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

# Konzept zur Langzeitverfügbarkeit digitaler Datensätze des FDZ am 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

KULTUSMINISTER KONFERENZ
HUMBOLDT-UNIVERSITÄT ZU BERLIN

**Bibliographische Informationen / Bitte zitieren als:**

Forschungsdatenzentrum am Institut zur Qualitätsentwicklung im Bildungswesen (FDZ am IQB) (2021). *Konzept zur Langzeitverfügbarkeit digitaler Datensätze des FDZ am IQB.* Berlin: IQB - Institut zur Qualitätsentwicklung im Bildungswesen. Verfügbar unter: https://www.iqb.hu-berlin.de/fdz/Grundlagen/Langzeitverfuegb.pdf

Unter Mitarbeit von: Lisa Pegelow, Claudia Neuendorf und Malte Jansen



# Inhalt

# 1 Begriffe und Definition zur Langzeitarchivierung

Unter Langzeitarchivierung wird die langfristige Aufbewahrung und der Erhalt der dauerhaften Verfügbarkeit von Informationen verstanden. Vor allem im Digitalen ist dies von nicht unerheblicher Relevanz, denn bei der analogen Archivierung liegt der Schwerpunkt auf der Erhaltung des physischen Mediums; bei der Langzeitarchivierung digitaler Daten hingegen ist es nicht notwendig, das Medium selbst zu erhalten, sondern durch rechtzeitige Kopierung Datenverluste zu verhindern. Langzeitarchivierung benötigt Strategien, um mit dem Wandel der Arten von Speichermedien, von Dateiformaten usw. umzugehen. Das Archivierungsnetzwerk nestor[1] definiert Langzeitarchivierung wie folgt:

> „Langzeit bedeutet für die Bestandserhaltung digitaler Ressourcen […] die verantwortliche Entwicklung von Strategien, die den beständigen, vom Informationsmarkt verursachten Wandel bewältigen können. […] Vielmehr schließt es die Erhaltung der dauerhaften Verfügbarkeit und damit eine Nachnutzung […] der digitalen Ressourcen mit ein.“[2]

Digitale Daten bestehen aus einer festgelegten Abfolge von Bits (=Bitstream), die auf Datenträgern gespeichert werden. Um digitale Daten dauerhaft verfügbar zu halten, müssen bestandserhaltene Maßnahmen getroffen werden (= **Bitstream Preservation**). Die Lebensdauer und Zuverlässigkeit von Datenträgern jedoch ist begrenzt und einzelne Bits können mit der Zeit nicht mehr oder verfälscht gelesen werden. Die Bitstream Preservation stellt durch technische Maßnahmen wie Prüfsummen und Redundanz sicher, dass der Bitstream über längere Zeiträume, auch nach Technologiewechseln, unverändert erhalten bleibt. Diese Art der Datensicherung ist eine elementare Maßnahme zur Datensicherheit; dabei es geht um den reinen physischen Erhalt der Daten und deren Lesbarkeit. Bitstream Preservation bildet eine Grundvoraussetzung für die digitale Langzeitarchivierung.

**Backup**s sind Teil der Bitstream Preservation. Eine sinnvolle Backupstrategie ist, mehrere Kopien des Backups redundant an verschiedenen Orten, auf verschiedenen Speichermedien verteilt zu archivieren. Beim Backup handelt es sich also um eine reine Datensicherung, bedeutet das Kopieren von Daten, um im Falle eines Datenverlustes die Daten wieder zurückkopieren zu werden können. Damit ist noch nichts über die Lesbarkeit oder Verwendungsmöglichkeit gesagt.

Um die dauerhaft erhaltenen digitalen Daten auch nachnutzen zu können, müssen sie interpretierbar bleiben und das ist der Kern der Langzeitarchivierung, was auch oft als digitale Kuratierung bezeichnet wird. **Digitale Kuratierung** bedeutet, digitale Forschungsdaten während ihres gesamten Lebenszyklus zu pflegen, zu erhalten und anzureichern. Diese Maßnahmen gehen über die reine technische Erhaltung des Bitstreams hinaus und erforderlich auch fachlich-inhaltliche Expertise. Beispielsweise geht es darum, dass die Dateien auch für neue im Feld eingesetzte Software lesbar bleiben. Für das Forschungsdatenmanagement heißt das dann auch,

[1] = deutsches Netzwerk für die Langzeitarchivierung und Langzeitverfügbarkeit digitaler Ressourcen
[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: nestor.sub.uni-goettingen.de/handbuch/nestor-handbuch_23.pdf, Kap. 1:3. Letzter Zugriff: 28.09.2021

dass der langfristige Forschungswert erhalten bleibt und das Risiko der digitalen Veralterung[3] gemindert wird.

Die Daten müssen nicht nur aufbewahrt und erhalten (kuratiert) werden, sondern müssen, um ihren Mehrwert zu entfalten, von der Scientific Community genutzt werden können. Dies kann durch vertrauenswürdige Datenzentren ermöglicht werden.[4] Das Forschungsdatenzentrum am IQB ist ein solches vertrauenswürdiges Datenzentrum. D. h., dass das FDZ am IQB Maßnahmen ergreift, um die Kuratierung und Erhaltung der digitalen Forschungsdaten während des gesamten Lebenszyklus zu ermöglichen sowie die digitalen Forschungsdaten am FDZ am IQB auf Dauer sicher zu archivieren und verfügbar zu machen.

Die **Langzeitarchivierung** ist damit von der reinen Datensicherung abzugrenzen.

[3] Mit der Zeit ändern sich die Gegebenheiten der verwendeten Software; neue Techniken, Anforderungen und Formate entstehen. Eine mangelnde Anpassung an diese Gegebenheiten kann zu Problemen, Fehlern oder einfach zu einem verringerten Nutzen des Programmes führen, dem wirkt die Langzeitarchivierung entgegen und daher sollten auch offene (open source) Formate als langzeitarchivierbare Formate genutzt werden.

[4] s. dazu auch http://www.dcc.ac.uk/digital-curation/what-digital-curation. Letzter Zugriff: 14.01.2020

## 2 Schaubild Langzeitarchivierung


erstellt von: Lisa Pegelow

# 3 Technische Infrastruktur

Das IQB und damit auch das FDZ am IQB greift auf die Infrastruktur des Rechenzentrums der Humboldt-Universität zu Berlin, auf den Computer- und Medienservice (CMS), zurück.

Das CMS ist im Falle von Speicherdiensten für die IQB-Dienste, seine Server und die zugrunde liegende Infrastruktur (Storage Area Network (SAN)[5], Backup, Netzwerk) verantwortlich. Es kümmert sich um Backups, Medienüberwachung und Refreshing. Zum Schutz der Daten werden folgende Backup- und Zugriffskontrollverfahren[6] eingesetzt, um die (physische) Sicherheit der digitalen Archivbestände zu gewährleisten:

- Rechenzentrum und Serverräume sind gegen unbefugten Zugriff gesichert
- Rauch- und Wassermelder sind installiert
- Temperaturen in den Serverräumen werden überwacht
- Redundante Datenhaltung an verschiedenen Standorten (Adlershof, Mitte)
- Häufige inkrementelle und vollständige Backups (Jede Nacht wird der Inhalt der IQB-Netzlaufwerke gebackupt, 60 Tage Aufbewahrungsfrist, 2 Wochen im Zugriff)[7]
- Vielfalt der Speichermedien und häufige Medienauffrischung

Auch für den Fall eines technischen Versagens, einer böswilligen Handlung oder eines menschlichen Fehlers verfügt das CMS über Spiegel der Server. Diese speichern die letzten 2 Wochen.

# 4 Open Archival Information System (OAIS)

Das FDZ am IQB arbeitet entlang definierter Workflows – von der Datenauswahl (gemäß der Collection Policy des FDZ am IQB) und dem Ingest (inkl. Prüfung und Validierung) über den Vertragsschluss und die Datenaufbereitung (inkl. Feinprüfung, Tiefenprüfung, Metadatenanreicherung und Kommunikation mit Datengebenden) bis hin zur Dokumentation und Datenverfügbarkeit (inkl. Veröffentlichung und Freigabe).

---

[5] Redundantes Netzwerk an Severn zur Bereitstellung ausfallsicherer Festplattenspeicher. Ein Netz, das Server und Speichersystem über dedizierte Leitungen miteinander verbindet. Strukturell ist ein SAN analog aufgebaut wie ein Local Area Network (LAN): es gibt Hubs, Switches und Router.
s. auch hierzu: www.cms.hu-berlin.de/de/dl#/svc14

[6] s. auch hier: https://www.cms.hu-berlin.de/de/publikationen/ordnungen/ZutrittCMS und https://www.cms.hu-berlin.de/de/publikationen/ordnungen/cbo_html

[7] s. auch hier: www.cms.hu-berlin.de/de/dl#/svc15 und www.cms.hu-berlin.de/de/dl/systemservice/fileservice/tsm, sowie zum Thema Backups: Erstellung und Aufbewahrungsdauer

**PostgreSQL:**

- alle Datenbanken werden zweimal täglich (01:00 Uhr und 13:00 Uhr) abgezogen
- die mittäglichen Backups werden 7 Tage lang aufbewahrt
- die nächtlichen Backups werden 60 Tage lang aufbewahrt
- da alle Backups zusätzlich beim zentralen Backup-Service des CMS auf Band gespeichert werden, erhöhen sich alle genannten Aufbewahrungszeiten nochmal um 60 Tage

**MySQL:**

- alle Datenbanken werden einmal täglich (01:00 Uhr) abgezogen
- alle Backups werden 60 Tage lang aufbewahrt
- da alle Backups zusätzlich beim zentralen Backup-Service des CMS auf Band gespeichert werden, erhöhen sich alle genannten Aufbewahrungszeiten nochmal um 60 Tage

In Anlehnung an das Funktionsmodell des Offenen Archivinformationssystems (OAIS)[8] ist das FDZ am IQB das Archiv/Repositorium bzw. die Verbindung (Konnektor) zwischen Datengebenden und Datennutzenden.

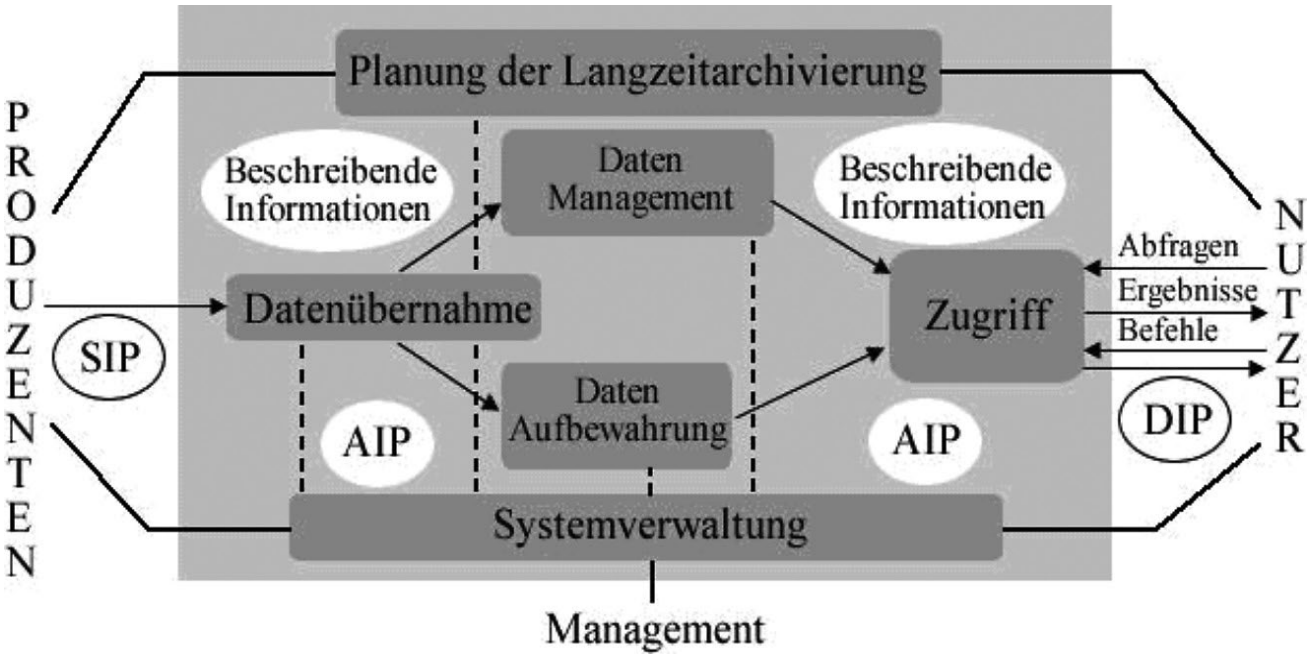


*Abb.: OAIS-Referenzmodell*[9]

Die Daten, die das FDZ am IQB von Datengebenden erhält, werden am FDZ am IQB als Originaldaten betrachtet und bezeichnen nach der OAIS-Terminologie die sog. Submissionsinformationspakete (=SIP). Die Forschungsdaten und Dokumentationsmaterialien, die schließlich für die Datennutzenden bereitgestellt werden, werden als Dissemination Information Package (=DIP) bezeichnet. Diejenigen Daten, die am FDZ am IQB langzeitarchiviert werden, werden als Archivinformationspaket (=AIP) bezeichnet. Das AIP enthält alle von den Datengebenden erhaltenen Materialien (also das SIP), die Materialien, die bereitgestellt werden (also das DIP), aber auch alle Dokumentations- und Aufbereitungsschritte, die am FDZ am IQB durchgeführt wurden, um vom SIP zum DIP zu kommen. Das AIP einer Studie setzt sich also aus den SIPs, den DIPs und zusätzlich aus den bereitgestellten Begleitmaterialien zusammen; ebenso sind alle vorliegenden Daten in langzeitverfügbaren Formaten vorliegend sowie wurde pro AIP eine Checksumme gebildet.

## 4.1 Datenübernahme: Ingest

Im Bereich Ingest geht es um die Übernahme der digitalen Daten.

Bei der Übergabe der Daten an das FDZ am IQB wird ein Datenbereitstellungsvertrag zwischen dem FDZ am IQB und den Datengebenden geschlossen; dieser behandelt und regelt alle rechtlichen Fragen. So versichert der Datengebende, dass Rechte Dritter durch die Weitergabe

[8] s. dazu: The Consultative Committee for Space Data Systems (SSCDS) (2012). Recommendation for Space Data System Practices. Reference Model for an Open Archival Information System (OAIS). Online verfügbar unter: https://public.ccsds.org/Pubs/650x0m2.pdf. Letzter Zugriff: 24.3.2020.
[9] s. dazu: urn:nbn:de:0008-2010061762, S. 11

ans FDZ am IQB und mit einer dortigen Archivierung der Daten nicht beeinträchtigt werden sowie alle Beteiligten die DFG-Regelung zur "Sicherung guter wissenschaftlicher Praxis" einhalten.

Das FDZ am IQB-Team und die Datengebenden vereinbaren ein Zugriffskonzept, welches regelt, welche Daten und Begleitmaterialien wie/auf welche Weise an Dritte weitergegeben werden (öffentlich auf der Website des FDZ am IQB oder standardmäßig nur auf Anfrage zusammen mit den Daten oder nie).

Um die Datenqualität zu gewährleisten, hat das FDZ im IQB Mindestanforderungen definiert, die die eingereichten Daten erfüllen müssen. Bei allen eingehenden Daten prüft das FDZ im IQB, ob das gelieferte Material vollständig, korrekt und in einem geeigneten Zustand (lesbar, virenfrei etc.) ist (=technisch-formale Prüfung). In der sich anschließenden inhaltlichen Bewertung wird geprüft, ob die Daten dem Scope des FDZ am IQB, der Collection Policy entsprechen sowie die Mindestanforderungen (=Prüfung, ob alle Datensätze sowohl den technischen und dokumentarischen Anforderungen (=Klarheit von Datensatz und Dokumentation) als auch den rechtlichen Anforderungen (=Einhaltung von Datenschutz, Urheberrecht(en), (Gesetz)) entsprechen) erfüllt sind. Zusätzlich wird das Nachnutzungspotenzial, sprich das Potenzial der Daten für die Sekundäranalyse, bewertet.

Alle Arbeitsschritte werden in einer Datenbank (=Metadaten-Management- und -Dokumentations-Tool) dokumentiert, die nur autorisierten Mitarbeitenden zur Verfügung steht. Zur Vereinheitlichung der Abläufe wird der gesamte Ingest-Prozess durch FDZ-interne Dokumente, Checklisten und Vorlagen gesteuert, die regelmäßig aktualisiert werden.

## 4.2 Datenaufbewahrung: Archival Storage

Die Datenaufbewahrung umfasst den digitalen Archivspeicher, seine Organisation sowie seinen Aufbau im engeren Sinne. Hier geht um das Prozedere bei der Erstellung von Archivpaketen. Im Archivspeicher des FDZ am IQB werden die AIPs so zusammengestellt, dass deren Wiederauffindbarkeit sowie deren Lesbarkeit und Interpretierbarkeit gewahrt ist. Dann werden die AIPs in den LZA-Speicher des Computer- und Medienservice (CMS) der Humboldt-Universität (HU) zu Berlin überführt und dort abgespeichert. Es gelten die dortigen weiteren Regeln, u. a. werden dort die Checksummen der AIPs regelmäßig überprüft (bitstream), so dass die Integrität und Authentizität der Daten gewahrt bleibt.

### 4.2.1 Welche digitalen Daten werden langzeitverfügbar gemacht?

Es werden die digitalen Daten der am FDZ am IQB veröffentlichten Studien digital langzeitverfügbar gemacht. Das FDZ am IQB stellt Datensätze von nationalen und internationalen Schulleistungsstudien sowie von nationalen Studien mit Kompetenzmessungen im Bildungsbereich der Scientific Community für Re- und Sekundäranalysen zur Verfügung. Bei den vorliegenden Datensätzen handelt es sich bisher im Wesentlichen um quantitative Daten.

Konkret heißt das, dass das FDZ am IQB alle digitalen Daten der Studien, die bereitgestellt werden, langzeitarchiviert. Zusätzlich werden die Originaldaten inkl. des Begleitmaterials in langzeitarchivierten Formaten archiviert sowie alle Dokumentations- und Aufbereitungsschritte, die am FDZ am IQB durchgeführt wurden, um vom SIP zum DIP zu kommen. Das umfasst bspw. Aufbereitungssyntaxen und Dokumente zur Bewertung der Daten.

### 4.2.2 Wann werden die Archivpakete erstellt?

Ein Archivpaket (=AIP) für eine Studie wird 1) nach Datenübernahme und der Entscheidung, dass die Studie bereitgestellt werden soll, und 2) bei einer neuen Datenversion angelegt und enthält die Originaldaten (=SIP) sowie diese in langzeitverfügbarem Format inkl. der Begleitmaterialien. Nach der Datenaufbereitung wird das Archivpaket mit den weiteren entsprechenden Objekten (=DIP, herauszugebende Begleitmaterialien, alles in langzeitverfügbaren Formaten) befüllt.

### 4.2.3 Wer ist verantwortlich für die Erstellung der Archivpakete?

Die wissenschaftliche Dokumentarin (data librarian) ist verantwortlich für die Erstellung der Archivpakete.

### 4.2.4 Wie erfolgt die Erstellung der Archivpakete?

Ziel ist es, dass die vom FDZ am IQB zur Verfügung gestellten Daten auf Dauer zum einen verfügbar und lesbar, sprich funktionsfähig und nutzbar sind und es bleiben, und zum anderen inhaltlich interpretierbar sind und bleiben.

#### *4.2.4.1 Datenformate*

Im Rahmen der Langzeitverfügbarkeit werden die digitalen Daten in unverschlüsselten, nicht komprimierten, nicht proprietären Formaten, sondern mit offenen, dokumentierten Standards gespeichert.

Redundanz bei den Daten ist hinsichtlich der Datensicherheit nützlich und sinnvoll, d. h., das FDZ am IQB speichert Textdateien sowohl im PDF/A[10]- als auch im .txt- Format sowie Tabellendaten im .csv-Format. Die .txt-Dateien werden nur für den Fall abgespeichert, dass das vorrangige PDF/A-Format doch einmal nicht mehr funktionstüchtig sein sollte. Daher werden Textdateien zusätzlich in .txt-Formate umgewandelt, mehr wird damit jedoch nicht gemacht.

Im SPSS portable-Format speichern wir nicht, da es bspw. von der Library of Congress nicht als Archiv-Format empfohlen wird[11].

#### *4.2.4.2 Vorgehen*

Am FDZ am IQB wird zur Erstellung von langzeitverfügbaren Daten je Studie ein sog. Archivpaket erstellt. Ein Archivpaket ist das sog. Archival Information Package (AIP) und setzt sich zusammen aus den Originaldaten (=SIP), den konvertierten Originaldaten, den bereitgestellten und zur Verfügung gestellten (Forschungs-)Daten (=DIP) sowie den in langzeitverfügbaren Formaten vorliegenden Begleitmaterialien und der Checksummendatei pro AIP.

Im Konkreten bedeutet das, dass nach Eingang der Originaldaten und dem Beschluss, dass diese bereitgestellt werden sollen, diese sofort in langzeitarchivierbare Formate gebracht werden, sprich Dateien im SPSS- und/oder im .xls-Format werden ins .csv-Format umgewandelt, Dateien, die in R vorliegen, werden identisch im AIP der Studie abgespeichert, ggf. mit den

[10] PDF/A ist ein Dateiformat zur Langzeitarchivierung digitaler Dokumente, das von der International Organization for Standardization (ISO) als Teilmenge des Portable Document Format (PDF) genormt wurde. Standardisierte Metadaten sind im Dokument eingebettet. PDF/A-Dokumente unterstützen Volltextsuche und sind selbständig und unabhängig: Elemente (Schriften, Farbprofile usw.), die für eine einwandfreie Wiedergabe benötigt werden, sind im Dokument enthalten. Ein PDF/A-Dokument darf keine Verweise zu externen Quellen aufweisen. Einfache informative Verweise wie zum Beispiel Links zu Webseiten sind erlaubt. PDF/A spart Speicherplatz. PDF/A-Dokumente bleiben fristlos gültig.

[11] https://www.loc.gov/preservation/digital/formats/fdd/fdd000468.shtml; Online-Zugriff: 03.07.2020

entsprechende R-Paketen sowie Begleitdokumentation (Text) in PDF/A- und ergänzend als .txt-Format und im AIP der Studie abgespeichert. Gleiches gilt für das Vorgehen bei neuen Datenversionen.

Nach der Datenaufbereitung werden im AIP die bereitgestellten Daten (=DIP), die herauszugebenden Begleitmaterialien (alle in langzeitverfügbaren Formaten) abgespeichert.

Für jedes AIP wird eine Checksummendatei über alle darin enthaltenen digitalen Objekte zur Sicherung der Datenintegrität erstellt.

Pro Studie und Version werden im AIP die dazugehörigen administrativen (z. B.: DOI, Zitationsvorschlag, Datengebende, Zugriffsbestimmungen, Rechte) sowie deskriptiven (z. B.: Studienzeitraum, Erhebungszeitraum, Auswahlverfahren) Metadaten[12] (im .xml- und .csv-Format) abgespeichert, denn auch die Metadaten werden ebenfalls langzeitverfügbar gemacht. Das FDZ am IQB erfasst als Metadaten NICHT Variablen- oder Wertelabel. Pro Studie sowie bei einer neuen Version des Datenpaketes werden die Metadaten als entsprechende .xml-Datei[13] mit in das Archivpaket mit aufgenommen.

Nach der Datenaufbereitung wird das Archivpaket mit weiteren Objekten befüllt: neben den bereitgestellten Datensätzen inkl. der Syntaxen[14] und den Begleitmaterialien sind das zentrale Dokumente wie Checklisten, Bewertungen, Verträge, die Korrespondenz mit den Datengebenden zur Aufbereitung der Daten, ein Zugriffskonzept sowie relevante Dokumentation – jeweils in langzeitverfügbaren Formaten.

#### 4.2.4.2.1 Archivpaket

Ein Archivpaket umfasst immer:

- Begleitmaterialien
  - vom Datengebenden mitgelieferte Begleitmaterialien (nur: Einwilligungserklärungen/Genehmigungsschreiben)
    - als PDF/A
    - als .txt
  - bereitgestellte Begleitmaterialien
    - als PDF/A
    - als .txt
- Datensätze
  - bereitgestellte/n Datensatz/-sätze
    - als .csv

[12] Zwei Arten von Metadaten sind voneinander zu unterscheiden:

- bibliographische bzw. administrative Metadaten =Informationen zur Verwaltung der Daten, Informationen zur Entstehung der Gesamtheit der Daten, allgemeinerer Natur, weitaus weniger community-spezifisch sowie
- inhaltsbeschreibende bzw. fachliche Metadaten = Beschreibung der Datensätze sowie zusätzliche Informationen, disziplin-/fachspezifisch

(s. dazu auch https://www.forschungsdaten.info/themen/aufbereiten-und-veroeffentlichen/metadaten-und-metadatenstandards/)

[13] Hierbei handelt es sich um die xml-Datei, die beim Registrierungsservice für DOIs – dalra – hochgeladen wird, um damit die DOI für die Studie zu erhalten.

[14] Die Syntaxen selbst werden über git versioniert, sowohl alle Zwischenschritte als auch die jeweils finale Version einer Syntax (=die, die die Grundlage bildet, welche Daten herausgegeben werden).

        - als SPSS
        - ggf. als R
        - ggf. als Stata
    - Originaldatensatz/-sätze
        - als .csv
        - als SPSS
- Metadaten
    - als .xml
    - als .csv
- Syntax/en der bereitgestellten Datensätze
- Zentrale Dokumente
    - jeweils als ein PDF/A je Dokument
- Checksummendatei

### 4.2.5 Wo werden die langzeitverfügbaren Daten aufbewahrt? Wer hat Zugriff?

Die Archivpakete der langzeitverfügbaren Daten werden in einem nach ID und dem Akronym der Studie benannten Ordner im Archivspeicher abgelegt. Lesenden Zugriff auf den Archivspeicher haben alle FDZ am IQB-Mitarbeitenden, schreibender Zugriff erfolgt über einen separaten Account, auf den die wissenschaftliche Dokumentarin (data librarian) und deren Vertretung Zugriff hat.

Von dort werden die AIP in den Langzeit-Speicher des CMS gegeben und aus dem direkten Zugriff des FDZ am IQB gelöscht.

### 4.2.6 Wie wird die Lesbarkeit der Daten im AIP gewährleistet?

#### *4.2.6.1 Migration*[15]

Ein Archivierungspaket (AIP) enthält die Originaldaten der Datengebenden (SIP), die Daten, die den Datennutzenden zur Verfügung gestellt werden (DIP) und die Daten in langzeitarchivierten Formaten. Die den Datennutzenden zur Verfügung gestellten Daten werden als .sav-Datei(en) geliefert. Wir sind uns bewusst, dass SPSS eine proprietäre Software ist, aber es ist diejenige, die in der wissenschaftlichen Gemeinschaft am häufigsten verwendet wird. Dieses Problem ist aber auch der Grund, warum das FDZ am IQB auf R umsteigt und sich verstärkt bemüht, die Prozesse entsprechend anzupassen/zu verändern.

Es kann überhaupt nur dann von einer Langzeitverfügbarkeit digitaler Daten gesprochen werden, wenn dauerhaft auf die digitalen Daten (hier Archivpakete) zugegriffen werden kann, diese dauerhaft lesbar sind und dauerhaft erhalten bleiben.

Die Erhaltungsstrategie ist abhängig von den signifikanten Eigenschaften, also den Eigenschaften eines Objektes, die unbedingt erhalten werden müssen. Als Erhaltungsstrategie wird am FDZ am IQB die Migration gewählt.

[15] s. dazu auch nestor Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung. Version 2.0; hg. v. H. Neuroth, A. Oßwald, R. Scheffel, S. Strathmann, M. Jehn; verfügbar unter: http://www.langzeitarchivierung.de/; hier: Kapitel 8 „Digitale Erhaltungsstrategien“ (Version 2.0): urn:nbn:de:0008-20090811378 sowie https://de.wikipedia.org/wiki/Migration_(Informationstechnik)#Medienmigration

#### 4.2.6.1.1 Medienmigration

Die Archivierungspakete werden vom CMS im Rahmen der regelmäßigen Datensicherung regelmäßig gesichert, so dass das CMS die Datensicherung (Bitstream Preservation) gewährleistet. Das bedeutet, dass das FDZ am IQB die Datenbestände nicht selbst migrieren muss (d. h., eine reine Datenträgermigration von Seiten des FDZ am IQB ist nicht nötig). Das CMS führt als Arten der Medienmigration Refreshment und Replication durch. Sie bezeichnen das Auswechseln einzelner Datenträger (refreshing) oder eine Änderung eingesetzter Speicherverfahren (replication). Der Umkopierprozess erfolgt in beiden Fällen mit der Absicht, das Trägermedium zu ersetzen, unabhängig davon, welche Inhalte auf ihm abgelegt sind. Es finden keine Änderungen an den Daten oder der Speicherinfrastruktur statt.

#### 4.2.6.1.2 Formatmigration

Das FDZ am IQB konvertiert die den Datennutzenden zur Verfügung gestellten Daten (=.sav-Dateien) in langzeitarchivierbare Formate (.csv, .pdf/A, .txt). Dies wird als Transformation bezeichnet – ein Migrationsprozess, der auch die Inhalte des AIP verändert. Die vom FDZ am IQB verwendeten Langzeitarchivierungsformate sind also nicht von der Version der Ursprungserstellungssoftware (derzeit SPSS) abhängig.

Bei (größeren) Versionssprüngen der Software werden die entsprechenden digitalen Daten auf die neue, aktuelle Version überführt.

Für den Fall, dass das bis dahin verfügbare Format der DIPs (bisher hauptsächlich SPSS) nicht mehr verwendet wird, werden die Daten in ein alternatives, anderes Format konvertiert. Dieses Vorgehen ist aufwendig und wird auch nur dann vorgenommen, wenn tatsächlich durch das Verfolgen und Beobachten der technischen/technologischen Entwicklungen[16] davon auszugehen ist, dass die digitalen Dokumente in Zukunft aufgrund des Formats nicht mehr nutzbar sind.

Einmal im Jahr wird geprüft, ob eine Migration notwendig wird.

Im Falle einer Migration muss sich eine Qualitätssicherung anschließen, d. h., es wird stichprobenartig geprüft, ob die digitalen Objekte noch lesbar und interpretierbar sind.

### *4.2.6.2 (Über-)Prüfung / Kontrolle*

Ein wahrscheinliches Szenario für Datenverlust ist menschliches Versagen. Es braucht demnach einen Mechanismus, um zu überprüfen, ob alle Daten noch unverändert vorliegen. Daher überprüft die wissenschaftliche Dokumentarin (data librarian) die AIPs und erstellt pro AIP über die darin enthaltenen Objekte/Dateien/files eine Checksummendatei (.sha256). Bevor ein AIP in die Langzeitarchivierung des CMS übergeht, verifiziert die IQB-IT die Checksummendatei, um Datenintegrität zu gewährleisten. Die Langzeitarchivierung des CMS stellt sicher, dass die dort liegenden Daten lesbar bleiben (=bitstream).

## 4.3 Datenaufbereitung

Zuständig und verantwortlich für die Datenaufbereitung sind die wissenschaftlichen Mitarbeitenden des FDZ am IQB. Das FDZ am IQB prüft, ob die Datendokumentation für Forschende, die nicht an der Datenerhebung beteiligt waren, zur Analyse der Datensätze

[16] als Teil des Preservation Planning, d. h., richtige Entscheidungen mit Blick auf die Erhaltung der digitalen Objekte zu treffen

ausreicht. Darüber hinaus prüft das FDZ am IQB stichprobenartig, ob die gemeldeten Ergebnisse mit den eingereichten Daten replizierbar sind und ob die Daten mit technischen Berichten und Skalen-Dokumentationen übereinstimmen. Weitere Prüfungen werden hinsichtlich der Plausibilität, Konsistenz und des Datenschutzes durchgeführt. Metadaten werden angereichert. Schließlich wendet das FDZ am IQB in enger Zusammenarbeit mit den Datengebenden Maßnahmen zur Datenbereinigung an (z. B. Korrektur von Tippfehlern, Hinzufügen von Variablenbeschriftungen, Wertebeschriftungen und fehlenden Kategorien), es werden Inkonsistenzen und Fehler gemeldet, notwendige Korrekturen vorgenommen und fehlende Informationen ergänzt, alles um die Datenqualität zu erhöhen sowie gleichzeitig ein Maximum an Informationen der Originaldaten zu erhalten. Durch dieses Vorgehen wird sichergestellt, dass die Daten vollständig, brauchbar und interpretierbar sind.
Alle Arbeitsschritte werden in einer Datenbank dokumentiert. Jede einzelne digitale Ressource, die im FDZ im IQB veröffentlicht wird, unterliegt einer Qualitätskontrolle.

Im Rahmen des Qualitätsmanagements sendet das FDZ im IQB nach Abschluss der Datenaufbereitung auch eine E-Mail an die Datengebenden, um sie über die Bereitstellung der Daten zu informieren und sie aufzufordern, die auf den Webseiten des FDZ am IQB und des VerbundFDB veröffentlichten Metadaten zu überprüfen.

### 4.3.1 Dokumentation

Aus Gründen der Nachvollziehbarkeit werden die Schritte zur dauerhaften Verfügbarkeit digitaler Daten dokumentiert. Metadaten tragen dabei zur langfristigen Erhaltung von digitalen Objekten bei. Von großer Bedeutung sind hierbei vor allem Angaben zum vorliegenden Format der Daten, die das FDZ am IQB a) von den Datengebenden erhalten hat, b) an die Datennutzenden herausgibt und c) archiviert. Denn ein digitales Objekt muss nicht nur lesbar, sondern auch korrekt interpretierbar bleiben. Daten und Dokumentation werden in klar definierten, standardisierten Dateiformaten archiviert. Zusätzlich werden Syntax- und Setup-Dateien aufbewahrt, die die Änderungen zwischen verschiedenen Datenversionen dokumentieren. Die vorhandene interne Dokumentation steht allen Mitarbeitenden zur Verfügung. Alle Transformationen der Daten werden dokumentiert.

Alle Arbeitsschritte werden in eigenentwickelten Datenbankanwendungen dokumentiert, die nur autorisierten Mitarbeitenden des FDZ am IQB zur Verfügung stehen.

Pro Studie wird in der Datenbank die ID des Archivpakets dokumentiert und wann das Archivpaket in die Langzeitarchivierung des CMS übergegangen ist.

Werden Variablen oder Datensätze zum Datenpaket hinzugefügt, die eine neue externe Version zur Folge haben, wird eine neue Versionsnummer erstellt und diese Datenversion ebenfalls langzeitarchiviert. Jede Datenversion hat ihre eigene DOI. Jedes Datenpaket mit einer DOI erhält ein eigenes AIP, d. h., neue Datenversionen haben ein eigenes AIP.

## 4.4 Zugriff (Access)

Das FDZ am IQB vergibt je bereitgestelltem Datenpaket einen Digital Object Identifier (=DOI), dadurch ist die dauerhafte Verfügbarkeit der bereitgestellten Forschungsdaten für die Scientific

Community gegeben. Ebenso ermöglicht die Vergabe von Persistent Identifiers die Zitierfähigkeit von Forschungsdaten und somit wird auch die Sichtbarkeit und Transparenz der Datenproduzierenden erhöht.

Jede Datenversion hat ihre eigene DOI, der es Forschenden ermöglicht, veröffentlichte Ergebnisse auf der Grundlage älterer Datenversionen zu replizieren.

Die Datensätze (=DIP) werden auf Antrag Forschenden, die eine akademische Zugehörigkeit haben, in verschiedenen Formaten (SUF, IZA), die von der Vertraulichkeit der Datensätze sowie vom Standort der Forschenden abhängen, für wissenschaftliche, nicht-kommerzielle Zwecke zur Verfügung gestellt.

Anträge auf Datenzugang werden online gestellt und enthalten eine kurze Projektskizze mit einer Zusammenfassung der theoretischen Grundlagen, Hypothesen und geplanten Analysen. Antragsrichtlinien, ein Musterprojektantrag und das Antragsformular sind online verfügbar. Das FDZ am IQB prüft die Anträge auf die Einhaltung der formalen Kriterien für die Genehmigung. Diese Kriterien umfassen im Wesentlichen vier Bereiche:

- Sollen die Daten für rein nicht-kommerzielle und wissenschaftliche Zwecke verwendet werden?
- Wird der Datenschutz respektiert?
- Entsprechen die vorgesehenen Analysen den vertraglichen Vereinbarungen mit dem Datengebenden?
- Ist sichergestellt, dass die geplanten Analysen die laufende Qualifizierungs- und Publikationsarbeit an den Daten nicht gefährden? (=Sind Sperrvermerke betroffen?)

Wenn die formalen Kriterien erfüllt sind, wird ein Datennutzungsvertrag geschlossen, erst dann erfolgt der Zugang zu den Daten.

Datennutzende erklären sich mit Folgendem einverstanden:

- Die Nutzung ist nur für wissenschaftliche Zwecke in Forschung und Lehre erlaubt.
- Eine Übertragung von Nutzungsrechten an den Daten auf Dritte durch den Datennutzer ist unzulässig.
- Es dürfen keine Versuche unternommen werden, Personen aus dem Datenbestand zu re-identifizieren. Bei einer versehentlichen Re-Identifikation ist das FDZ am IQB zu benachrichtigen.
- Es dürfen keine Daten von einzelnen Personen oder Gruppen von weniger als fünf Personen berichtet werden.
- Die Daten müssen nach Abschluss des Projekts (bzw. spätestens nach Ablauf der Vertragslaufzeit des Datennutzungsvertrags) gelöscht werden.

Bei einem Verstoß gegen die Vertragsbedingungen erlischt das Nutzungsrecht sofort und der Datennutzende muss eine Vertragsstrafe in Höhe von 10.000 € zahlen.

Die vom FDZ im IQB erhaltenen Daten müssen nach Abschluss der Analysen, für die sie zur Verfügung gestellt wurden, vernichtet werden.

Zu den Bedingungen der Datennutzung gehört, dass Datennutzende die vom FDZ am IQB erhaltenen Daten nur zugangsgeschützt sowie auf passwortgeschützten Speichermedien speichern dürfen. Die Daten dürfen nur in Länder gebracht werden, die über ein angemessenes Datenschutzniveau verfügen.

Der Arbeitsablauf zur Aufbereitung und Versendung der Datensätze in verschiedenen Formaten an Datennutzende ist in einer Datenbank dokumentiert, die nur autorisierten Mitarbeitenden des FDZ am IQB zur Verfügung stehen.