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

KonsortSWD Measure TA.2-M.12: FDMontheground

Projektbericht

Kathrin Behrens[1], Markus Quandt[1], Wolfgang Zenk-Möltgen[1], Kerstin Beck[1], Ruben Brück[2]

Dezember 2023

[1] GESIS – Leibniz-Institut für Sozialwissenschaften
[2] Leibniz-Institut für Psychologie (ZPID)

Abstract

Das Projekt FDMontheground hat sich zum Ziel gesetzt, die Datenkurationstätigkeiten in Forschungsdatenzentren zu unterstützen, indem ein überarbeitetes Portfolio an Dokumentations- und Trainingsmaterialien für zukünftige und bestehende FDZ-Mitarbeitende bei GESIS erarbeitet wird. Exemplarische Auszüge aus den Materialien sollen auf RDM Compas zur Verfügung gestellt werden.

Keywords:

Forschungsdatenmanagement, Datenkuration, Forschungsdatenzentrum, FDM-Training

Inhaltsverzeichnis

1 Einleitung

Das im Rahmen von KonsortSWD[1] geförderte Kurzprojekt „FDMontheground"[2] zielt auf die Entwicklung eines systematischen Aufbaus für interne Prozessdokumentation und Trainingsmaterialien in sozialwissenschaftlichen Datenkuratorien ab.

Angesichts der kurzen Projektlaufzeit von ca. 10 Monaten erfolgte eine Konzentration auf die Entwicklung einer Grundstruktur für Schulungs- und prozessbeschreibende Dokumente, die für die Abteilung(en) „Survey Data Curation" (SDC) und „Data Services for the Social Sciences" (DSS) bei GESIS und die dort angesiedelten Forschungsdatenzentren geeignet sind, aber auch ein Potenzial zur Verallgemeinerung auf andere Organisationen und insbesondere Forschungsdatenzentren besitzen. Die Projektergebnisse sind neben einem Entwurf dieser Struktur auch einige beispielhafte Pilotmaterialien.

Diese Materialien können zunächst innerhalb von GESIS, aber ggf. auch außerhalb als Vorlage für die künftige Entwicklung neuer oder weiterer Dokumentations- und Schulungsmaterialien dienen. Im Hinblick auf einen Nutzen jenseits der beteiligten FDZ sind die erstellten Dokumente für sich genommen eher als Muster denn zur direkten Übernahme nützlich, weil andere FDZ oft andere Software oder Detailprozesse anwenden. Ein mindestens teilweise höheres Verallgemeinerungspotenzial gibt es jedoch bei einem im Projekt niedergelegten Raster von FDM-Tätigkeiten für sozialwissenschaftliche Umfragedaten, das für interne Dokumentationsunterlagen prototypisch inhaltliche Zuschnitte, einen inneren Aufbau und eine pragmatische Granularität aufführt. Das Raster vermittelt dadurch auch zwischen allgemeineren FDM-Richtlinien und spezifischen technischen Lösungen und Daten-Typ bezogenen Standards.

Es ist vorgesehen, die Projektergebnisse 2024 zudem auf der RDM Compas-Plattform[3] zu veröffentlichen und ausgewählte Materialien dort zur Verfügung zu stellen. [4]

[1] https://www.konsortswd.de/

[2] KonsortSWD wird im Rahmen der NFDI durch die Deutsche Forschungsgemeinschaft (DFG) gefördert - Projektnummer: 442494171.

[3] Research Data Management Competence Base (RDM Compas): https://rdm-compas.org/

[4] Bedanken möchten wir uns bei Horst Baumann, Katharina Blinzler, Petra Brien, Markus Cziesla, Serap Firat, Sophia Kratz, Christina Laßka, Ina Lendowski, Marlies Post, Christian Prinz, Ivet Solanes Ros und Oliver Watteler für ihre Unterstützung und wertvollen Beiträge. Ein besonderer Dank geht an Boris Heizmann für seine umfassende inhaltliche Mitarbeit bei der Erstellung der Beispieldokumente (Anhang 2).

2 Idee und Zielsetzung von FDMontheground

Die Bedeutung eines umfassenden, generischen und fachspezifischen Forschungsdatenmanagements (FDM) in Forschung, Wissenschaft und Forschungsdatenzentren ist in den letzten Jahren signifikant gestiegen (RatSWD 2023). FDM ist ein wesentlicher Bestandteil der guten wissenschaftlichen Praxis und des Forschungsprozesses. Die Erstellung eines Datenmanagementplans (DMP) schon in der Planungsphase von Forschungsprojekten wird durch Forschungsförderer vorausgesetzt. (Forschungs-) Daten sollten, den FAIR-Prinzipien (Wilkinson/Dumontier/Aalbersberg et al. 2016) folgend, langfristig gesichert, verfügbar und nachnutzbar gemacht werden, rechtliche und ethische Aspekte im Umgang mit Daten bekannt sein und befolgt werden. Richtlinien zum FDM und der Datenkuration werden zunehmend detaillierter und konkreter (Corti/van den Eynden/Bishop et al. 2020; Jensen/Netscher/Weller 2019). Dies beinhaltet sich stetig verändernde Anforderungen an Forschende und Mitarbeitende in Forschungsdatenzentren. Denn nicht nur die Menge an erhobenen Daten steigt rasant, auch sind Forschende und FDZ-Mitarbeitende mit verschiedensten Datentypen und Datenformaten in ihrem Alltag konfrontiert. Dies erstreckt sich auf die Forschungsdaten selbst, sowie auf die Metadaten zur Dokumentation und Nachnutzung (Riley 2017). Infolgedessen steigt die Anzahl und Spezifität von Softwareprogrammen und FDM-Tools zur Dokumentation, Sicherung und Nachnutzbarkeit von Forschungsdaten in den verschiedenen wissenschaftlichen Disziplinen zusehends. Das Wissen um die Abfolge der Archivierungs- und Aufbereitungsschritte, die erforderliche Detailtiefe in der Aufbereitung und die Anwendung diverser Softwaretools sind essenziell und geht weit über leicht memorierbare Handlungsanleitungen hinaus. Daher ist von der Notwendigkeit zur Verschriftlichung dieses Wissens auszugehen.

Jedoch zeigt sich im Alltag, wie schnell entsprechende Dokumente veralten, z.B. durch bewusste Änderungen im Vorgehen, beim Umstieg auf ein anderes Softwareprodukt oder auch durch erzwungene Brüche in der Softwarelandschaft (etwa nach IT-Angriffen auf Wissenschaftsorganisationen [Technische Universität Berlin 2021]). Die Einbindung von neuen Mitarbeitenden mit den damit verbundenen Schulungsbedarfen ist ein weiterer Punkt, der die Relevanz einer aktuellen Dokumentation untermauert. Zudem entwickelt sich der Bereich des FDM stetig und dynamisch zugleich. Diese überlappenden, aber oft asynchronen Entwicklungen verändern die Arbeitsabläufe in der Datenkuration und erfordern damit gelegentlich tiefgreifende Revisionen von internen Dokumentations- und Trainingsmaterialien für FDZ-Mitarbeitende.

Das Projekt hatte zum Ziel, anhand von Beispielen zur Daten- und Metadatenbereitstellung die Umsetzung von generellen Richtlinien des Forschungsdatenmanagements in der Kuratierungspraxis von Forschungsdatenzentren (FDZ) auszuleuchten. Dazu sollten zunächst Einsichten über den Abgleich von allgemeinen Richtlinien im Forschungsdatenmanagement (FDM) mit lokalen Umsetzungen gewonnen werden. Die Erkenntnisse sollten in neue und verbesserte Dokumentations- und Trainingsmaterialien für zukünftige und bestehende

Mitarbeitende der FDZ münden und komprimiert im RDM Compas (TA2.M4) für andere FDZ zur Verfügung gestellt werden.

Ein konkreter Nutzen ergibt sich zunächst für die teilnehmenden FDZ („ALLBUS", „Wahlen", „Internationale Umfrageprogramme"), die nicht nur durch die Nutzung von „GESIS Search", sondern auch durch ihre Schwerpunkte bei thematisch breit angelegten quantitativen Umfragedaten, durch überwiegend geringe Zugangsrestriktionen, und durch sehr hohe Nutzung verbunden sind und daher ähnliche Prozessstrukturen aufweisen. Dennoch verbleiben u.a. durch unterschiedliche Datensatzstrukturen (Querschnitt, Längsschnitt, komparativ, Multilevel…) und –inhalte, sowie unterschiedlich reichhaltiger Metadaten auch erhebliche Unterschiede in den Prozessen, die spezifische Anpassungen der internen Dokumente mit sich bringen. Hier sind aktuelle Dokumentationen notwendiger Teil der Qualitätssicherung und reflektieren letztlich auch den Anspruch, dass alle Bearbeitungsschritte an Forschungsdaten größtmöglicher Transparenz und Regelhaftigkeit unterliegen sollen.

3 Projektablauf und Ergebnisse

3.1 Materialsammlung

Praktischer Hintergrund des aktuellen Projektes war die Beobachtung, dass für die teilnehmenden GESIS-FDZ der Bestand an Dokumentationsmaterialien durch die kontinuierliche Fortentwicklung von Arbeitsprozessen, aber auch durch größere interne Softwareumstellungen, potenziell inkonsistent und unvollständig geworden war. Ein systematisches, gemeinsames Dokumentationsraster, welches eine Diagnose solcher Probleme erlauben würde und zudem die Mehrfachnutzung bestimmter Dokumentationselemente zwischen den FDZ erleichtern würde, lag jedoch nicht vor. Daher wurde im ersten Schritt der Status Quo der aktuellen Arbeitsmaterialien erfasst. Dazu wurden Kolleg:innen aus den GESIS-Forschungsdatenzentren „Internationale Umfragen", „Nationale Umfragen" und „Wahlen" sowie aus den Teams „Data Acquisitions and Access", „Archiving" und „Metadata Standards and Interoperability" angeschrieben. Ferner wurde eine existierende ältere Dokumentations-Ressource von GESIS, das so genannte „DAS-Wiki" als internes Online-System für FDM-Prozessdokumentationen, hinsichtlich Struktur und Inhalt ausgewertet.
Den Kolleg:innen wurde der Projekthintergrund erläutert und sie waren aufgefordert, aktuelle Dokumente, die Relevanz für den jeweiligen Arbeitsbereich haben, zusammenzustellen und an das Projektteam zu senden. Zur Veranschaulichung, welche Dokumente beispielsweise relevant sind, wurde eine Tabelle mit Erläuterungen zum Forschungsdatenzyklus angehängt (s. Anhang A**Anhang**). Zu jedem Dokument sollten in einer weiteren Tabelle auch die folgenden Details erfasst werden (s. Tabelle 1).

Tabelle 1: Tabelle zur Erfassung weiterer Dokumentendetails

Dokument 1: [Dokumentenname]				
Dokumentenkategorie	Typ	Ablageort / Kontakt	Bewertung und Kommentar	Kurations-/Prozesschritt (Mehrfachauswahl möglich)
☐ Vertragsdokument ☐ Prozessdokumentation ☐ Schulungsmaterial ☐ Generelle Information ☐ Andere(s)	☐ digital ☐ Papier		☐ ist auf aktuellem Stand ☐ beschreibt den Prozess für mich hinreichend Weitere Bemerkungen:	☐ Erstellung & Empfang ☐ Auswahl & Bewertung ☐ Datenübernahme ☐ Erhaltungsmaßnahmen ☐ Speicherung & Sicherung ☐ Zugang & Nachnutzung ☐ Transformation ☐ unklar

Auf diese Weise kamen 189 Dokumente zusammen (davon 158 Dokumentformate [doc, xls, pdf, html] und 31 SPSS- oder Stata-Syntaxen/Setups). Die Materialien wurden incl. der Detailinformationen erfasst und gesichtet. Bei der Sichtung wurde darauf geachtet, dass das im Dokument beschriebene Vorgehen nicht übermäßig speziell ist und dadurch eine Anwendung in einem generischeren Kontext denkbar scheint.

3.2 Das Tätigkeitsschema

3.2.1 Organisationsprinzipien für Dokumentations- und Trainingsmaterialien

Sowohl Dokumentations- als auch Trainingsmaterialien für FDM- und Datenkuratierungsprozesse beschreiben *Tätigkeiten*, die Mitarbeitende regelmäßig in Bezug auf Forschungsdaten ausführen. Nach unserer Erfahrung können solche Beschreibungen von Tätigkeiten so gestaltet werden, dass sie gleichzeitig der Dokumentation dienen können, womit wir die Herstellung von Transparenz und Replizierbarkeit über die Methoden und Zwecke der Datenbehandlung meinen, wie auch dem Training, also der Schulung von Mitarbeitenden darin, diese Methoden sinnvoll und effektiv anzuwenden. Daher sprechen wir im Weiteren zusammenfassend von „DT-Materialien". DT-Materialien benötigen eine interne Struktur und Aufbereitung, die es den Zielgruppen erlaubt, Information einfach aufzufinden und aufzunehmen. Zudem müssen sie hinreichend vollständig und konsistent sein, ohne aber durch überzogenen Detailgehalt oder Gliederungstiefe die kognitive Aufnahme zu erschweren. Schließlich ist der Aufwand bei der Erstellung und Pflege von Dokumenten zu berücksichtigen, der es normalerweise ausschließt, ein vollständiges „Handbuch für alles" über komplexe Prozesse vorzuhalten. Eine Kernfrage ist daher, wie eine gute Balance zwischen den genannten Zielgrößen ausfallen kann.

Eine unmittelbare Folgerung ist, dass eine Modularisierung der DT-Materialien nötig ist, um diese im Fall von Änderungen selektiv aktualisieren zu können. Es bietet sich an, als ‚Module' bzw. separate Dokumente dabei solche Einheiten zu wählen, die zugleich auch den Nutzer:innen – also den Mitarbeitenden in FDM-Einrichtungen – die Navigation im gesamten Gefüge des Materials erleichtern. Daneben braucht jedes Modul auch noch eine

Binnengliederung, die einerseits dem jeweiligen Gegenstand entsprechen muss, andererseits über die Module möglichst ähnliche Strukturen haben sollte.
Für die Gliederung des DT-Materials kann es allerdings verschiedene Ansatzpunkte geben, die oft zu überlappenden Einordnungen führen dürften:

A. Die Mitarbeitenden selbst, bzw. ihre Rollen im FDM-Prozess (also nach Zuständigkeit oder nach Expertise, aber auch nach evtl. fixen organisatorischen Strukturen in der Einrichtung);
B. die eingesetzten (Software-) Werkzeuge, die oft ohnehin eigenständige Dokumentationen haben;
C. Eigenheiten der Tätigkeiten, die eine Gruppierung nahelegen (bspw. logisch in sich geschlossene ‚Blöcke ' in der Bearbeitung, wie etwa eine Prüfung von Copyright-Fragen oder die Vergabe von Schlagwörtern);
D. wenn sich solche Blöcke als wiederholbare sequenzielle Abfolge darstellen lassen: die Stellung in der Sequenz oder ‚Phase' der Bearbeitung;
E. Eigenheiten der Daten/Produkte (bspw. Regeln für die Behandlung fehlender Antworten oder die Filterprüfung bei Daten aus standardisierten Fragebögen, die sich je nach Datenerhebungsprojekt unterscheiden können).

Ein praktikabler Gliederungsansatz sollte unserer Auffassung nach nun nicht etwa versuchen, diese überlappenden Zuordnungsmöglichkeiten vollständig in eindeutige Zuordnungen aufzulösen, sondern sollte es vielmehr erleichtern, je nach konkreter Fragestellung an das Material, die Perspektive wenigstens partiell zu wechseln. Wir haben als erstrangiges Gliederungskriterium Punkt C, die Eigenheiten der Tätigkeiten, ausgewählt, weil diese auf einer höheren Abstraktionsebene grundsätzlich die größte Generalisierbarkeit über spezifische Datentypen/-produkte und institutionelle Abläufe hinweg aufweisen. Eine Spezifikation auf Institute und auf Datenprodukte erfolgt dann separat in beide Richtungen.

3.2.2 Entwicklung und Struktur des Tätigkeitsschemas

Als orientierender Rahmen zur Erarbeitung des Gliederungs- bzw. Tätigkeitsschemas diente uns das „DCC Curation Lifecycle Model" (Higgins 2008, s. Abbildung 1). Dieses definiert eine Grundstruktur aus einzelnen Phasen bzw. Stufen in der Behandlung von Forschungsdaten: Erstellung und Empfang, Auswahl und Bewertung, Datenübernahme, Erhaltungsmaßnahmen, Speicherung und Sicherung, Zugang und Nachnutzung, Transformation. Innerhalb dieser Stufen müssen nun spezifische Methoden und Maßnahmen benannt werden, um ein adäquates Forschungsdatenmanagement zu gewährleisten.

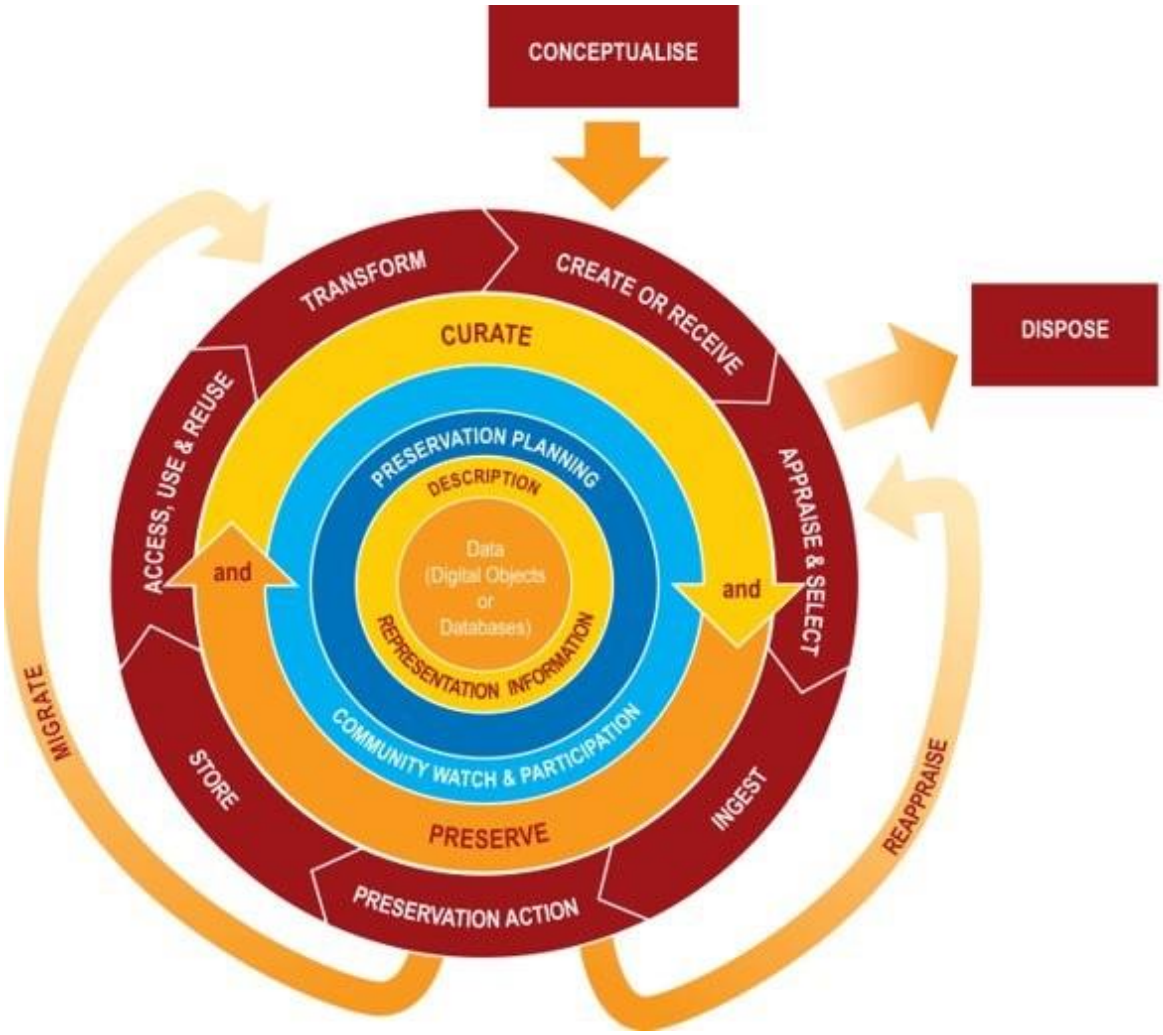

Abbildung 1: DCC Curation Lifecycle Model (Higgins 2008)

Die Bearbeitungsphasen des Modells finden sich auf *Ebene 1* unseres Schemas wieder (s. Tabelle 2 für einen Auszug sowie Anhang 1). Daran anknüpfend wurden, mittels der in den zur Verfügung gestellten Dokumenten beschriebenen Tätigkeitsschritte und den organisationsspezifischen Anforderungen, weitere Ebenen hinzugefügt. So entstand ein Schema mit vier Ebenen, von denen allerdings nur drei Tätigkeitsgruppen im engeren Sinne bezeichnen. Auf der obersten Ebene sind *Rollen* definiert. Diese Rollen sind organisationsspezifisch zu sehen, d.h. wir gehen davon aus, dass sich zwischen Instituten die Bezeichnungen der Rollen unterscheiden, vor allem aber auch die Zuordnungen von einzelnen Tätigkeiten zu den Rollen anders sein können (besonders an den ‚Grenzen' zwischen Rollen), drittens auch Hierarchie-Verhältnisse zwischen Rollen bestehen können (solche sind hier nicht abgebildet, sind aber leicht denkbar, wenn z.B. etwa nach rechtlichen Prüfungen noch Freigaben von Vorgesetzten o.ä. erforderlich wären). *Ebene 1* gibt, wie erwähnt, weitestgehend den Datenkurationszyklus des DCC-Modells wieder und wurde nur leicht angepasst. *Ebene 2* beschreibt funktional geschlossene Schritte von Einzeltätigkeiten und stellt damit eine Konkretisierung und Differenzierung der Kurationsphasen auf den Datentypus Umfragedaten dar. Auf *Ebene 3* sind die einzelnen Tätigkeitsblöcke zu finden, die typischerweise die letzte Strukturierungsebene der DT-Materialien ergeben. Die konkreten textlichen Inhalte unterhalb

dieser Tätigkeitsblöcke sind dann stark spezifisch für Datentypen oder sogar für einzelne Datensätze; mit einigen beispielhaften inhaltlichen Dokumenten werden wir uns weiter unten in diesem Bericht befassen. Die unterste Ebene 3 umfasst auch Tätigkeitsblöcke, die selbst beim Datentyp Umfragedaten nicht immer durchlaufen werden. Obwohl sich Ebene 3 auch an den typischen Abfolgen orientiert, in denen Tätigkeiten zu erledigen sind, kann es zudem vorkommen, dass die im Schema angelegte Reihenfolge verlassen werden kann, also keine strikte Sequenzialität gilt. Das wird z.B. oft der Fall sein, wenn bereits einmal aufbereitete Datensätze Updates oder Korrekturen erfahren – dann ist zunächst an den vom Update betroffenen Stellen in die Sequenz hineinzuspringen, und es sind nicht zwangsläufig alle Folgeschritte wieder zu durchlaufen. Für solche Fälle kann es sich lohnen, eine weitere Dokumentationsebene *außerhalb* unseres Tätigkeitsschemas einzuziehen, für die separate Dokumente erstellt werden (etwa ToDo-Listen für Updates wegen Datensatzergänzungen oder Fehlerkorrekturen). Dabei sollte es sich jedoch meist um *Workflows* handeln, die auf schon beschriebene Tätigkeiten innerhalb des Schemas verweisen, statt selbst neue Tätigkeitsbeschreibungen anzubieten. Dabei kann in verschiedenen Institutionen oder für verschiedene Datenkollektionen ein Workflow auch unterschiedlich aussehen und sich doch der gleichen Elemente bedienen, wie konkret auch für Längsschnittdaten von Barkow, Block, Greenfield et al. 2013, Figure 6 gezeigt.

Technisch wurde das Tätigkeitsschema in einem Excel-Worksheet erfasst und lässt sich damit leicht weiter pflegen und bei Bedarf in Open Source-Formate übertragen. Der aktuelle Umfang liegt bei ca. 75 inhaltlichen Zeilen über den gesamten Workflow mit Umfragedaten in den projektbeteiligten FDZ, wobei eine Zeile einem Aktivitätsblock auf unterster Ebene entspricht. Diese Liste ist nicht im engeren Sinne erschöpfend für alle Tätigkeiten (selbst bei den beteiligten FDZ), da sie maßgeblich aus einer Sichtung bereits vorhandener Materialien erzeugt wurde, die nicht analytisch vollständig sein muss. Sie ist jedoch gerade deshalb praktisch und von einer handhabbaren Granularität; zudem sind in der Zählung bereits Tätigkeitsblöcke enthalten, für die aus der lokalen Materialsichtung keine Dokumente berücksichtigt wurden (obwohl solche möglicherweise in irgendeiner Form vorlagen - in diesen Fällen ist in der Übersichtstabelle in Anhang 2 ein Eintrag in Ebene 2 vorhanden, aber keiner in Ebene 3).

Tabelle 2: Tätigkeitsschema (Auszug)

Rolle	Ebene 1	Ebene 2	Ebene 3
Akquise	Outreach & Beratung [...]	Beratung von Datengebenden/ Nutzenden [...]	
Datenspezia-list*innen	Datenübernahme/ Ingest	Dateneingangskontrolle	Metadaten im Datensatz (Paradaten wie IDs etc., Variablennamen und -label vollständig/verständlich)
			Archivvariablen erstellen (Syntax erarbeiten)
	Erhaltungs- & Aufwertungs-maßnahmen [...]	Datenaufbereitung	[...] Dublettenchecks auf Fallebene [...]
			[...] Anonymisierung/Pseudonymisierung [...]
		Dokumentation [...]	[...] Variablendokumentation [...]
			[...] Variablenreports, Methodenberichte
LZA	Speicherung & Sicherung	Transfer in LZA-System [...]	Digital Preservation Policy [...]
Datenbereit-stellung [...]	Zugang & Nachnutzung [...]	[...] Transfer in öffentliches Downloadsystem [...]	

3.2.3 Evaluation und vorläufige Weiterentwicklung des Tätigkeitsschemas

In einem Workshop mit acht Kolleg*innen aus den GESIS-Abteilungen SDC und DSS, die im Vorfeld auch bereits bei der Dokumentensammlung behilflich waren und überwiegend langjährige Erfahrung im praktischen FDM haben, wurde das Tätigkeitsschema evaluiert.

Ziel des Workshops war, Feedback zu unseren Arbeiten von den Datenkurator:innen zu erhalten, um dadurch die aktuellen Arbeitsergebnisse zu verbessern und Perspektiven für die kommende Projektphase zu eruieren.

Zur Vorbereitung auf den Termin wurde eine Einführung/Erläuterung (PDF) und ein Datenblatt (Excel) mit dem ersten Entwurf des Tätigkeitsschemas an die Eingeladenen versendet. Zusätzlich wurden fünf Fragen formuliert, die die Kolleg:innen bei der Durchsicht der Materialien berücksichtigen sollten:

- Ist die Struktur als solche vollständig und in ihrer Komplexität angemessen, zu wenig oder zu viel?
- Zu den einzelnen Elementen der Struktur: Was fehlt, ist unklar oder falsch platziert?
- Welche Form der Präsentation sollten wir wählen, um das Material später zugänglich zu machen? Eine statische Dokumentation (z.B. PDFs) oder dynamischere Funktionen (z.B. ein Online-Tool/eine Informationsplattform)?
- Wie sollte die Erstellung und Verwaltung von Materialien auf der untersten Ebene organisiert werden - wer, wie, wann?

- Wie viel Konsistenzprüfung (über Datensammlungen hinweg) ist mittelfristig auf der untersten Ebene der Materialien erwünscht/durchführbar?

Die Erörterung der Fragen stand im Mittelpunkt des Workshops, jedoch wurde auch zu weiteren Kommentaren eingeladen. Sofern eine Teilnahme am Workshop nicht möglich war, konnten Antworten und Kommentare schriftlich an das Projektteam gesendet werden.

In der Bearbeitung der Fragen zeigte sich, dass das Schema an manchen Stellen unvollständig war, so fehlte z.B. zunächst der Punkt „Akquise". Dies lag einerseits daran, dass uns zu diesem Bereich wenig Arbeitsdokumentation erreichte, aber auch an der organisationbedingten Arbeitsteilung bei GESIS - die im Projekt direkt involvierten FDZ erhalten ihre Daten zumeist über seit Jahren feststehende Kanäle, wodurch aktive Akquise im Gegensatz zu anderen GESIS-Bereichen und FDZ außerhalb von GESIS praktisch weniger aktiv bewirtschaftet wird.

Des Weiteren wurden die Reihenfolge und Vollständigkeit der identifizierten Einzeltätigkeiten (Ebene 3) optimiert und Abhängigkeiten berücksichtigt.

Daneben gibt es Workflows, die nicht gleichförmig entlang des Zyklus verlaufen, wie z.B. die versionsbezogene Bearbeitung von Datensätzen. Sie bedienen sich, wie bereits erwähnt, punktuell der im Schema gelisteten Tätigkeiten oder Tätigkeitsbündel (Barkow, Block, Greenfield et al. 2013, Figure 6). Ein korrektes Abbilden dieser Prozesse wird durch den angestrebten modularen Aufbau der DT-Materialien ermöglicht.

3.3 Die Trainingsdokumente

3.3.1 Entwicklung der Dokumente

Im Anschluss an die Vervollständigung des Tätigkeitsschemas wurden aus den vorab gesichteten Materialien bestehende Arbeitsanweisungen ausgewählt, die es erlauben, exemplarisch einen Workflow mit seinen Einzeltätigkeiten zu beschreiben und die vorhandenen Dokumente in eine optimierte Form zu bringen. Der Zweck dabei war, einen Einblick zu gewinnen, ob das Präsentationsformat der Dokumente und der Detailgrad und Fluss der Anweisungen im eigentlichen Dokumentationstext ihren Informationszweck erfüllen und hinreichend zugänglich sind. Dabei wurde versucht, pro Tätigkeit (Ebene 3 des Schemas) je ein Dokument aus den vorhandenen Materialien auszuwählen und an das Tätigkeitsschema anzupassen (soweit notwendig).

Die interne Gliederung der Dokumente wurde folgendermaßen vorgenommen:

-	Titel	
-	Versionshistorie	Versionsnummer, Datum, Bearbeitende
1	Einleitung	Kurze, prägnante Beschreibung des beschriebenen Workflows
2	Zuständigkeiten und Berechtigungen	Ansprechpartner, die in den Prozess involviert sind; Zuständigkeiten;

		Berechtigungen für Laufwerke, Software etc.
3	Voraussetzungen	Vorarbeiten/Voraussetzungen, die erfüllt sein müssen, um den beschriebenen Arbeitsschritt durchzuführen
4	Beschreibung der Workflows	Erläuterung der Einzelschritte als Fließtext (ggf. mit Syntax)

Zusätzlich zu den beschreibenden Dokumenten wurde ein Workflowdokument erstellt, das die Abfolge der Dokumente gemäß der Reihenfolge der Tätigkeiten im Schema enthält und die Dokumente entsprechend verlinkt.

Im Rahmen des Projekts wurden sieben Beispieldokumente erarbeitet, die die Datensatzaufbereitung eines konkreten komparativen Surveys beschreiben (Anhang 2, um die Zugänglichkeit im Rahmen dieses Berichts zu gewährleisten wurden die als Einzeldokumente konzipierten Beschreibungen in einem PDF-Dokument zusammengestellt).

3.3.2 Evaluation und Verbesserung der Trainingsdokumente

Durch ein internes Testtraining mit drei Proband*innen wurden die neu entworfenen Arbeitsdokumente auf ihre Zugänglichkeit und ihre Verständlichkeit geprüft. Alle drei Testpersonen erhielten das Material erstmalig einen Tag vorab zur Lektüre, waren aber vorher weder mit in diesem Format aufbereitetem Material noch mit den im Training aufzubereitenden Daten irgendwie näher vertraut. Es handelte sich um eine sehr erfahrene Mitarbeiterin sowie zwei erst seit wenigen Wochen im Institut tätige Personen.

Es zeigte sich, dass die Teilnehmenden mit den Dokumenten einen überwiegend direkten Start in die Bearbeitung fanden und den Arbeitsablauf gemäß der Dokumente in der vorgesehenen Zeit weitgehend erledigen konnten (es war vorab erwartet worden, dass nicht alle dokumentierten Arbeitsschritte vollständig durchgeführt werden könnten, dies trat auch so ein).

Fragen, die während der Bearbeitung auftraten, hingen zumeist mit dem Verständnis der im Datensatz vorhandenen Variablen zusammen. Da auch die Beschreibung der Dublettenprüfung Probleme bereitete, wurde der Text des betreffenden Dokuments (Anhang 2, Kapitel „Dublettenprüfung“) im Nachgang konkretisiert. Ebenso wurde die Detailtiefe in der Beschreibung der einzelnen Tätigkeitsschritte angeglichen.

Des Weiteren wurde angemerkt, dass eine inhaltliche Auseinandersetzung mit dem Datensatz unbedingt vor Beginn der Arbeit mit den Dokumenten erfolgen sollte. Da dieser Punkt bisher nicht im Tätigkeitsschema aufgeführt war, wurde das Schema dahingehend erweitert.

Die Kapitel „Zuständigkeiten und Berechtigungen“ sowie „Voraussetzungen“ wurden als sinnvoll erachtet. Zusätzlich würde es sich anbieten, Links und Hinweise auf weitere, bei der Bearbeitung ggf. relevante Ressourcen (z.B. Gewichtungshinweise) ebenfalls in den Dokumenten aufzunehmen.

Kontroll- und Prüfschritte sollten ebenfalls in die Dokumente aufgenommen werden, sofern diese sinnvoll formuliert werden können (z.B. bei den Demographievariablen).

Da es sich um Trainingsdokumente handele, sei es sinnvoll, ein Dokument voranzustellen, in dem beschrieben wird, wie in die Bearbeitung des Datensatzes eingestiegen werden kann. Dies könne generisch oder auf eine spezielle Kollektion zugeschnitten sein. Ein solches Dokument wäre jedoch aktuell nicht mit unserem Tätigkeitsschema verbunden.

4 Zusammenfassung und Ausblick

4.1 Transfer

Die Ergebnisse des Projekts FDMontheground mit den dort entwickelten Trainings- und Dokumentationsworkflows zeigen, welche weitgehenden Spezifikationen allgemeiner FDM-Richtlinien notwendig sind, um institutsspezifische FDM-Aufgaben adäquat und zielgerichtet dokumentieren zu können. In ihrer Grundstruktur orientieren sich die Workflows an den einzelnen Phasen des Datenkurationszyklus. Daran anschließend wurden dann, auf einer Unterebene, institutsspezifische Aufgaben abgeleitet und entsprechende Maßnahmen benannt. Hieraus ergeben sich folgende Vorteile: Die detaillierte Struktur der Workflows macht die einzelnen Maßnahmen des Datenkurationsprozesses transparent und bietet den Mitarbeitenden Orientierung sowohl innerhalb der jeweiligen Schritte als auch zu den jeweils benachbarten Schritten. Sie können damit nach einer initialen Trainingsphase weitgehend autonom, ohne weitere externe Anleitung oder Hilfestellung, die Erfassung, Bearbeitung und Dokumentation durchführen.

Die im Projekt entwickelten Trainingsmaterialien sind demnach valide Vorlagen. Ihre feingranulare Struktur ermöglicht es, je nach Bedarf Anpassungen auf der Makro- oder Mikroebene durchzuführen. Somit können und sollen die Vorlagen als Referenzmaterialien für andere FDZ und deren spezifische Erfassungsprozesse dienen. Daher möchten wir es ermöglichen, die Ergebnisse des Projekts FDMontheground u.a. über RDM Compas einer größeren Öffentlichkeit und der KonsortSWD-Gemeinschaft zugänglich zu machen. Die in diesem Projekt gesammelten Erfahrungen und Ergebnisse können auch als Ausgangsbasis für weitere Projekte dienen, die sich analytisch mit internen FDM-Prozessen in Forschungsdatenzentren befassen.

4.2 Weitere Überlegungen zu Implementierung und Pflege eines Dokumentationssystems

Über die oben beschriebene Erfüllung der im Arbeitsplan deklarierten Projektziele hinaus ist das längerfristige Ziel der bisherigen Arbeiten, die nachhaltige Pflege eines hinreichend vollständigen Bestands an DT-Materialien vorzubereiten. Neben den hier behandelten Fragen der Strukturierung und ansatzweise der Gestaltung solcher Materialien ist dazu zuvorderst die Erstellung der konkreten Inhalte zu klären, die in höchstem Maße von spezifischen fachlichen, datenbezogenen Kriterien geprägt sein muss und daher hier nicht angemessen diskutiert werden kann. Generischer ist hingegen die Frage, *wie* DT-Materialien praktikabel zu erstellen, zu pflegen und zugänglich zu machen sind. Grundlegende Gedanken dazu wollen wir abschließend kurz präsentieren.

- Für eine lokale Implementierung praktisch relevant sind vor allem die Ebenen 2 (Phasen) und 3 sowie ggfs. Zusatzinformationen, welche Tätigkeiten aus Ebene 3 in welcher Reihenfolge auszuführen sind, wenn Workflows auch von Standardsequenzen abweichen können.
- Eine aktive Erstellung der eigentlichen Texte unterhalb von Ebene 3 ist bei der ersten Implementierung naturgemäß die aufwändigste Aufgabe. Diese Erstellung erfolgt idealerweise durch erfahrene Mitarbeiter:innen und wird von verantwortlichen Wissenschaftler:innen mindestens auf Korrektheit geprüft.
- Herausfordernd kann es sein, wenn die in einer Organisation kuratierten Daten mit großer Detailtiefe bearbeitet werden und in sich heterogen sind – dann müssen auch die Vorgehensregeln entsprechend detailliert und vielgestaltig sein. Allerdings kann sich die Ausdifferenzierung oft auf enge Tätigkeitsbereiche beschränken. Eine explizite schriftliche Dokumentation bietet dann auch im Vergleich über die ‚parallelen' Einzeldokumente die Gelegenheit zu reflektieren, wie viel an Varianz in den Bearbeitungstätigkeiten eigentlich notwendig und angemessen ist.
- Je größer der Bestand an DT-Material und je mehr Mitarbeiter:innen, desto wichtiger wird eine explizite Redaktionsregelung. Dabei liegt eine Trennung der Verantwortung für die Gliederungsebene – nur wenige Personen – und für die Textebene – mehr Personen, auf der Bearbeitungsebene – nahe. Dazu können auch Regelungen über explizite Aktualisierungsintervalle gehören, die jedoch je nach Gebiet unterschiedlich ausfallen werden.
- Vermutlich sehr relevant für jede Implementierung ist auch die Frage, auf welcher Art von Plattform die Dokumente vorgehalten und organisiert werden. Dateibasierte Lösungen wie Fileserver-Laufwerke können grundlegende Bedürfnisse wie Strukturierung und Volltext-Durchsuchbarkeit erfüllen und benötigen in der Regel keine oder nur wenige nicht ohnehin in einer Büroumgebung vorhandenen Software-Tools. Dafür sind sie wenig interaktiv, bieten keine grafische Benutzerführung und die Verwaltung von Zugriffsrechten für Dokumentänderungen etc. muss eventuell über IT-Administratoren erfolgen. Alternativ bieten sich Wiki-artige Lösungen an, die eine visuelle Baumstruktur zur Navigation anbieten, auch während die Nutzer:innen ‚im'

Detail-Text sind, gut durchsuchbar sind, Hypertext/Verlinkungen mit Querverweisen zwischen Dokumenten erlauben, und ein flexibles Online-Editieren sowohl der Struktur als auch der Inhalte ermöglichen. Ein zentraler Vorteil dürfte auch sein, dass diese Systeme für kollaboratives Editieren konstruiert sind, womit die Arbeitsteilung bei der Dokumentbearbeitung wesentlich einfacher und transparenter wird. Nachteilig können der höhere erste Implementationsaufwand einer entsprechenden Server-basierten internen Plattform und die von allen Redakteur:innen zu meisternde Lernkurve für das aktive Editieren von Inhalten sein. In den IT-affinen Arbeitsumgebungen der FDZ wird dies jedoch häufig kein Problem sein.

- Zu schließen ist mit dem erfahrungsbasierten Hinweis, dass die Pflege des DT-Materials nur nachhaltig erfolgen kann, wenn dies unter aktivem Monitoring der Organisationsleitung erfolgt und mit der entsprechenden Zuweisung von Arbeitszeit unterlegt ist.

5 Referenzen

Barkow, Ingo, William Block, Jay Greenfield, Arofan Gregory, Marcel Hebing, Larry Hoyle, and Wolfgang Zenk-Möltgen. 2013. Generic longitudinal business process model: DDI – documenting the helix. DDI Working Paper Series – Longitudinal Best Practices No. 5. doi: https://doi.org/10.3886/DDILongitudinal05.

Corti, Louise, van den Eynden, Veerle, Bishop, Libby, and Woollard, Matthew. 2020. Managing and Sharing Research Data: A Guide to Good Practice. 2nd ed. Los Angeles: SAGE Publications.

Higgins, Sarah. 2008. The DCC Curation Lifecycle Model. *International Journal of Digital Curation* 3 (1): 134–40. https://doi.org/10.2218/ijdc.v3i1.48.

Jensen, Uwe, Netscher, Sebastian, and Weller, Katrin (Hrsg.). 2019. Forschungsdatenmanagement sozialwissenschaftlicher Umfragedaten: Grundlagen und praktische Lösungen für den Umgang mit quantitativen Forschungsdaten. Opladen, Berlin, Toronto: Barbara Budrich. https://doi.org/10.3224/84742233.

RatSWD [Rat für Sozial- und Wirtschaftsdaten]. 2023. Forschungsdatenmanagement in kleinen Forschungsprojekten - Eine Handreichung für die Praxis. (RatSWD Output Series, 7. Berufungsperiode Nr. 3). https://doi.org/10.17620/02671.72

Riley, Jenn. 2017. Understanding Metadata: What is Metadata, and What is it For? https://www.niso.org/publications/understanding-metadata-2017.

Technische Universität Berlin. 2021. Informationen zu den eingeschränkten IT-Services an der TU Berlin. https://www.tu.berlin/themen/einschraenkung-it-services/, letzter Zugriff: 06.12.2023.

Wilkinson, Mark D., Dumontier, Michel, Aalbersberg, IJsbrand Jan et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018. https://doi.org/10.1038/sdata.2016.18.

6 Anhang

Die Dokumente

- Anhang 1_Tätigkeitsschema und
- Anhang 2_Trainingsdokumente

sind als Dateien in Zenodo hinterlegt (https://doi.org/10.5281/zenodo.10260847).

Anhang A

Information zum Datenlebenszyklus-Modell incl. Beispielen für mögliche Arbeitsdokumente

Prozessschritt	Beschreibung	Beispiele für mögliche Arbeitsprozesse und -dokumentationen
Erstellung & Empfang	Entgegennahme von Forschungsdaten von Datenerzeugern (ggf. auch anderen Archiven, Repositorien oder anderen Datenzentren) in Übereinstimmung mit den dokumentierten Richtlinien und ggf. Zuweisung geeigneter Metadaten.	• Workflowbeschreibung „Dateneingang“ • Metadatenkonventionen
Auswahl & Bewertung	Bewertung der eingereichten Forschungsdaten und Auswahl der Forschungsdaten, die für die langfristige Archivierung und Bewahrung in Frage kommen. Unter Befolgung der dokumentierten Leitlinien, internen Richtlinien oder rechtliche Anforderungen.	• Surveyrichtlinien • Bewertungskriterien Anonymisierung
Datenübernahme	Transfer der Forschungsdaten in das Archiv des Forschungsdatenzentrums unter Befolgung der dokumentierten Leitlinien, internen Richtlinien oder rechtliche Anforderungen.	• Workflowbeschreibung zur Datenübergabe an das Archivteam

Erhaltungsmaßnahmen	Ergreifen von Maßnahmen, die die langfristige Bewahrung und Beibehaltung des maßgeblichen Datencharakters gewährleisten. Die Bewahrungsmaßnahmen sollten sicherstellen, dass die Daten authentisch, zuverlässig und nutzbar bleiben und ihre Integrität erhalten bleibt. Zu den Maßnahmen gehören Datenbereinigung, Validierung, Zuweisung von Bewahrungsmetadaten, Zuweisung von Darstellungsinformationen und Gewährleistung akzeptabler Datenstrukturen oder Dateiformate.	• Beschreibung des Datenaufbereitungsworkflows (Beteiligte? Übergabedokumente? Hilfsmittel)? • Konvention zur Variablenbenennung • Regeln zur Variablencodierung (Schemata, Behandlung von Missings) • Datenaufbereitung (Dokumentation, Prüfungen, Analyseskripte, Fehlerbearbeitung [Dokumentation und Behandlung von Datenfehlern]) • Regeln zur Anonymisierung • Methodenbericht
Speicherung & Sicherung	Speicherung der Daten auf sichere Weise unter Einhaltung der einschlägigen Normen und Standards. Prüfregeln und –workflows. Erstellung von persistenten Identifikatoren (PIDs)	• Konzept der Datensicherung • Beschreibung einer speziellen Ordnerstruktur zur Ablage der Arbeitsfiles • Anforderungen an Dateiformate, Vollständigkeit der Dateien
Zugang & Nachnutzung	Sicherstellung, dass die Forschungsdaten für die Nachnutzung zugänglich sind. Dies kann je nach Datentyp in Form von öffentlich zugänglichen Datensätzen erfolgen, oder durch strenge Zugangskontrollen und Authentifizierungsverfahren geregelte Nachnutzungsverfahren.	• Dokumentation der Maßnahmen zur Langzeitarchivierung • Lizensierung oder Zuweisung von Zugangsklasse • Nutzungsstatistik
Transformation	Erstellen neuer Daten aus den Originaldaten, zum Beispiel • durch Migration in ein anderes Format oder • durch die Erstellung eines Teildatensatzes.	• Nachbearbeitung • Versionierungsdokumentation (Erstellung von Errata)

[eigene Darstellung in Anlehnung an: Higgins, 2008]

Impressum

Kontakt:

Dr. Markus Quandt
GESIS – Leibniz-Institut für Sozialwissenschaften
Unter Sachsenhausen 6-8
50667 Köln
www.gesis.org
markus.quandt@gesis.org

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

Zitation:

Behrens, K., Quandt, M., Zenk-Möltgen, W., Beck, K., & Brück, R. (2023). KonsortSWD Measure TA.2-M.12: FDMontheground - Projektbericht. Zenodo.
https://doi.org/10.5281/zenodo.10260847