

Konzept zur Langzeitverfügbarkeit von Forschungsdaten

überarbeitete Version des Vorläufer-Dokuments: 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.

Stand: Juli 2025

Bibliographische Informationen / Bitte wie folgt zitieren:

Forschungsdatenzentrum am Institut zur Qualitätsentwicklung im Bildungswesen (FDZ am IQB) (2025). *Konzept zur Langzeitverfügbarkeit von Forschungsdaten*. Berlin: IQB - Institut zur Qualitätsentwicklung im Bildungswesen. http://doi.org/10.5159/IQB_LZA-Konzept_v1

Das Werk steht unter der Creative-Commons-Lizenz Namensnennung - Weitergabe unter gleichen Bedingungen 4.0 International (<https://creativecommons.org/licenses/by-sa/4.0/legalcode.de>). Ausgenommen von der oben genannten Lizenz sind Teile, Abbildungen und sonstiges Drittmaterial, wenn anders gekennzeichnet.



Inhalt

1	Begriffe und Definition zur Langzeitarchivierung	4
2	Schaubild Langzeitarchivierung	6
3	Technische Infrastruktur	7
4	Langzeitverfügbarkeit am FDZ am IQB	8
4.1	Datenübernahme: Ingest.....	9
4.2	Datenaufbewahrung: Archival Storage.....	10
4.2.1	Welche digitalen Objekte werden langzeitverfügbar gemacht?.....	10
4.2.2	Wann werden die Archivpakete erstellt?	10
4.2.3	Wer ist verantwortlich für die Erstellung der Archivpakete?	10
4.2.4	Wie erfolgt die Erstellung der Archivpakete?.....	10
4.2.5	Wo werden die langzeitverfügbaren digitalen Objekte aufbewahrt? Wer hat Zugriff?..	13
4.2.6	Wie wird die Lesbarkeit der Daten im AIP gewährleistet?.....	13
4.3	Datenaufbereitung	14
4.3.1	Dokumentation.....	15
4.4	Zugriff (Access).....	15

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 nestor¹ 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.

Backups 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.

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: urn:nbn:de:0008-2010071949, Kap. 1:3. Letzter Zugriff: 5.6.2025

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, dass der langfristige Forschungswert erhalten bleibt und das Risiko der digitalen Veralterung³ gemindert wird.

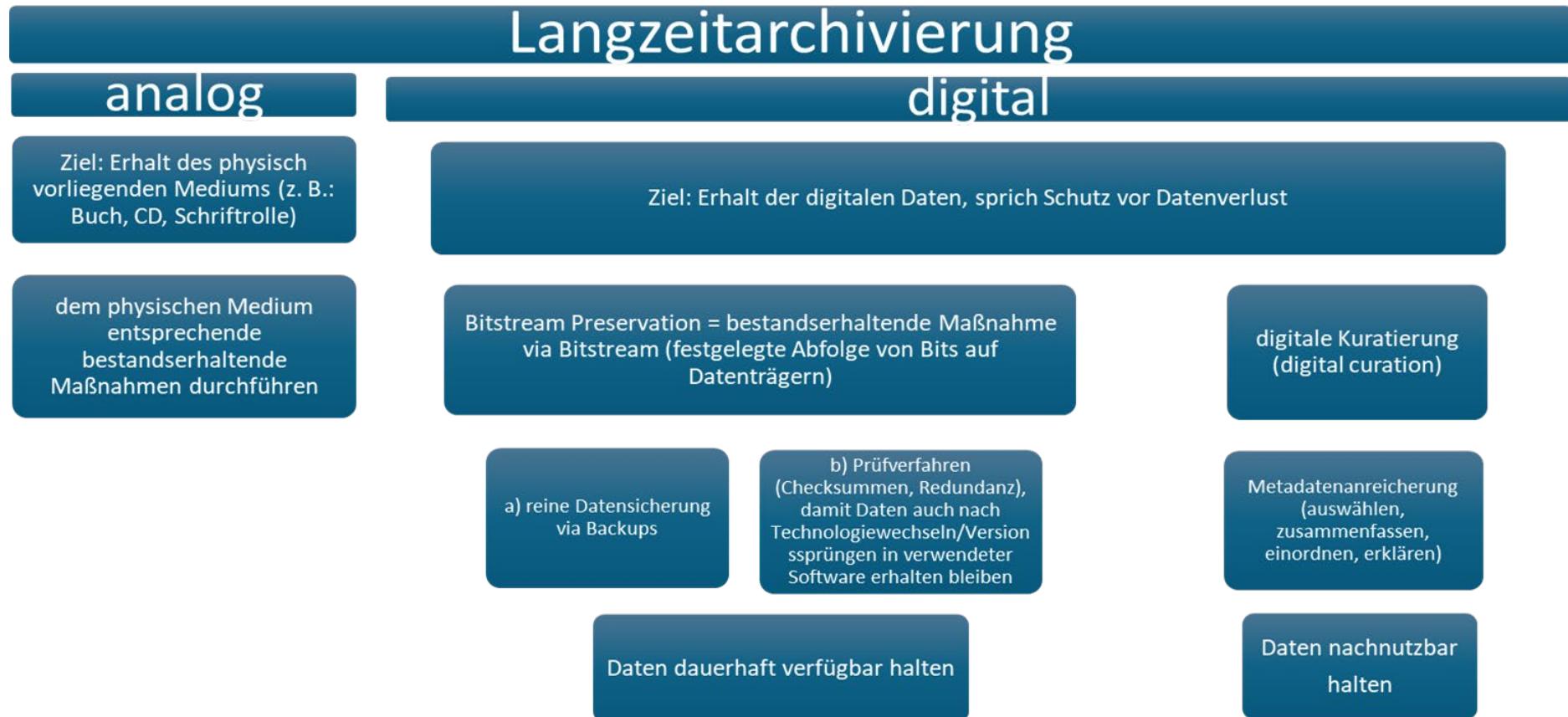
Die Daten müssen nicht nur aufbewahrt und erhalten (kuratiert) werden, sondern müssen, um ihren Mehrwert zu entfalten, von der Scientific Community genutzt werden können. Dies kann durch vertrauenswürdige Datenzentren ermöglicht werden.⁴ Das Forschungsdatenzentrum am IQB ist ein solches vertrauenswürdige 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: <https://www.dcc.ac.uk/about/digital-curation>. Letzter Zugriff: 5.6.2025

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 Zugriffskontrollverfahren⁶ 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.

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: <https://www.cms.hu-berlin.de/de/dl#/svc14>. Letzter Zugriff: 5.6.2025

⁶ siehe auch: <https://www.cms.hu-berlin.de/de/publikationen/ordnungen/ZutrittCMS>. Letzter Zugriff: 5.6.2025

⁷ siehe auch: <https://www.cms.hu-berlin.de/de/dl#/svc15> und <https://www.cms.hu-berlin.de/de/dl/systemservice/fileservice/tsm>. Letzter Zugriff: 5.6.2025; 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

4 Langzeitverfügbarkeit am FDZ am IQB

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).

In Anlehnung an das Funktionsmodell des Offenen Archivinformationssystems (Open Archival Information System OAIS)⁸ ist das FDZ am IQB das Archiv/Repository bzw. die Verbindung (Konnektor) zwischen Datengebenden und Datennutzenden.

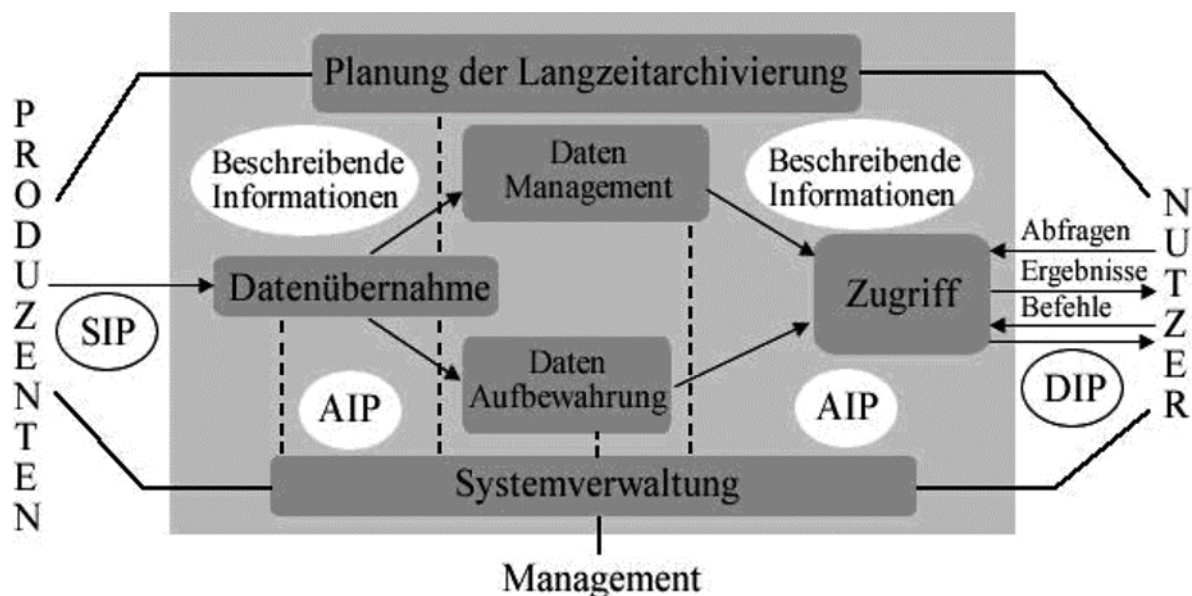


Abb.: OAIS-Referenzmodell⁹

Alle digitalen Objekte (d. h.: Daten in Form von Datensätzen sowie weitere Kontextinformationen in Form von begleitendem Dokumentationsmaterial und Metadaten), die das FDZ von den Datengebern erhält, stellen das nach der OAIS-Terminologie sog. Informationspaket *Submission Information Package* (=SIP) dar.

Das Informationspaket *Archival Information Package* (=AIP) umfasst die – vom FDZ mit Metadaten ergänzt – digitalen Objekte, die langfristig am FDZ am IQB aufbewahrt, also langzeitarchiviert werden.

⁸ s. dazu: 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>. Letzter Zugriff: 5.6.2025

⁹ s. dazu: urn:nbn:de:0008-2010061762, S. 11

Die Forschungsdaten und Dokumentationsmaterialien, die schließlich für die Datennutzenden bereitgestellt werden, werden als Informationspaket *Dissemination Information Package* (=DIP) bezeichnet.

Das AIP enthält 1) alle von den Datengebenden erhaltenen digitalen Objekte (=SIP), 2) die für die Datennutzenden vorbereiteten bereitzustellenden digitalen Objekte (=DIP) sowie 3) alle Dokumentations- und Aufbereitungsschritte (inkl. Syntax/en), die am FDZ am IQB durchgeführt wurden, um ein oder mehrere Datenprodukte für die Datennutzenden bereitstellen zu können; aber auch Verträge, Zugriffsregelungen, Kommunikation mit den Datengebenden sowie weitere begleitende Dokumentationsmaterialien. D. h., vom SIP kommend über die AIP-Erstellung generiert das FDZ am IQB das DIP. Alle digitalen Objekte liegen in langzeitverfügbaren Formaten vor, ebenso wird pro AIP eine Checksummendatei gebildet.

4.1 Datenübernahme: Ingest

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

Bei der Datenübergabe an das FDZ am IQB wird ein Vertrag zur Datenaufnahme 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 ans FDZ am IQB und mit einer dortigen Archivierung der Forschungsdaten 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 Forschungsdaten und Begleitmaterialien wie/auf welche Weise an Dritte weitergegeben werden (a. standardmäßig öffentlich auf der Website des FDZ am IQB, b. nur auf Anfrage zusammen mit den Forschungsdaten oder c. nie).

Um die Datenqualität zu gewährleisten, hat das FDZ am IQB Mindestanforderungen definiert, die die eingereichten Forschungsdaten erfüllen müssen. Bei allen eingehenden digitalen Objekten prüft das FDZ am 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 Forschungsdaten 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 Forschungsdaten für die Re- und Sekundäranalyse, bewertet.

Alle Arbeitsschritte werden in einer inhouse-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 es 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. Anschließend 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 digitalen Objekte gewahrt bleibt.

4.2.1 Welche digitalen Objekte werden langzeitverfügbar gemacht?

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

Konkret heißt das, dass das FDZ am IQB alle digitalen Objekte der Studien, die bereitgestellt werden, langzeitarchiviert. Zusätzlich werden die Originaldatensätze inkl. des Begleitmaterials in langzeitarchivierten Formaten archiviert sowie alle Dokumentations- und Aufbereitungsschritte, die am FDZ am IQB durchgeführt wurden (bspw. Aufbereitungssyntaxen und Dokumente zur Bewertung der Forschungsdaten), um das DIP pro Studie für Datennutzende zu generieren.

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 Version angelegt und enthält das SIP sowie die darin enthaltenden digitalen Objekte in langzeitverfügbarem Format. Nach der Datenaufbereitung, dessen Ergebnis das DIP für die an Datennutzende herauszugebende Studie ist, wird das AIP eben damit befüllt – auch hier alles ebenfalls in langzeitverfügbaren Formaten.

4.2.3 Wer ist verantwortlich für die Erstellung der Archivpakete?

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

4.2.4 Wie erfolgt die Erstellung der Archivpakete?

Ziel ist es, dass die vom FDZ am IQB zur Verfügung gestellten digitalen Objekte 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 Objekte in unverschlüsselten, nicht komprimierten, nicht proprietären Formaten, sondern mit offenen, dokumentierten Standards gespeichert.

Redundanz bei den digitalen Objekten ist hinsichtlich der Datensicherheit nützlich und sinnvoll, d. h., das FDZ am IQB speichert Textdateien sowohl im PDF/A¹⁰- als auch im .txt- Format sowie Tabellendaten auch 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 wird¹¹.

4.2.4.2 Vorgehen

Am FDZ am IQB wird zur Erstellung von langzeitverfügbaren Daten je Studie ein Archivpaket erstellt. Ein Archivpaket ist das sog. Archival Information Package (AIP).

Im Konkreten bedeutet das, dass nach Eingang der digitalen Objekte durch Datengebende und dem Beschluss, dass die Forschungsdaten der Studie bereitgestellt werden sollen, diese sofort in langzeitarchivierbare Formate gebracht werden, sprich Datensätze 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 entsprechenden 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 Versionen.

Nach der Datenaufbereitung werden im AIP die bereitzustellenden digitalen Objekte (=DIP) abgespeichert; ebenfalls alle in langzeitverfügbaren Formaten.

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-Angabe, Zitationsvorschlag, Datengebende, Zugriffsbestimmungen, Rechte) sowie deskriptiven (z. B.: Studienzeitraum, Erhebungszeitraum, Auswahlverfahren) Metadaten¹² (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, d. h., diese sind zwar in den Datensätzen enthalten, werden aber nicht separat als Metadatum erfasst. Pro Studie sowie bei

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>. Letzter Zugriff: 5.6.2025

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://forschungsdaten.info/themen/beschreiben-und-dokumentieren/metadaten-und-metadatenstandards>. Letzter Zugriff: 5.6.2025)

einer neuen Version werden die Metadaten als entsprechende .xml-Datei¹³ mit in das AIP mit aufgenommen.

Nach der Datenaufbereitung wird das AIP mit weiteren Objekten befüllt: neben den bereitzustellenden digitalen Objekten – also inkl. der Syntax/en¹⁴ 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.

4.2.4.3 Archivpaket

Ein AIP umfasst immer folgende digitalen Objekte:

Begleitmaterialien

- ◆ vom Datengebenden mitgelieferte Begleitmaterialien (nur: Einwilligungserklärungen/Genehmigungsschreiben)
 - als PDF/A
 - als .txt
- ◆ bereitzustellende Begleitmaterialien
 - als PDF/A
 - als .txt

Datensätze

- ◆ bereitzustellende/r Datensatz/-sätze
 - als .csv
 - als SPSS
 - ggf. als R
 - ggf. als Stata
- ◆ Originaldatensatz/-sätze vom Datengebenden
 - als .csv
 - als SPSS

Metadaten

- ◆ als .xml
- ◆ als .csv

¹³ Hierbei handelt es sich um die xml-Datei, die beim Registrierungsservice für DOIs 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 (=die, die die Grundlage bildet, welche Daten herauszugeben sind).

Syntax/en der bereitgestellten Datensätze

Zentrale Dokumente

- ♦ jeweils als ein PDF/A je Dokument

Checksummendatei

4.2.5 Wo werden die langzeitverfügbaren digitalen Objekte aufbewahrt? Wer hat Zugriff?

Die AIPs der langzeitverfügbaren digitalen Objekte 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) sowie deren Vertretung Zugriff hat.

Von dort werden die AIPs 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 Migration¹⁵

Ein AIP enthält die digitalen Objekte im originären sowie in langzeitarchivierten Formaten. Die den Datennutzenden zur Verfügung gestellten Daten werden als .sav-Datei(en) geliefert. Das FDZ am IQB ist sich bewusst, dass SPSS eine proprietäre Software ist, aber es ist diejenige, die in der designated community des FDZ am IQB am häufigsten verwendet wird. Dieses Problem ist aber auch der Grund, warum das FDZ am IQB die Datenaufbereitung in R vornimmt; und sich ebenso verstärkt bemüht, Prozesse anzupassen/zu verändern.

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

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

4.2.6.1.1 Medienmigration

Die AIPs 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 digitalen Objekte 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,

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](https://de.wikipedia.org/wiki/Migration_(Informationstechnik)#Medienmigration)

unabhängig davon, welche Inhalte auf ihm abgelegt sind. Es finden keine Änderungen an den digitalen Objekten an sich oder der Speicherinfrastruktur statt.

4.2.6.1.2 Formatmigration

Das FDZ am IQB konvertiert die den Datennutzenden zur Verfügung zu stellenden Forschungsdaten (= .sav-Dateien) in langzeitarchivierbare Formate (.csv). 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 Forschungsdaten 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 Entwicklungen¹⁶ davon auszugehen ist, dass die vorliegenden Forschungsdaten 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 digitalen Objekte noch unverändert vorliegen. Daher überprüft die wissenschaftliche Dokumentarin (data librarian) die AIPs und erstellt pro AIP über die darin enthaltenen digitalen Objekte eine Checksummendatei (.sha). 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 Forschungsdaten ausreicht. Darüber hinaus prüft das FDZ am IQB stichprobenartig, ob die gemeldeten Ergebnisse mit den eingereichten Forschungsdaten replizierbar sind und ob diese mit technischen Berichten und Skalendokumentationen 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

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

Variablenbeschriftungen, Wertebeschriftungen, konsistente Codierung von fehlenden Werten), 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 originären digitalen Objekte zu erhalten. Durch dieses Vorgehen wird sichergestellt, dass die digitalen Objekte vollständig, brauchbar und interpretierbar sind. Alle Arbeitsschritte werden in einer Datenbank dokumentiert. Jedes einzelne digitale Objekt, welches im FDZ am IQB veröffentlicht wird, unterliegt einer Qualitätskontrolle.

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

4.3.1 Dokumentation

Aus Gründen der Nachvollziehbarkeit werden die Schritte zur dauerhaften Verfügbarkeit digitaler Objekte dokumentiert. Metadaten tragen dabei zur langfristigen Erhaltung von digitalen Objekten bei. Von großer Bedeutung sind hierbei vor allem Angaben zum vorliegenden Format der digitalen Objekte, 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. Datensätze und Dokumentation werden in klar definierten, standardisierten Dateiformaten archiviert. Zusätzlich werden Syntax- und Setup-Dateien aufbewahrt, die die Änderungen zwischen verschiedenen Versionen dokumentieren. Die vorhandene interne Dokumentation steht allen FDZ am IQB-Mitarbeitenden zur Verfügung. Alle Transformationen der digitalen Objekte werden dokumentiert.

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

Pro Studie wird in der Datenbank die ID des AIP dokumentiert, ebenso wann das AIP in die Langzeitarchivierung des CMS übergegangen ist.

Werden Variablen oder Datensätze einer Studie hinzugefügt, was eine neue externe Version zur Folge hat, wird eine neue Versionsnummer erstellt und diese Version ebenfalls langzeitarchiviert. Jede Version hat eine eigene DOI. Eine neue Version erhält ein eigenes AIP.

4.4 Zugriff (Access)

Das Ergebnis eines Kuratierungsprozesses ist ein Datenprodukt, repräsentiert durch einen PID (hier: DOI), der die persistente Auffindbarkeit und dadurch die dauerhafte Verfügbarkeit der Forschungsdaten für die Scientific Community sichert – eine Standardanforderung der LZA.

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

Im Detail: Jedem Datenprodukt entspricht eine DOI. Ein Datenprodukt kann mehrere Datensätze enthalten – dabei handelt es sich oftmals um mehrere Einzeldatensätze, die inhaltlich aufeinander

bezogen sind (z. B. einer Studie entstammen) und die gemeinsam in einem oder mehreren ZIP-Files (z. B. für unterschiedliche Datenformate, also .dta, .sav) zur Verfügung gestellt werden –, hat aber immer die gleiche Anonymisierungsstrategie zugrunde liegend. Das bedeutet, dass es pro Studie sowohl ein Datenprodukt mit sensiblen Datensätzen (SUF Remote) als auch ein Datenprodukt ohne sensible Datensätze (SUF Off-site, CUF) geben kann. Ein Datenpaket stellt nun einen Zusammenschluss mehrerer Datenprodukte der gleichen Studie aufgrund der unterschiedlichen Granularität in der Sensibilität der Datensätze dar. Ein AIP kann aus einem oder mehreren Datenprodukten einer Studie bestehen, je nachdem, wie viele Datenprodukte die Studie aufgrund der Granularität in der Sensibilität der Datensätze hat.

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

Die Forschungsdaten (=DIP) werden auf Antrag Forschenden, die eine akademische Zugehörigkeit haben, über verschiedene Zugangswege (off-site, remote), 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 Forschungsdaten 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 Forschungsdaten nicht gefährden? (=Sind Sperrvermerke betroffen?)

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

Datennutzende erklären sich mit Folgendem einverstanden:

- ♦ Die Nutzung ist nur für wissenschaftliche Zwecke in Forschung und Lehre erlaubt.
- ♦ Eine Übertragung von Nutzungsrechten an den Forschungsdaten auf Dritte durch Datennutzende 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 Forschungsdaten 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 Datennutzende müssen eine Vertragsstrafe in Höhe von 10.000 € zahlen.

Die vom FDZ am IQB erhaltenen Forschungsdaten 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 Forschungsdaten nur zugangsgeschützt sowie auf passwortgeschützten Speichermedien speichern dürfen. Die Forschungsdaten dürfen nur in Länder gebracht werden, die über ein angemessenes Datenschutzniveau verfügen.

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

Weiterführende Kontextinformationen sowie begleitendes Dokumentationsmaterial zur Studie finden Datennutzende auf den entsprechenden Landing Pages, welche über die DOI öffentlich zugänglich sind.