sets / studies; rescaling testcores, etc.)
- Deleting and / or adding variables
- Adding a recoding key
- Deleting or adding data set(s) to a study (e. g. adding new survey waves in longitudinal studies such as BiKS, BilWiss, BRISE)
- Changes to the sample (exclusion or addition of persons)

2.2 Naming convention

Data product

The version number **of the entire data product** is incremented by 1.

Data set(s)

- **Only the version number of the edited data sets** is increased by 1.
- If one or more data sets in a data product need to be modified in terms of content, these modified data sets are given the version identifier v[No.+1 to the previous version]. However, the version numbers and variables of the other data sets in the data product do not change.

[1] https://fdz.iqb.hu-berlin.de/documents/106/Langzeitverfuegbarkeit_externe_Version_english_Stand-2025-Juli.pdf

- If one or more new data sets are added to a data product, the corresponding data sets are given the version identifier v1. However, the version numbers and variables of the other data sets in the data product do not change.
- A version variable is inserted into the corresponding data set or the existing version variable is updated.

2.3 Storage

- Versioned data sets are stored in the corresponding study.
- Previous versions of the corresponding data sets are moved to the archive in the corresponding study.

3. New internal version

Alternatively, administrative changes are documented in the administrative metadata. No new DOI is assigned. The internal version variable is used to track which data set version was used in users' analyses when they make requests. Since the internal version variable is assigned to the data set rather than the data product, this increases reproducibility and traceability for user requests.

Internal versions are also archived for the long term.

3.1 Criteria for when a new internal version is created

Administrative changes to a data set require an internal version for that data set.

- Spelling corrections in the data set (not in the file name)
- Additions or changes to value and variable labels
- Adjustments to the scale level, format, etc. of the variables in the data set

3.2 Naming convention

The version information in the file name of an edited data record is structured as follows: v[No]_[internal previous version+1]

3.3 Storage

- Versioned data sets are stored in the corresponding study.
- Previous versions of the corresponding data sets are moved to the archive in the corresponding study.

4. With regard to long-term preservation

Both new external and internal versions are preserved long-term, i.e. new versions of data products are stored in a new AIP. If there are several data products per study, these are stored together in one AIP, as an AIP is always created at the study level.

The conversion into long-term available formats aims to preserve the content of the data, i.e. to keep it readable and interpretable over time, as these are essential properties of the data. The preservation of other aspects, such as the layout of the input format (the 'look and feel'), is considered less important.

In the event of data conversion to another file format for preservation or access purposes, the original file(s) are retained.