RatSWD Working Paper Series

RatSWD.
Rat für Sozial- und WirtschaftsDaten

275

Forschungsdaten sichtbar machen: Der VerbundFDB-Harvester

Karoline Harzenetter,
Lisa Pegelow,
Dirk Weisbrod

Juni 2021

www.ratswd.de

Working Paper Series
des Rates für Sozial- und Wirtschaftsdaten (RatSWD)

Die *RatSWD Working Papers*-Reihe startete Ende 2007. In dieser Online-Publikationsreihe werden konzeptionelle und historische Arbeiten, die sich mit der Gestaltung der statistischen Infrastruktur und der Forschungsinfrastruktur in den Sozial-, Verhaltens- und Wirtschaftswissenschaften beschäftigen, publiziert. Dies sind insbesondere Papiere zur Gestaltung der Amtlichen Statistik, der Ressortforschung und der akademisch getragenen Forschungsinfrastruktur sowie Beiträge, die Arbeit des RatSWD selbst betreffend. Auch Papiere, die sich auf die oben genannten Bereiche außerhalb Deutschlands und auf supranationale Aspekte beziehen, sind besonders willkommen.

RatSWD Working Papers sind nicht-exklusiv, d. h. einer Veröffentlichung an anderen Orten steht nichts im Wege. Alle Arbeiten können und sollen auch in fachlich, institutionell und örtlich spezialisierten Reihen erscheinen.

Die Inhalte der *RatSWD Working Papers* stellen ausdrücklich die Meinung der jeweiligen Autorinnen bzw. Autoren dar und nicht die des RatSWD. Die Zuwendungsgeber des RatSWD haben die Publikationen nicht beeinflusst.

Forschungsdaten sichtbar machen: Der VerbundFDB-Harvester

Von Karoline Harzenetter (GESIS - Leibniz-Institut für Sozialwissenschaften, ORCID: 0000-0003-0097-7500), Lisa Pegelow (Institut zur Qualitätsentwicklung im Bildungswesen (IQB), Berlin, ORCID: 0000-0003-4148-6978) und Dirk Weisbrod (DIPF | Leibniz-Institut für Bildungsforschung und Bildungsinformation, Frankfurt am Main, ORCID: 0000-0002-9455-4527)

in Zusammenarbeit mit Sonja Bayer (DIPF), Marcus Eisentraut (GESIS), Nadeshda Jung (DIPF), Alexia Meyermann (DIPF), Claudia Neuendorf (IQB), Maike Porzelt (DIPF) und Jessica Trixa (GESIS)

doi: 10.17620/02671.62

Abstract

Die produktive Vielfalt der Datentypen der interdisziplinären und multimethodischen Bildungsforschung führt zu einer starken Heterogenität bei der Datenbeschreibung. Darunter leidet besonders die Auffindbarkeit der Forschungsdaten, was deren Nachnutzung erschwert. Ein Ziel des Verbund Forschungsdaten Bildung (VerbundFDB) ist es deshalb, die Sichtbarkeit von Daten über das VerbundFDB-Portal zu erhöhen und Forschungsdaten aus verschiedenen Forschungsdatenzentren (FDZ) an zentraler Stelle nachzuweisen.

Der VerbundFDB hat zu diesem Zweck einen gemeinsamen Metadatenstandard zur Beschreibung von Studien entwickelt, der sich stark an den Bedarfen der Community der Bildungsforschung orientiert. Dazu gehören etwa die Harmonisierung und Standardisierung der Metadatendokumentation, mit deren Hilfe die dezentral vorliegenden Datenbestände an zentraler Stelle sichtbar und damit recherchierbar gemacht werden können. Für den Ausbau seines zentralen Nachweissystems entwickelte der VerbundFDB mit und für seine Kooperationspartner eine technisch skalierbare Lösung (Harvesting), die die fortlaufende Integration verlässlicher Metadaten über die DOI-Registrierungsagentur da|ra ermöglicht.

Seit Mitte 2020 steht dieses niedrigschwellige Angebot zur Metadatenintegration allen da|ra-Nutzer*innen zur Verfügung und wird bereits von mehreren FDZ genutzt. Voraussetzung für die Integration der Datennachweise über da|ra ist die Verwendung des vom VerbundFDB definierten Metadatenstandards und die Bereitstellung von Informationen über die Daten im Umfang und Form eines definierten Metadaten-Kernsets (*VerbundFDB-Kernset Bildungsforschung*), mit dem die Qualität der Datenbeschreibung institutionenübergreifend sichergestellt werden soll. Die skizzierte Lösung ist grundsätzlich auch auf andere Forschungsfelder adaptierbar.

1. Einleitung

Der Verbund Forschungsdaten Bildung (VerbundFDB) entstand 2013 mit der Zielsetzung, Angebote und Services für die empirische Bildungsforschung in den Bereichen Datenarchivierung und Datenbereitstellung zur Verfügung zu stellen sowie Beratung zum Forschungsdatenmanagement zu entwickeln und anzubieten. Seitdem wurde unter anderem ein effektiver Workflow der verteilten Archivierung realisiert, der es Forschenden ermöglicht, Forschungsdaten an den VerbundFDB zu melden und zu übermitteln, die dann von den Fachleuten der an dem VerbundFDB beteiligten Forschungsdatenzentren (FDZ) unter Einhaltung der dafür geltenden gesetzlichen Bestimmungen hinsichtlich Datenschutz und Urheberrecht bearbeitet, erschlossen und veröffentlicht werden. Leitgedanke ist, dass die Forschungsdaten an jenes Partner-FDZ weitergegeben werden, welches für die Erschließung des jeweiligen Datentyps die größte Expertise besitzt. Derzeit besteht der Kreis der föderierten FDZ aus dem FDZ am IQB (Institut zur Qualitätsentwicklung im Bildungswesen) für Daten der Kompetenz- und Leistungsmessung, dem GESIS (Leibniz-Institut für Sozialwissenschaften e. V.) für Umfrage- und Aggregatdaten, dem FDZ Bildung am DIPF (Leibniz-Institut für Bildungsforschung und Bildungsinformation) für qualitative Daten, dem FDZ des DZHW (Deutsches Zentrum für Hochschul- und Wissenschaftsforschung) für Daten aus dem Bereich der Hochschulforschung und dem FD-LEX (Forschungsdatenbank Lernertexte) für Lernertexte[1].

Neben der Archivierung ist der zentrale Nachweis von Forschungsdaten für den deutschsprachigen Raum, aber in Zukunft auch international, eine bedeutsame Aufgabenstellung des VerbundFDB. Dabei sind nicht nur die Metadaten der Forschungsdaten zu berücksichtigen, die von Forschenden an den VerbundFDB übergeben wurden, sondern auch jene, die bereits andere FDZ der Bildungsforschung in ihren Repositorien vorhalten.

Deswegen erarbeitet der VerbundFDB in Kooperation mit den FDZ technische Lösungen zum Einsammeln relevanter Metadaten über Schnittstellen. Zuletzt wurde ein verbundseitiger Harvester entwickelt, der die OAI-PMH[2]- Schnittstelle der DOI-Registrierungsagentur da|ra[3] zum Zweck der Nachweiserweiterung des VerbundFDB abfragt. Diese Lösung ermöglicht es fortan, dass FDZ, die da|ra für die DOI-Registrierung ihrer Forschungsdaten nutzen, Datennachweise auch direkt an den VerbundFDB übergeben können.

Um eine Mindestqualität der solcherart übermittelten Metadaten zu gewährleisten, hat der VerbundFDB zudem ein Kernset für die Bildungsforschung, das so genannte *VerbundFDB-Kernset Bildungsforschung*, entwickelt und verabschiedet, das aus 17 Metadaten-Angaben zu Provenienz, Methodik und Inhalt besteht. Nur wenn dieses *VerbundFDB-Kernset*

[1] Lernertexte sind z. B. die im BMBF-Projekt *Unterrichtliche Förderung von Teilkomponenten der Schreibkompetenz* entstandenen Schülertexte, die von der Forschungsdatenbank FD-Lex (https://fd-lex.uni-koeln.de) archiviert und bereitgestellt werden.
[2] Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) ist ein web-basiertes standardisiertes Protokoll, das es über eine Client-Anwendung ermöglicht, Metadaten abzufragen und einzusammeln (Harvesting) http://www.openarchives.org/pmh/).
[3] https://www.da-ra.de/home/

Bildungsforschung im da|ra-Metadatensatz vollständig vorliegt, wird der Datensatz in die VerbundFDB-Datenbank importiert.

Nachfolgend sollen die Entwicklung und die Funktionalitäten des *VerbundFDB-Harvester*s und des *VerbundFDB-Kernsets Bildungsforschung* sowie die Evaluation der Harvesting-Lösung beschrieben werden.

2. Auf dem Weg zum VerbundFDB-Harvester

Die Entwicklung des *VerbundFDB-Harvester*s ist das Ergebnis eines längeren Abstimmungsprozesses zwischen führenden FDZ der Bildungs- und Sozialforschung. Die drei Kernpartner des VerbundFDB (DIPF, GESIS und IQB) luden im März 2017 die Institutionen ApaeK, BIBB, DIE, DIW Berlin, DJI, DZHW, IAB, LIfBi, Qualiservice Bremen und ZPID sowie die Leitung der Geschäftsstelle des RatSWD ein, um mit ihnen im Rahmen eines Netzwerkstreffens über das Thema „Harmonisierung und Bündelung von Metadaten in der Bildungsforschung" zu diskutieren.[4] Ein Ergebnis des Netzwerktreffens war die Gründung einer Arbeitsgruppe zur Metadaten-Harmonisierung (*AG Metadaten*) unter Beteiligung des DZHW, DIW Berlin, LIfBi und Qualiservice Bremen und der VerbundFDB-Kernpartner (DIPF, GESIS, IQB). Der VerbundFDB richtete die Treffen dieser Arbeitsgruppe aus, die sich die folgenden Ziele gesetzt hatte:

1. den Aufwand für die Metadaten-Harmonisierung zwischen dem VerbundFDB und den FDZ sowie die damit verbundenen Konsequenzen – z. B. für die Metadatenkuratierung und Doppelerfassung von Forschungsdaten in unterschiedlichen Systemen – detailliert zu beschreiben,
2. Lösungen für den Metadatenaustausch zwischen FDZ und VerbundFDB zu entwickeln und
3. ein Metadaten-Kernset für die Bildungsforschung zu erarbeiten, welches vom gesamten VerbundFDB-Partnerkreis[5] als Qualitätsstandard zur Beschreibung von Forschungsdaten genutzt werden kann.

Ergebnis der Beratungen war schließlich zum einen die Konzeption und Entwicklung einer Clientanwendung (VerbundFDB-Harvester), um die Schnittstelle der Plattform da|ra für die Nachweiserweiterung zu nutzen, sowie zum anderen das *VerbundFDB-Kernset Bildungsforschung*, welches Mindestanforderungen für die Auszeichnung eines Datensatzes mit Metadaten festlegt, der über den VerbundFDB in dessen Suchportal angezeigt werden soll.

[4] Karoline Harzenetter: Bericht vom ersten Netzwerktreffen des Verbunds Forschungsdaten Bildung. Online unter: https://www.forschungsdaten-bildung.de/get_files.php?action=get_file&file=VFDB_NT1_Bericht01_201705.pdf
[5] https://www.forschungsdaten-bildung.de/vernetzung

2.1. Warum da|ra?

da|ra ist eine Registrierungsagentur für die Vergabe von dauerhaften digitalen Identifikatoren (= Persistent Identifier, PID) – in diesem Fall von Digital Object Identifiern (DOI) – im Bereich der Sozial- und Wirtschaftsdaten. Sie ist somit Anlaufstelle vieler FDZ mit Bildungsforschungsdaten in Deutschland. da|ra bietet ihren Service darüber hinaus aber auch für Institutionen weltweit an. Bei der DOI-Registrierung von Forschungsdaten ist die Angabe von Metadaten in einem geringen Umfang (6 Elemente[6]) verpflichtend. Das bedeutet, in da|ra liegt eine große Menge[7] an persistenten Datennachweisen vor, die teilweise thematisch für das VerbundFDB-Portal von Interesse sind.

2.2. Zielgruppe

Besonders vorteilhaft für die Fragestellungen der vom VerbundFDB eingerichteten Arbeitsgruppe *AG Metadaten* war aber, dass da|ra darüber hinaus ein komplettes Metadatenschema (32 Hauptelemente mit zahlreichen Subelementen) für eine erweiterte Beschreibung der zu registrierenden Forschungsdaten bereitstellt.[8] Dieses Schema verwendet kontrollierte Vokabulare (orientiert am Standard der Data Documentation Initiative, DDI[9]), die ebenfalls vom VerbundFDB genutzt werden. Es verfügt zudem über die Option, Metadaten stark zu strukturieren und so Forschungsdaten differenziert und übersichtlich zu beschreiben. Das Schema unterstützt außerdem die Mehrsprachigkeit von Metadaten. Über eine Schnittstelle, die gemäß der Open Archives Initiative (OAI) die Interoperabilität von Metadaten nach internationalem Standard gewährleistet[10], können bei daIra Metadaten jederzeit abgefragt, geharvestet und weiterverwendet werden. Hier konnte der VerbundFDB andocken und den FDZ anbieten, ihre bei da|ra hinterlegten Metadaten direkt auch an den VerbundFDB weiterzureichen. Da da|ra bei GESIS, einem Kernpartner des VerbundFDB, angesiedelt ist, konnten sehr schnell erste Gespräche für die Entwicklung einer für den VerbundFDB erweiterten OAI-PMH-Schnittstelle aufgenommen werden. Schließlich entschloss sich die Arbeitsgruppe, die Entwicklung einer solchen Harvesting-Lösung in Angriff zu nehmen. Die technische Entwicklung der Harvesting-Anwendung lag dabei beim DIPF.

Zur Zielgruppe des Harvesters gehören:

1. Netzwerkpartner des VerbundFDB, die zur Integration ihrer Datennachweise in die VerbundFDB-Datenbank automatisierte Prozesse nutzen, dafür aber nicht eine eigene Entwicklung (z. B. offene Schnittstelle für den VerbundFDB Harvester,

[6] In erster Line Informationen bibliografischer Natur zum Zwecke der Datenzitation: resourceType', 'title', 'creators', 'publicationDate', 'availability' und 'dataURL'.
[7] Derzeit über 60.000 Registrierungen (Stand: Dezember 2020)
[8] Koch, Ute et al.: da|ra Metadata Schema. Documentation for the Publication and Citation of Social and Economic Data. GESIS Papers, 2017-25. Version: 4.0. GESIS Leibniz Institute for the Social Sciences. Text. URL: https://doi.org/10.4232/10.mdsdoc.4.0
[9] https://ddialliance.org/
[10] Siehe hierzu auch die Selbstbeschreibung der Open Archives Initiative (OAI): "The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) is a low-barrier mechanism for repository interoperability. Data Providers are repositories that expose structured metadata via OAI-PMH. Service Providers then make OAI-PMH service requests to harvest that metadata. OAI-PMH is a set of six verbs or services that are invoked within HTTP." Online unter: https://www.openarchives.org/pmh.

Metadatenmapping mit dem Metadatenset des VerbundFDB) anstoßen wollen. Das ist etwa dann sinnvoll, wenn generell nur ein sehr kleiner Bestand oder eine Teilmenge eines größeren Bestands übernommen werden soll.

2. Institutionen/Datenzentren, die nicht mit dem VerbundFDB kooperieren, aber im VerbundFDB-Portal[11] eigene Daten der Bildungsforschung nachweisen wollen.

Zusammenfassend kann man sagen, dass sich das Harvesting-Angebot des VerbundFDB als niedrigschwelliges Angebot an FDZ richtet, die bereits ihre Forschungsdaten bei da|ra zum Erhalt einer DOI registrieren und dabei ohne großen Eigenaufwand, aber unter Wahrung einer gewissen Mindestqualität, direkt ihre Metadaten an den VerbundFDB übermitteln wollen. Zusätzliche technische Eigenentwicklungen oder gar Anpassungen des eigenen Metadatenschema sind nicht notwendig. In der Regel reichen geringe Anpassungen und Aktualisierungen der Metadatennachweise bei der Registrierung in da|ra aus, um die Vorgaben des VerbundFDB zu erfüllen.

Die beschriebene Vorgehensweise hat für den Verbund FDB zudem den Vorteil, dass ein Großteil der vom RatSWD akkreditierten FDZ ihre Forschungsdaten bei da|ra bereits registrieren oder die Registrierung planen. Er erhält damit auch Datennachweise, die ihm sonst womöglich entgehen würden – und dies bereits bei deren Veröffentlichung oder Aktualisierung.

Gleichzeitig ist jedoch zu erwähnen, dass sich die Datendokumentation in da|ra im Gegensatz zu der bildungsforschungsspezifischen Ausrichtung der Dokumentation im VerbundFDB-Portal nicht ausschließlich an den Bedarfen der Bildungsforschung orientiert und deswegen nur bedingt flexibel auf Anpassungen in diesem Bereich reagieren kann. Derzeit wird deshalb auch an einer direkten Schnittstelle zwischen VerbundFDB und einigen anderen Partnern (DZHW und FD-LEX) gearbeitet. Dieser direkte Weg lohnt sich insbesondere dann, wenn Ressourcen auf Seiten des FDZ für die Erarbeitung eines eigenen Mappings und die technische Umsetzung vorhanden sind und langfristig zur Übermittlung von Datennachweisen mit dem VerbundFDB kooperiert werden soll.

2.3. Mapping der Metadatensets

Voraussetzung für das automatisierte Harvesting via da|ra war das Festsetzen der Regeln, nach denen die Metadaten in da|ra in das VerbundFDB-Metadatenschema[12] überführt und in die VerbundFDB-Datenbank integriert werden: das so genannte Metadaten-Mapping. Dabei geht es um das Abstimmen und In-Bezug-Setzen von Datenelementen aus zwei verschiedenen Datenmodellen.

[11] https://www.forschungsdaten-bildung.de/studienliste.php
[12] vgl. Verbund Forschungsdaten Bildung (2019): Metadatenset des VerbundFDB. Version 1.0, fdbinfo Nr. 8. URL: https://www.forschungsdaten-bildung.de/files/fdbinfo_8_Metadatenset_v1.0.pdf

Beim Mapping musste geprüft werden,

- welche Informationen in den verschiedenen Schemata vorliegen (Identifikation),
- wie diese zueinander passen und von Seiten des VerbundFDB bzw. da|ra passend gemacht werden können (Transformation) sowie
- welche Felder verpflichtend vorliegen müssen, um eine Mindestqualität der Datensätze im Zielsystem, in diesem Fall VerbundFDB, zu erhalten (Konvention).

Der letzte Punkt war eng mit der Entwicklung des *VerbundFDB-Kernsets Bildungsforschung* verbunden, das für den *VerbundFDB-Harvester* entwickelt wurde.

Besondere Herausforderungen beim Mapping waren die Prüfung der Kompatibilität und die Festschreibung der Transformation. So mussten die von da|ra genutzten Versionen kontrollierter Vokabulare in das Schema des VerbundFDB überführt und die unterschiedlichen Dokumentationssprachen und Grade der Metadatenstrukturierung in den beiden Systemen auf Kompatibilität geprüft werden. Kontrollierte Vokabulare, die bei der Metadatenerfassung im VerbundFDB obligatorisch sind, aber im daIra-Metadatenset nicht verwendet werden, mussten umgekehrt in da|ra abbildbar gemacht werden.[13]

Wenn etwa auf Seiten von da|ra kein vergleichbares kontrolliertes Vokabular für ein Metadatum vorlag, wie beispielswiese bei der Erhebungseinheit, beschloss man als Behelfslösung, den kontrollierten VerbundFDB-Term in da|ra als Freitext-String im da|ra-Metadatenschema unter „notes" eintragen zu lassen[14]. Umgekehrt standardisiert und strukturiert da|ra Informationen zum Untersuchungsgebiet[15], die der VerbundFDB in dieser Tiefe und Struktur bisher nicht benötigte. Hier war dann eine Anpassung auf Seiten des VerbundFDB notwendig.

Da in da|ra kontrollierte Vokabulare nur in englischer Sprache vorliegen, wurden diese für die VerbundFDB-Datenbank ins Deutsche übersetzt. Bei Versionsunterschieden der verwendeten

[13] Die Kontrollierten Vokabulare sind in der Metadatenregistry des Verbund Forschungsdaten Bildung (https://mdr.iqb.hu-berlin.de/#/) hinterlegt und öffentlich zugänglich:
- Resource Typ (KV VerbundFDB: Datenformat Bildungsforschung: https://mdr.iqb.hu-berlin.de/#/catalog/7e812b6c-ee57-a075-eb33-596e4e58f8ab)
- Resource Typ Free (KV VerbundFDB: Datentyp Bildungsforschung: https://mdr.iqb.hu-berlin.de/#/catalog/5c4748e0-bcaa-c3c3-e5af-dd481fddbf0c)
- Time Dimension (KV VerbundFDB: Erhebungsdesign Bildungsforschung): https://mdr.iqb.hu-berlin.de/#/catalog/56cc4164-6731-7d54-c97f-ad9bd24bf1b7)
- Time Dimension Free (KV VerbundFDB: Forschungsdesign Bildungsforschung: https://mdr.iqb.hu-berlin.de/#/catalog/89724b81-2aab-d0dc-29ca-87f29b1c9e01)
- Collection Mode (KV VerbundFDB: Erhebungsverfahren Bildungsforschung: https://mdr.iqb.hu-berlin.de/#/catalog/48f296c7-fba6-475c-e53b-3cee66b27453 bzw. KV VerbundFDB: Erhebungsmethode Bildungsforschung: https://mdr.iqb.hu-berlin.de/#/catalog/dfa9cfa4-b86f-42d0-d370-684694e5bc80
- Sampling (KV VerbundFDB: Auswahlverfahren Bildungsforschung: https://mdr.iqb.hu-berlin.de/#/catalog/1d791cc7-6d8d-dd35-b1ef-0eec9c31bbb5)
- Note (KV VerbundFDB: Erhebungseinheit Bildungsforschung: https://mdr.iqb.hu-berlin.de/#/catalog/94d1ae4f-a441-c728-4a03-adb0eb4604af)

[14] Siehe Verbund Forschungsdaten Bildung (2019): Kernset und da|ra-Harvesting im VerbundFDB. Version 1.0, fdbinfo Nr. 7 (S. 15). Online unter: https://www.forschungsdaten-bildung.de/files/fdbinfo_7.pdf

[15] In da I ra kann Untersuchungsgebiet bis auf die Ebene von Regionen und subnationalen Verwaltungseinheiten weltweit beschrieben werden. Im VerbundFDB wurde bisher nur die Staatenebene (Deutschland) bis maximal auf Ebene der Bundesländer ohne standardisierte ISO-Codierung erfasst.

kontrollierten DDI-Listen mussten zudem Terme für die VerbundFDB-Datenbank ggf. aggregiert oder subsumiert werden. Dabei handelte es sich aber um sehr geringfügige Abweichungen mit minimalem Informationsverlust.

Die kontrollierten Vokabulare des VerbundFDB sind für alle Nutzenden transparent in der Metadatenregistry (MDR)[16] des VerbundFDB hinterlegt. Sie sind frei zugänglich und nutzbar. Die MDR dient grundsätzlich als zentrale Dokumentationsquelle für die kontrollierten Vokabulare des VerbundFDB sowie für das *VerbundFDB-Kernset Bildungsforschung*. Die Inhalte sind unter einer Creative Commons (CC BY-SA 4.0)[17] lizenziert und können von jedem FDZ auch für die eigenen Metadatensets übernommen werden. Details des Mappings der Elemente von da|ra mit dem Vokabular des VerbundFDB sind ebenfalls in der MDR und in einer Veröffentlichung des VerbundFDB beschrieben.[18]

3. Entwicklung und Anwendung des VerbundFDB-Kernsets Bildungsforschung

Die Durchführung des Mappings war eng mit der Erarbeitung des *VerbundFDB-Kernsets Bildungsforschung* verbunden. Im Vordergrund stand dabei die Auswahl von Metadatenfeldern, die für eine aussagekräftige Mindestauszeichnung eines Datennachweises zwingend erforderlich sind. In intensiven Diskussionen wurden formale/bibliografische (etwa Titel, Primärforschende/Autor*innen, Veröffentlichungsdatum der Daten, persistenter Identifikator/DOI-Name), inhaltliche (etwa Zusammenfassung/Abstract, Schlagwörter) und methodische Informationen (etwa Forschungsdesign, Auswahl- und Erhebungsverfahren, Stichprobe) identifiziert, die diese Anforderung erfüllen. Dabei wurde auf die Expertise der *AG Metadaten*-Mitglieder, aber auch von Forschenden zurückgegriffen. Am Ende definierte die *AG Metadaten* 16 plus 1 Metadatenfelder, die für die an den VerbundFDB übermittelten Datennachweise obligatorisch sein sollten. Das zusätzliche Metadatum ist nicht im Sinne der Beschreibung, jedoch im Hinblick auf das Harvesting selbst ein obligatorischer Bestandteil des *VerbundFDB-Kernsets Bildungsforschung*: Es ist der so genannte Flag „VerbundFDB", der in da|ra über das Metadatum „Alternative Identifier" gesetzt werden kann. Nur solcherart ausgezeichnete Datensätze werden vom Harvester abgefragt und an den VerbundFDB übermittelt.

Des Weiteren wurden optionale Felder definiert, die zwar zwischen da|ra und VerbundFDB gemappt wurden, aber nicht zwingend Bestandteil des zu harvestenden Datennachweises sein müssen.

[16] https://mdr.iqb.hu-berlin.de/#/catalogs
[17] https://creativecommons.org/licenses/by-sa/4.0/deed.de
[18] Verbund Forschungsdaten Bildung (2019): Kernset und da|ra-Harvesting im VerbundFDB.

Das Metadatenset *VerbundFDB-Kernset Bildungsforschung* umfasst insgesamt die folgenden Felder:

Zentrale (verpflichtend) zu erfassende Elemente

- Resource Type (Datentyp/Datenformat)
- Title (Studientitel)
- Creator (beteiligte Forschende)
- DOI (Datenzugang/Identifier)
- Publication Date (Veröffentlichungsdatum)
- Availability (Verfügbarkeit/Zugangsbedingungen)
- Contributor/ Institution (Archivierende Einrichtung)
- Sampled Universe (Grundgesamtheit/Population/Stichprobe)
- Sampling (Auswahlverfahren)
- Temporal Coverage (Erhebungszeitraum)
- Time Dimension (Erhebungsdesign)
- Collection Mode (Erhebungsverfahren/Erhebungsmethode)
- Descriptions (Abstract)
- Geographic Coverage (Untersuchungsgebiet)
- Keywords (Schlagwörter)
- Note (Erhebungseinheit)
- Alternative Identifiers (Flag setzen)

Weitere wichtige (optionale) Elemente

- Angaben zum Projekt in dessen Rahmen die Forschungsdaten generiert wurden
- Time Dimension Free (Forschungsdesign)
- Beteiligte Institutionen
- Note (Anmerkungen zu den Daten / zur Studie)
- Publikationen

FDZ, die Datennachweise über da|ra an den VerbundFDB übermitteln möchten, müssen somit folgendes tun:

1. Den betreffenden da|ra-Datensatz mit dem Alternative Identifier „VerbundFDB“ auszeichnen und
2. den Datensatz daraufhin überprüfen, ob er die anderen 16 obligatorischen Metadaten des *VerbundFDB-Kernsets Bildungsforschung* bereits enthält und diese gegebenenfalls nachtragen / in da|ra aktualisieren.
3. Falls ein FDZ für die DOI-Registrierung der Daten eine Schnittstelle zwischen dem eigenen System und da|ra entwickelt hat, muss es ggf. diese Schnittstelle und das Mapping (FDZ-eigenes Schema – da|ra-Schema) anpassen, um die für den Verbund obligatorischen Metadaten korrekt zu liefern.

4. Falls ein FDZ für die DOI-Registrierung das da|ra-Webformular nutzt, kann es die Metadaten direkt manuell über die daIra-Benutzeroberfläche eintragen und bearbeiten.[19]
5. Zudem sollte jedes FDZ, welches seine Daten harvesten lassen will, den VerbundFDB darüber informieren und für Rückmeldungen dort seine Kontakt-E-Mail-Adresse hinterlegen.

Nach dem Setzen des Flags und der Aktualisierung der Änderungen wird der Datensatz beim nächsten Harvesting-Job des VerbundFDB über die OAI-Schnittstelle von da|ra abgefragt.

Ein FDZ muss somit unter Umständen seine Dokumentations- und Registrierungs-Workflows nachjustieren, fehlende Informationen für den VerbundFDB-Mindeststandard nachrecherchieren und – wenn es die DOI-Registrierung über eine automatisierte Schnittstelle zu daIra durchführt – kleinere technische Modifikationen vornehmen, um seine Metadaten über da|ra im VerbundFDB nachzuweisen. Erste Rückmeldungen zeigen jedoch, dass sich der Aufwand in Grenzen hält. Der Lohn für diese Arbeit sind sichtbarere und auffindbare Forschungsdaten.

4. Wie funktioniert der VerbundFDB-Harvester?

Der *VerbundFDB-Harvester* fragt in regelmäßigen Abständen die OAI-Schnittstelle von da|ra ab. Dies geschieht durch eine „Arbitrary Question", die durch da|ra definiert wurde. Dabei werden während eines Harvesting-Jobs alle Datensätze in der da|ra-Datenbank abgefragt, die im Metadatum „Alternative Identifier" für das VerbundFDB-Harvesting ausgezeichnet sind.

Verbundseitig wurde für Verarbeitungs- und Prüfschritte, die vor dem endgültigen Import des geharvesteten Datensatzes in die VerbundFDB-Datenbank notwendig sind, ein Zwischenspeicher entwickelt. Dieser Zwischenspeicher verfügt über eine Benutzeroberfläche, worüber der *VerbundFDB-Harvester* verwaltet werden kann. Hier können auch neue Abfragen definiert oder weitere Schnittstellen eingebunden werden. Das Harvesting kann sowohl manuell (zu Testzwecken) als auch automatisch, in vorher definierten Zeitintervallen, ausgeführt werden. Der Regelbetrieb erfolgt automatisch und wird durch die folgenden Cronjobs gesteuert:

- do-jobs: Abfrage und Harvesting von Datensätzen über die da|ra-Schnittstelle, die für das VerbundFDB-Harvesting markiert wurden. Die Datensätze werden im Zwischenspeicher abgelegt.
- perform: Prüfroutinen im Zwischenspeicher auf Vollständigkeit des *VerbundFDB-Kernsets Bildungsforschung*, auf Dubletten in der VerbundFDB-Datenbank und Metadaten-Updates.
- import: Import geprüfter Datensätze in die VerbundFDB-Datenbank und Freischaltung im VerbundFDB-Portal.

[19] Das vom VerbundFDB veröffentlichte Manual beschreibt, wie Metadaten über die Benutzeroberfläche von da|ra bearbeitet werden können: vgl. Verbund Forschungsdaten Bildung (2019): Kernset und da|ra-Harvesting im VerbundFDB, a.a.O.

- clean: Löschung von unvollständigen Datensätzen, die nicht in die Datenbank importiert wurden, sowie der als „geprüft“ markierten Dubletten und Updates aus dem Zwischenspeicher.

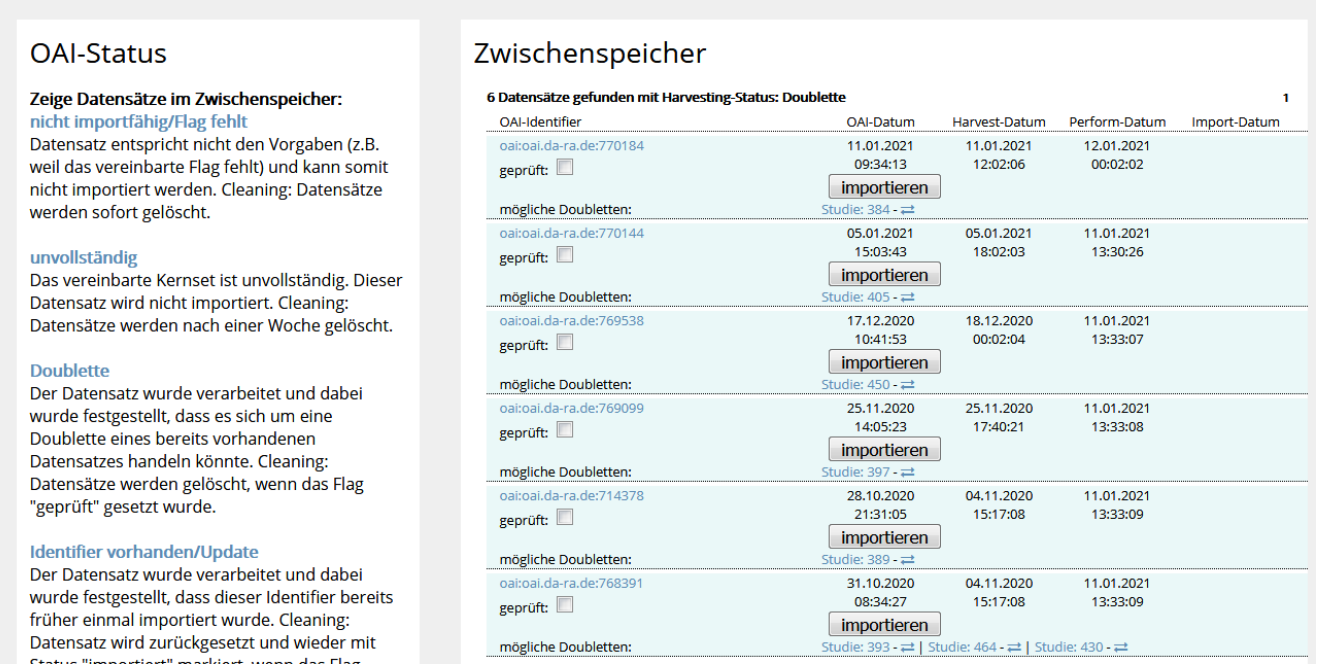

Abbildung 1: Prüfansicht eines Datensatzes im Zwischenspeicher des *VerbundFDB-Harvesters*. Rechts sind die Prüffälle aufgelistet, die bei Bedarf angeklickt und bearbeitet werden können.

Die Nachweise von da|ra werden als XML-Dateien geharvestet, die maximal 50 Datensätze enthalten können. Stehen während eines Jobs mehr Datensätze zur Verfügung, wird eine zweite Datei angelegt. Diese XML-Dateien werden in den Zwischenspeicher importiert und dort automatisch verarbeitet (Cron-Job „perform“).

Diese Verarbeitungs-Routine prüft zuerst die OAI-ID, die jeder Datensatz in da|ra erhält, und dessen Zeitstempel. Weiterverarbeitet bzw. zwischengespeichert werden nur Datensätze, die seit dem letzten Harvesting-Job neu hinzugekommen sind (d. h.: eine neue OAI-ID haben) oder upgedatet wurden (d. h.: eine bereits vorhandene OAI-ID haben sowie einen Zeitstempel, der aktueller ist als beim letzten Harvesting-Job). Alle anderen Datensätze werden nicht verarbeitet.

Zudem werden die folgenden Prüfschritte durchgeführt:

- **Datensatz unvollständig:** Dieser Prüfschritt erkennt Datensätze, bei denen ein Metadatum oder mehrere der obligatorischen Metadaten des *VerbundFDB-Kernsets Bildungsforschung* fehlen. Der unvollständige Datensatz wird nicht in die VerbundFDB-Datenbank importiert. Der „Publication Agent“ (Bezeichnung für das FDZ, von dem der Datennachweis stammt) wird durch eine automatisch generierte E-Mail über die unvollständigen Datensätze informiert. Hierfür hinterlegt jedes FDZ eine E-Mail-Adresse im Harverster-System des VerbundFDB, die mit dem

entsprechenden Publication Agent verknüpft ist. Für jeden unvollständigen Datennachweis (Angabe der OAI-ID und der DOI) sind in dieser E-Mail die fehlenden Metadaten aufgelistet – mit der Bitte an das jeweilige FDZ, ihn entsprechend zu ergänzen. Nach Ergänzung des Datensatzes wird dieser beim nächsten Harvesting-Job via Zeitstempel als aktualisiert erkannt und – wenn er vollständig ist – in die VerbundFDB-Datenbank importiert.

- **Dublette:** Die Prüfung erfolgt anhand eines Abgleichs der DOI und des Titels des Datensatzes. Sind diese bereits in der VerbundFDB-Datenbank vorhanden, wird ein geharvesteter Datennachweis nicht automatisch importiert. Stattdessen muss er vom VerbundFDB manuell daraufhin überprüft werden, ob tatsächlich eine Dublette vorliegt oder doch neue, relevante Informationen übermittelt wurden.
- **Update:** Die Prüfung erfolgt anhand der OAI-ID und des Zeitstempels. In der Regel werden Updates automatisch in die VerbundFDB-Datenbank importiert. Für eine manuelle Prüfung werden nur jene Datennachweise herausgefiltert, deren schon importierter Vorgänger manuell in der VerbundFDB-Datenbank bearbeitet wurde, etwa um Schreibfehler zu bereinigen. Ein solcher Datensatz wird in der VerbundFDB-Datenbank entsprechend markiert. Liegt eine solche Markierung vor, erfolgt kein automatischer Import. In diesem Fall muss der Datensatz manuell geprüft und gegebenenfalls manuell importiert werden.

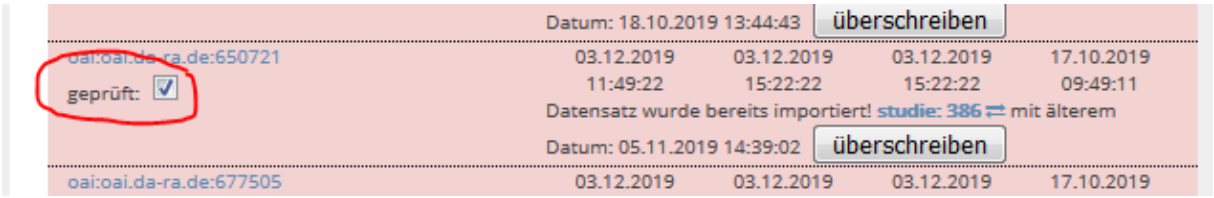

Abbildung 2: Datensatz wurde geprüft und entsprechend markiert

Dubletten und Updates werden im Zwischenspeicher aufgelistet. Anhand dieser Liste können Mitarbeitende des VerbundFDB jeden einzelnen Datensatz aufrufen, prüfen und durch Anklicken eines Buttons manuell in die VerbundFDB-Datenbank importieren. Geprüfte Datensätze, die nicht importiert werden sollen, werden manuell geflaggt, damit die nächste Cron-Job-Routine sie automatisch löschen kann. Zudem gibt es eine Liste mit allen importierten Datensätzen. Diese wird stichprobenartig gesichtet, um sicherzustellen, dass es keine Fehler bei der automatisierten Verarbeitung gegeben hat.

Der Cron-Job „import" sorgt dafür, dass Datensätze, die in „perform" als importfähig erkannt wurden, in die Datenbank importiert und direkt für die Suche freigeschaltet werden. „clean" sorgt dafür, dass alte XML-Dateien regelmäßig gelöscht werden.

Es ist vorgesehen, dass die Datensätze vollautomatisch in den VerbundFDB importiert werden. Die über den *VerbundFDB-Harvester* integrierten Datensätze werden auch – mit Ausnahme der Stichproben und der Dubletten- und Updateprüfung – nicht weiter manuell bearbeitet, denn

die Verantwortung für die Vollständigkeit und Qualität der Beschreibung liegt bei den Forschungsdatenzentren, die ihre Datennachweise bei da|ra verwalten. Die über da|ra geharvesteten Datennachweise werden durch das Metadatum „da|ra-Import" für Forschende und andere Nutzende in der VerbundFDB-Suche sichtbar gemacht.[20]

5. Evaluation und Testdaten von GESIS

Derzeit läuft der *VerbundFDB-Harvester* noch im manuellen Betrieb, d. h., die oben aufgeführten Cron-Jobs werden manuell durch einen Administrator einmal wöchentlich ausgeführt. Währenddessen wird der *VerbundFDB-Harvester* evaluiert und seine Funktionen anhand konkreter Metadaten getestet. Dabei wird auch die korrekte Umsetzung des Metadaten-Mappings überprüft. Dies geschieht unter anderem in enger Zusammenarbeit mit GESIS, dessen circa 450 Studiennachweise aus dem Bereich der Bildungsforschung via da|ra-Schnittstelle und *VerbundFDB-Harvester* in die VerbundFDB-Datenbank sukzessive integriert werden. Da die Datenregistrierung bei GESIS sehr technikgestützt ist, erfordert die Integration des Testkorpus neben der Anpassung der inhaltlichen Beschreibung der Studien auch einige wenige technische institutsinterne Anpassungen bei der DOI-Vergabe.[21] Mithilfe des Testkorpus ist es möglich, das Mapping, den Harvesting-Workflow und die technische Umsetzung anhand unterschiedlicher konkreter Fälle zu prüfen und ggf. nachzubessern. Neben Nachbesserungen im Workflow sind auch nach wie vor Anpassungen in der Programmierung des *VerbundFDB-Harvesters* möglich und notwendig. Ziel ist es, weitere Anwendungsfälle von VerbundFDB-Partnern und Repositorien, die bei da|ra registrieren, zu testen, um zukünftig einen reibungslosen, automatisierten Betrieb des *VerbundFDB-Harvesters* zu gewährleisten.

Geplant ist, den *VerbundFDB-Harvester* in der zweiten Jahreshälfte 2021 auf den automatischen Betrieb umzustellen. Dann soll die OAI-Schnittstelle von da|ra täglich abgefragt werden.

6. Fazit und Ausblick

Während des manuellen Betriebs des *VerbundFDB-Harvesters* wurden Datensätze auch schon erfolgreich in die VerbundFDB-Datenbank integriert und zwar nicht nur von GESIS, sondern auch von anderen VerbundFDB-Partnern (DZHW und BIBB). Insgesamt konnten auf diesem Weg bislang 70 Studien im VerbundFDB nachgewiesen werden (Stand: 11.02.2021).

[20] In der Verbundsuche werden daIra-Importe mit der folgenden Information angezeigt: „Die Informationen zu dieser Studie kommen von einem externen Anbieter, der Metadaten entsprechend seiner geltenden Standards erfasst."

[21] Die Anpassungsschritte bei GESIS umfassten im Einzelnen:

- Recherche fehlender Informationen in der Studienbeschreibung,
- Anpassung des institutseigenen Datendokumentationstools DBKEdit (z.B. Einbindung kontrollierter Vokabulare, Ergänzung fehlender Metadatenelemente für VerbundFDBFlag),
- Überarbeitung der Übermittlung der Metadaten an da|ra (d.h. Anpassung des Mappings sowie eigener Institutsstandards auf das da|ra-Metadatenschema in Hinblick auf VerbundFDB Kernset),
- die Aktualisierung der Metadaten bei da|ra.

Dem VerbundFDB ist es somit in enger Zusammenarbeit mit da|ra und seinen Partnern in der *AG Metadaten* erfolgreich gelungen, eine Schnittstelle für die Erweiterung von Datennachweisen zu entwickeln und zu etablieren. Da da|ra von vielen FDZ der Bildungsforschung für die DOI-Registrierung ihrer Forschungsdaten genutzt wird, konnte der Zugang zum Harvesting besonders einfach und auch für kleinere FDZ erreichbar gestaltet werden. Die ersten Nutzungsfälle zeigen, dass die Auszeichnung der Daten in da|ra entsprechend des *VerbundFDB-Kernsets Bildungsforschung* ohne großen Mehraufwand möglich ist.

Somit bietet das da|ra-Harvesting Forschungsdatenzentren eine unkomplizierte aber auch skalierbare Möglichkeit, Forschungsdaten deutschlandweit zentral nachzuweisen und sichtbar zu machen. Zudem verfolgt der VerbundFDB das Konzept des zweistufigen Datennachweises. Im Portal des VerbundFDB werden Forschungsdaten auf Studienebene sowie der mit einer DOI versehenen Datenkollektionsebene nachgewiesen, für die tiefergehende Erschließung – z. B. auf Instrumentenebene – und die Bereitstellung der Daten sind die einzelnen Forschungsdatenzentren verantwortlich. Die Verbindung zwischen VerbundFDB-Nachweis und FDZ erfolgt in der Regel über die DOI. Dieses Konzept stellt neben der höheren Sichtbarkeit der Daten auch eine höhere Nutzendenfrequenz für die einzelnen Forschungsdatenzentren sicher, da die Nutzenden vom VerbundFDB-Portal zu deren Webauftritt weitergeleitet werden. Das Suchportal des VerbundFDB festigt damit gleichzeitig seine Stellung als zentrales Datennachweisinstrument im deutschsprachigen Raum für die Bildungsforschung.

Eine weitere nicht zu unterschätzende Leistung war die Entwicklung des *VerbundFDB-Kernsets Bildungsforschung*. Damit wurde erstmals ein Korpus an Metadaten definiert, der für eine aussagekräftige Beschreibung von Forschungsdaten der Bildungsforschung mindestens notwendig ist. Die Zukunft wird zeigen, ob sich dieser Standard wie beabsichtigt dauerhaft etabliert. Erste Rückmeldungen legen dies nahe.

Mit Harvester und Kernset wurden zudem beispielhafte Lösungen entwickelt, die auch für andere Forschungsfelder – etwa im Rahmen der Nationalen Forschungsdateninfrastruktur (NFDI) – attraktiv sein könnten. Die in diesem Beitrag skizzierten Standards und Workflows sind disziplinunabhängig und damit adaptionsfähig.

Literatur

Harzenetter, Karoline: Bericht vom ersten Netzwerktreffen des Verbunds Forschungsdaten Bildung. Mai 2017. URL: https://www.forschungsdaten-bildung.de/get_files.php?action=get_file&file=VFDB_NT1_Bericht01_201705.pdf (zuletzt geprüft am 08.02.2021).

Institut zur Qualitätsentwicklung im Bildungswesen (Hrsg.): Metadaten-Registry des Verbund Forschungsdaten Bildung (VerbundFDB). Kataloge. URL: https://mdr.iqb.hu-berlin.de/#/catalogs (zuletzt geprüft am 08.02.2021).

Koch, Ute; Akdeniz, Esra; Meichsner, Jana; Hausstein, Brigitte; Harzenetter, Karoline (2017): da|ra Metadata Schema. Documentation for the Publication and Citation of Social and Economic Data. GESIS Papers, 2017-25. Version: 4.0. GESIS Leibniz Institute for the Social Sciences. Text. URL: https://doi.org/10.4232/10.mdsdoc.4.0 (zuletzt geprüft am 08.02.2021).

Open Archives Initiative (OAI): Open Archives Initiative. Protocol for Metadata Harvesting. URL: https://www.openarchives.org/pmh/ (zuletzt geprüft am 08.02.2021).

Verbund Forschungsdaten Bildung (2019): Kernset und da|ra-Harvesting im VerbundFDB. Version 1.0, fdbinfo Nr. 7. URL: https://www.forschungsdaten-bildung.de/files/fdbinfo_7.pdf (zuletzt geprüft am 08.02.2021).

Verbund Forschungsdaten Bildung (2019): Metadatenset des VerbundFDB. Version 1.0, fdbinfo Nr. 8. URL: https://www.forschungsdaten-bildung.de/files/fdbinfo_8_Metadatenset_v1.0.pdf (zuletzt geprüft am 08.02.2021).

Impressum

Kontakt:
Rat für Sozial- und Wirtschaftsdaten (RatSWD)
Geschäftsstelle
Am Friedrichshain 22
10407 Berlin
office@ratswd.de
https://www.ratswd.de

Die Geschäftsstelle des RatSWD wird als Teil von KonsortSWD im Rahmen der NFDI durch die Deutsche Forschungsgemeinschaft (DFG) gefördert - Projektnummer: 442494171.

Berlin, Juni 2021

doi: 10.17620/02671.62

Zitationsvorschlag:
Harzenetter, Karoline; Pegelow, Lisa und Weisbrod, Dirk (2021): Forschungsdaten sichtbar machen: Der VerbundFDB-Harvester. RatSWD Working Paper 275/2021. Berlin, Rat für Sozial- und Wirtschaftsdaten (RatSWD). https://doi.org/10.7620/02671.62.