8 Working Paper 2023

KonsortSWD
Konsortium für die Sozial-, Verhaltens-, Bildungs- und Wirtschaftswissenschaften

Workshopdokumentation: Langzeitarchivierung (LZA) in Forschungsdatenzentren (FDZ)

Ute Hoffstätter und Kerstin Beck

Februar 2024

www.konsortswd.de

Workshopdokumentation: Langzeitarchivierung (LZA) in Forschungsdatenzentren (FDZ)

Ute Hoffstätter,[1] Kerstin Beck[2]

Februar 2024

https://doi.org/10.5281/zenodo.10261654

[1] Deutsches Zentrum für Hochschul- und Wissenschaftsforschung (DZHW)
[2] GESIS – Leibniz-Institut für Sozialwissenschaften

Abstract

Es folgt ein Bericht zum Workshop „Langzeitarchivierung (LZA) in Forschungsdatenzentren (FDZ)" vom 14. November 2023, der sich an interessierte Forschungsdatenzentren (FDZ), die durch den RatSWD akkreditiert und/oder Teil von KonsortSWD sind, richtete. Ziel war es, den Forschungsdatenzentren das Thema Langzeitarchivierung näherzubringen sowie Einblicke in die Praxis zu geben. Dies fand als Unterstützungsangebot zur Erlangung des CoreTrustSeal (CTS) statt. Das CoreTrustSeal ist eine disziplinunabhängige Zertifizierung für digitale Langzeitarchive. Die Durchführung des Workshops erfolgte im Rahmen des KonsortSWD-Measure TA.2-M.1 „Forschungsdatenzentren unterstützen".

Keywords: Langzeitarchivierung (LZA), Forschungsdatenzentren (FDZ), CoreTrustSeal (CTS), Workshop

Inhaltsverzeichnis

1. Überblick Workshop

Workshop-Titel:	Langzeitarchivierung (LZA) in Forschungsdatenzentren (FDZ)
Organisation:	KonsortSWD[1], TA.2-M.1: Kerstin Beck; Ute Hoffstätter; Daniel Buck[2]; Dr. Pascal Siegers[3]
Teilnehmende:	- Interessierte Forschungsdatenzentren (FDZ) aus KonsortSWD, akkreditierte Forschungsdatenzentren (FDZ) - 25 Vertreter*innen von 21 Institutionen haben teilgenommen

2. Beschreibung

Das KonsortSWD-Measure TA.2-M.1 hat unter anderem das Ziel, Forschungsdatenzentren (FDZ) bei der CoreTrustSeal-Zertifizierung zu unterstützen. Das CoreTrustSeal ist eine disziplinunabhängige Zertifizierung für die Vertrauenswürdigkeit digitaler Langzeitarchive. Aufbauend auf Bedarfen der FDZ (Hoffstätter & Beck, 2023) wurde daher eine Veranstaltung zum Thema Langzeitarchivierung (LZA) virtuell durchgeführt. Zu Beginn der Veranstaltung wurde eine Definition von Langzeitarchivierung sowie in Bezug zu den Forschungsdatenzentren gesetzt.

„Immer mehr für die wissenschaftliche Sekundärnutzung potenziell interessante Daten sind in digitaler und damit leicht speicherbarer und weitergebbarer Form verfügbar. […] Damit diese wertvollen Datensätze nicht verloren gehen, sind nachhaltige Konzepte der Langzeitarchivierung erforderlich."

(Altenhöner & Oellers, 2012, S. 11)

Langzeitarchivierung bedeutet dabei mehr als die Sicherung von Dateien über einen bestimmten Zeitraum, sondern die Erhaltung der physischen und inhaltlichen Interpretierbarkeit der Objekte über Zeit und Wandel hinweg. Dies geht mit der dauerhaften Evaluierung und Weiterentwicklung der Strategien für den Erhalt digitaler Objekte einher (Liegmann & Neuroth, 2010). Für Organisationen wie FDZ, die sich mit der Archivierung und Bereitstellung von Daten beschäftigen, ist es von großer Bedeutung, die langfristige Erhaltung digitaler Objekte sicherzustellen. Dies erfordert eine Beschäftigung mit dem Thema der Langzeitarchivierung. Die letztendliche Ausgestaltung hängt vom Selbstverständnis des betreffenden FDZ ab. FDZ, die sich als eine längerfristige Infrastruktur verstehen, sind u. U. eher motiviert, sich um eine eigene Langzeitarchivierung und somit eine mögliche Voraussetzung für eine CoreTrustSeal-Zertifizierung zu bemühen, als FDZ, die aufgrund befristeter

[1] https://www.konsortswd.de/
[2] Deutsches Zentrum für Hochschul- und Wissenschaftsforschung (DZHW)
[3] GESIS – Leibniz-Institut für Sozialwissenschaften

Bereitstellung von Daten eine Langzeitarchivierung eher mit Hilfe eines externen CoreTrustSeal-zertifizierten Partners erreichen würden.

Anschließend standen die Beiträge von fünf Referent*innen aus verschiedenen Bereichen sowie anschließende Diskussionsrunden im Zentrum der Veranstaltung.

Referent*in (Institution)	**Thema**	**Foliensatz**
Dr. Kai Naumann (Landesarchiv Baden-Württemberg)	Langzeitarchivierung in der Institution, OAIS	https://doi.org/10.5281/zenodo.10276163
Dr. Jonas Recker (GESIS - Leibniz-Institut für Sozialwissenschaften)	Langzeitarchivierung in der Institution, Preservation Levels	nicht veröffentlicht
Thomas Bähr (Leibniz-Informationszentrum Technik und Naturwissenschaften und Universitätsbibliothek)	Langzeitarchivierung in der Institution	https://doi.org/10.5281/zenodo.10276075
Micky Lindlar (Leibniz-Informationszentrum Technik und Naturwissenschaften und Universitätsbibliothek)	Rosetta, PREMIS	https://doi.org/10.5281/zenodo.10164340
Xiaoyao Han (Deutsches Institut für Wirtschaftsforschung)	OpenDataFormat	https://doi.org/10.5281/zenodo.10262513

Ein geplanter Beitrag des FDZ am IQB musste leider ausfallen.

Kai Naumann stellte verschiedene Aspekte der Langzeitarchivierung am Beispiel des Landesarchives Baden-Württemberg vor - von einem Überblick über die Archivalien, die rechtlichen Hintergründe, die eingesetzte Software bis hin zu den verwendeten Preservation Levels (National Digital Stewardship Alliance [NDSA], 2022) als Unterstützung für den Einstieg in sowie die Weiterentwicklung einer Langzeitarchivierung in den FDZ.

Jonas Recker gab einen Überblick über den Archivierungsworkflow bei GESIS sowie die dafür verwendeten Metadaten. Darüber hinaus stellte er die Umsetzung der Preservation Levels bei GESIS angelehnt an eine weitere Systematik von Preservation und Curation Levels (CoreTrustSeal Standards And Certification Board, 2023) dar und ging auf aktuelle technische und organisatorische Herausforderungen ein.

Thomas Bähr stellte zunächst Beweggründe für eine Langzeitarchivierung sowie den Bestand der TIB vor. Zudem wurden die Services der TIB, insbesondere „Preservation as a service", vorgestellt sowie welches Know-how auf Seiten der auftragnehmenden und der auftraggebenden Akteure erforderlich ist.

Micky Lindlar ging auf die Archivierungssoftware Rosetta sowie den Metadatenstandard PREMIS ein. Es wurden jeweils ein allgemeiner Überblick sowie anwendungsorientierte Beispiele zur Illustration gezeigt.

Xiaoyao Han stellte in ihrem Vortrag das Open Data Format (OpenDS) des DIW vor. Hierbei handelt es sich um ein mit Metadaten angereichertes offenes Dateiformat, das den Austausch zwischen unterschiedlichen Analysetools (z. B. Stata, SPSS, R, Python) ermöglicht. Eine Anwendung und Weiterentwicklung im Rahmen einer Langzeitarchivierung könnte die Interoperabilität, die Unabhängigkeit von proprietären Formaten, die Metadatenspeicherung und langfristige Verfügbarkeit von Datenmaterial unterstützen.

3. Diskussionspunkte

Es zeigte sich, dass die Teilnehmer*innen viele Detailfragen zu praktischen Aspekten der Langzeitarchivierung an die Vortragenden hatten. Beispielsweise wurde die Frage gestellt, wie genau die Ordnerstruktur eines *Archival Information Package*[4] im jeweiligen Archiv aussieht, welche Dateiformate bei verschiedenen Dateitypen genutzt und welche Dateien genau versioniert werden sollten.

Jonas Recker merkte im Verlauf der Diskussion an, dass die Forschungsdatenzentren durch ihre hohe Kuratierungsleistung bei den Kontextinformationen der Langzeitarchivierung bereits gut aufgestellt seien, insbesondere in Bezug auf den Erhalt der inhaltlichen Interpretierbarkeit der Daten. Dies sei eine große Stärke im Vergleich zu vielen anderen Institutionen und Repositorien. Dies könnte für FDZ eine Motivation darstellen, sich mit den weiteren Aspekten einer möglichen Langzeitarchivierungslösung auseinanderzusetzen.

Bei der technischen Umsetzung der Langzeitarchivierung zeigte sich in den Vorträgen ein breites Spektrum – von einer Langzeitarchivierung via Dateistruktur mit Automatisierungen bis zu spezifischen Softwareprodukten wie DIMAG oder Rosetta. Hier sei ein Austausch mit anderen Archiven wichtig, um die für das jeweilige FDZ passende Lösung zu finden. Mehrfach wurde dazu ermutigt, auf andere Archive und Communities mit Fragen zuzugehen.

Die Referent*innen betonten, dass eine Zusammenarbeit mit anderen Archiven bzw. anderen Communities (z. B. die Community der Bibliothekswissenschaften) angestrebt werden sollte. So haben die Bibliothekswissenschaften beispielsweise viel Erfahrung mit Metadatenstandards wie PREMIS, haben aber einen anderen Schwerpunkt bei den Archivobjekten und Dateiformaten. Die Entwicklung community-spezifischer Dateiformate aus den Reihen der FDZ wie das Open Data Format kann wiederum eine Bereicherung für die Nachbarcommunities sein. Somit können die FDZ und Archive von den gegenseitigen Erfahrungen profitieren, und haben die Möglichkeit, die eigenen Prozesse zu hinterfragen und weiterzuentwickeln.

[4] Ein *Archival Information Package (AIP)* ist ein Begriff aus dem *Open Archival Information System (OAIS)*-Schema, das die Zusammenstellung an Informationen beschreibt, die tatsächlich in die Langzeitsicherung geht.

Als Vision wurde die Frage aufgeworfen, ob langfristig ggf. eine Zusammenarbeit mit öffentlichen Archiven (z. B. Bundes- und Landesarchiven) denkbar wäre. Aktuell bedürfen jedoch noch einige Punkte einer weitergehenden Diskussion, so ist etwa der bisherige Aufgabenbereich der Archive gesetzlich eingegrenzt. Weiterhin müsste z. B. definiert werden, welche Daten für eine solche Archivierung in Frage kommen und wer darüber entscheidet. Im Sinne der Archivierungswürdigkeit sollten Qualitätskriterien für die zu archivierenden Objekte festgelegt werden. Bei dem Stichwort „Nachfolgeregelung" können die vergleichsweise jungen FDZ von den Erfahrungen der Bundes- und Landesarchive profitieren.

4. Wie fange ich an? Was könnten erste Schritte sein?

Zusätzlich zur Vorstellung der Archivierung in den einzelnen Institutionen wurden die Referent*innen gebeten, Tipps und Hilfestellungen unter dem Motto „Wie fange ich an? Was könnten erste Schritte sein?" für die FDZ zusammenzustellen.

Die Anregungen reichen von organisatorisch-praktischen bis hin zu umsetzungsrelevanten eher technischeren Aspekten der Langzeitarchivierung. Vielfach betont wird, dass die Langzeitarchivierung eine Aufgabe ist, die längerfristig und auch bereichsübergreifend konzipiert und koordiniert werden muss.

Zusammenfassung[5]:

Kernfragen	▪ Was ist die Motivation und warum soll archiviert werden? ▪ Archivbestände kennen – was genau soll archiviert werden? ▪ Ziele und Zielgruppen (Designated Community) definieren: Warum, für wen und für welche Nutzungsszenarien soll archiviert werden? (im Dialog mit der Community entwickeln) ▪ Einlesen in Standards (z. B. OAIS6, nestor7, Preservation Levels8, PREMIS9) ▪ das Ziel vor Augen haben und die Aktivitäten daran orientieren ▪ Diskutieren: Langzeitarchivierung mit externer Institution oder selbst umsetzen ▪ Ressourcen (Personal und materiell) einplanen – auch langfristig ▪ eine koordinierende Person benennen ▪ Für die Erarbeitung einer Langzeitarchivierung notwendige Einheiten sind nicht nur im FDZ, sondern auch in der übergeordneten Institution zu finden, diese müssen aktiv und frühzeitig in den Prozess eingebunden werden ▪ Prozesse und Workflows verschriftlichen ▪ Verantwortlichkeiten festlegen (intern und ggf. externen Parteien) ▪ mit anderen Institutionen und Communities Kontakt aufnehmen
Archivierungs-pakete	▪ Wie sehen die Daten beim Ingest technisch aus? ▪ Welche Dateiformate sollen für eine Langzeitarchivierung vorliegen? ▪ In welchem Umfang sollen Metadaten genutzt werden?
Metadaten-standard PREMIS	▪ Wie sollen die Metadaten (Informationen über die Daten) erfasst und gespeichert werden (z. B. Datenbank, XML etc.)? ▪ Welche der Ebenen der Entität „Object" sollen vorgehalten werden? Pflichtangaben (objectIdentifier, objectCategory, objectCharacteristics) und ggf. weitere archivierungsrelevante Angaben? ▪ Welche Prozesse kann ich auf diesen Metadaten aufbauen? ▪ ggf. zusätzliche Entitäten, z. B. „Agent" und „Events" zu berücksichtigen? Welche Tools / Organisationen / Personen führen Prozesse an den Daten aus und welche dieser Prozesse stellen wichtige Informationen dar, um die Daten zu dokumentieren / interpretieren? (z. B. Datenaufbereitung, Anonymisierung)

[5] Eine detailliertere Einsicht in die Tipps der Referent*innen, kann in der veröffentlichten Präsentation (Folien 13-18) vorgenommen werden.

[6] vgl. The Consultative Committee For Space Data Systems (2012) sowie Lesetipp: Lavoie (2014).

[7] siehe https://www.langzeitarchivierung.de/Webs/nestor/DE/Publikationen/publikationen_node.html

[8] siehe CoreTrustSeal Standards And Certification Board (2023) und National Digital Stewardship Alliance [NDSA] (2022).

[9] siehe https://www.langzeitarchivierung.de/Webs/nestor/SharedDocs/Downloads/DE/berichte/pREMISverstehen2021.pdf?__blob=publicationFile&v=3

Kai Naumann ergänzte, dass es wichtig sei, sich nicht von der Vielfalt der Ziele und Methoden verunsichern zu lassen. Langzeitarchivierung könne sehr unterschiedlich umfangreich und auf unterschiedliche Art und Weise umgesetzt und weiterentwickelt werden. Wichtig sei, ein einheitliches Verständnis von Begriffen und Prozessen voranzutreiben (z. B. orientiert am OAIS – Open Archival Information System bzw. Offenes Archiv-Informations-System) und im Austausch zu bleiben.

5. Zusammenfassung und Schlussfolgerungen

Diese Informationsveranstaltung zum Thema Langzeitarchivierung wurde primär im Rahmen der FDZ-Unterstützung zu CoreTrustSeal (CTS) angeboten, jedoch profitieren auch FDZ, die eine derartige Zertifizierung gerade nicht anstreben, von der Auseinandersetzung damit.

Die Beiträge machten deutlich, dass die Umsetzung einer Langzeitarchivierung sehr unterschiedlich ausgestaltet sein kann. Jedes FDZ muss selbst die Bedarfe an eine Langzeitarchivierung reflektieren, erarbeiten und weiterentwickeln.

Sofern im Rahmen der Beschäftigung mit der Langzeitarchivierung festgestellt wird, dass diese nicht im eigenen Hause geleistet werden kann, sollte eine Idee vorhanden sein, ob und wie durch eine*n externe*n Dienstleister*in (z. B. GESIS oder TIB) der angestrebte Archivierungsumfang erreicht werden kann. Im FDZ muss nichtsdestotrotz ein grobes Verständnis der notwendigen Definitionen und Prozesse einer Langzeitarchivierung vorhanden sein, um die Ausgestaltung der Langzeitarchivierung zu verstehen. Unterschiedliche Datentypen können dabei unterschiedliche Leistungsumfänge (z. B. Speicherplatz oder Metadaten) erfordern.

Für die Zertifizierung durch das CoreTrustSeal[10] ist nach aktuellem Stand eine Langzeitarchivierung in der Institution nötig. Soll die Langzeitarchivierung (teilweise) ausgelagert werden, sollte auf eine Zertifizierung der Partnerinstitution geachtet werden (z. B. CoreTrustSeal oder nestor-Siegel).

6. Ziel/Ausblick

Es ist angedacht, im Jahr 2024 Workshops zur Unterstützung bei einer CoreTrustSeal-Zertifizierung mit gemeinsamer Zeitplanung, Diskussion und Reviews durchzuführen sowie ggf. weitere inhaltliche Workshops entlang der Requirements der CTS-Zertifizierung zu organisieren, um die FDZ mit den inhaltlichen Anforderungen des Zertifikats weiter vertraut zu machen.

[10] Requirements CoreTrustSeal 2023-2025 Extended Guide: https://zenodo.org/records/7051096.

Literatur

Altenhöner, R. & Oellers, C. (Hrsg.). (2012). Langzeitarchivierung von Forschungsdaten. Standards und disziplinspezifische Lösungen. Berlin: Scivero.

The Consultative Committee for Space Data Systems. (2012). Recommendation for Space Data System Practices. Reference Model for an Open Archival Information System (OAIS). https://public.ccsds.org/Pubs/651x0m1.pdf.

CoreTrustSeal Standards and Certification Board. 2023, 1. January. Curation & Preservation Levels: CoreTrustSeal Discussion Paper. doi:10.5281/ZENODO.8083359.

Hoffstätter, U. & Beck, K. (2023). Workshopdokumentation: CoreTrustSeal: Erstveranstaltung. doi:10.5281/ZENODO.10213806.

Lavoie, B. (2014). The Open Archival Information System (OAIS) Reference Model: Introductory Guide (2nd Edition). doi:10.7207/twr14-02.

Liegmann, H. & Neuroth, H. (2010). Einführung. In H. Neuroth, A. Oßwald, R. Scheffel, S. Strathmann & K. Huth (Hrsg.), nestor Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung. Göttingen: Hülsbusch/Univ.-Verl. Göttingen.

National Digital Stewardship Alliance. (2022). 2019 Levels of Digital Preservation. doi:10.17605/OSF.IO/QGZ98.

Anhang

Weiterführende Informationen und Ressourcen

PREMIS

- PREMIS Zenodo Community: https://zenodo.org/communities/premis/
- Foliensatz: Crashkurs Digitale Langzeitarchivierung - Einführung in PREMIS: https://zenodo.org/records/4761923
- PREMIS Implementation Registry bei COPTR: https://coptr.digipres.org/index.php/PREMIS_(Preservation_Metadata_Implementation_Strategies)

Open Data Format

- Spezifikation: https://git.soep.de/opendata/specification;
- R Paket von Open Data Format: https://git.soep.de/opendata/r-package;

Weitere Fragen zur Langzeitarchivierung (LZA)

Gibt es Tools, die PREMIS-Metadaten integriert haben?

PREMIS Implementation Registry bei COPTR: https://coptr.digipres.org/index.php/PREMIS_(Preservation_Metadata_Implementation_Strategies)

Wieviel Speicherkapazität wird für den Anfang benötigt?

Näherungsweise: jährlichen Output auf 10 Jahre hochrechnen mal 2 (wg. doppelter Speicherung) bzw. mal X bei mehrfacher Sicherung plus 20 % Puffer. Je nachdem, ob datenintensivere Datentypen in Zukunft angedacht sind, mehr.

Impressum

Kontakt:
Ute Hoffstätter
Deutsches Zentrum für Hochschul- und Wissenschaftsforschung (DZHW)
Lange Laube 12
30159 Hannover
hoffstaetter@dzhw.eu

Hannover/Köln, Februar 2024

KonsortSWD Working Paper:
KonsortSWD baut als Teil der Nationalen Forschungsdateninfrastruktur Angebote zur Unterstützung von Forschung mit Daten in den Sozial-, Verhaltens-, Bildungs- und Wirtschaftswissenschaften aus. Unsere Mission ist es, die Forschungsdateninfrastruktur zur Beforschung der Gesellschaft zu stärken, zu erweitern und zu vertiefen. Sie soll nutzungsorientiert ausgestaltet sein und die Bedürfnisse der Forschungscommunities berücksichtigen. Wichtiger Grundstein ist dabei das seit über zwei Jahrzehnten durch den Rat für Sozial- und Wirtschaftsdaten (RatSWD) aufgebaute Netzwerk von Forschungsdatenzentren. In dieser Reihe erscheinen Beiträge rund um das Forschungsdatenmanagement, die im Kontext von KonsortSWD entstehen. Beiträge, die extern und doppelblind begutachtet wurden, sind entsprechend gekennzeichnet.

KonsortSWD wird im Rahmen der NFDI durch die Deutsche Forschungsgemeinschaft (DFG) gefördert – Projektnummer: 442494171.

DOI: 10.5281/zenodo.10261654

Zitationsvorschlag:
Hoffstätter, U. & Beck, K. (2024). Workshopdokumentation: Langzeitarchivierung (LZA) in Forschungsdatenzentren (FDZ). KonsortSWD Working Paper Nr. 8/2024. Konsortium für die Sozial-, Verhaltens-, Bildungs- und Wirtschaftswissenschaften (KonsortSWD). https://doi.org/10.5281/zenodo.10261654.