Integration von Forschungsdaten in Open-Access-Publikations- und Suchsysteme

Birte Lindstädt[1]

1 ZB MED Informationszentrum Lebenswissenschaften, Köln

Zusammenfassung. Forschungsdaten sollen dem Wissenschaftssystem Open Access zur Verfügung gestellt werden, um Transparenz und Nachnutzbarkeit zu ermöglichen. Eine Vielzahl dezentraler, teilweise fachspezifischer Infrastrukturen zu Speicherung, Archivierung, Nachweis und Zugriff auf Forschungsdaten existieren bereits bzw. sind im Aufbau begriffen.
Es geht jedoch nicht nur um die isolierte Betrachtung von Forschungsdaten, sondern um ihre Integration in vorhandene bzw. aufzubauende Publikations- und Suchsysteme. Als zitierfähiger wissenschaftlicher Output sollten Forschungsdaten in ihre Zusammenhänge mit Textpublikationen wie Journalartikeln oder mit anderen Forschungsdatensätzen gestellt werden. Diese Zusammenhänge sind wiederum in den Publikationssystemen abzubilden. Nach der Publikation ist auch dafür zu sorgen, dass der Nachweis der Forschungsdaten in relevante Suchportale integriert wird.
Als Beispiel eines solchen integrierten Ansatzes wird das Konzept von ZB MED - Informationszentrum Lebenswissenschaften dargestellt. PUBLISSO als Open-Access-Publikationsportal und LIVIVO als lebenswissenschaftliches Suchportal bilden die Grundlagen dieses Ansatzes. In beiden künftig miteinander verzahnten Systemen werden die Spezifika lebenswissenschaftlicher Forschungsdaten berücksichtigt, sofern sie bei der Publikation und Suche eine Rolle spielen.

Schlagwörter. Open Access, Forschungsdaten, Publikationsportal, Suchportal, Lebenswissenschaften

Besonderheiten lebenswissenschaftlicher Forschungsdaten

Bevor auf das Publikationsportal eingegangen wird, werden kurz die Spezifika lebenswissenschaftlicher Forschungsdaten dargestellt. Nach Definition von ZB MED umfassen die Lebenswissenschaften die Fächer Medizin, Gesundheitswesen, Umwelt-, Ernährungs- und Agrarwissenschaften. Eine Erweiterung dieser Kerndisziplinen stellen beispielsweise die Biologie oder die Psychologie dar, für die ebenfalls Publikationsdienstleistungen erbracht werden können.

Die Forschungsdaten der relevanten Fächer unterscheiden sich inhaltlich in hohem Maße: auf der einen Seite stehen medizinische, meist personenbezogene Daten wie Blutprobenergebnisse, Röntgenbilder oder Ergebnisse von Ernährungs- oder Klinischen Studien. Auf der anderen Seite Bodenmesswerte, Emissionswerte in Tierställen oder Wetterdaten.

Von der Art der Daten finden sich jedoch durchaus Gemeinsamkeiten für den Umgang bei der Verarbeitung, Speicherung und Publikation: Bilder (z.B. MRT, Satellitenaufnahmen), Videos (z.B. Operationsfilme, Interviews), statistische Daten, sog. Big Data (z.B. genomische Sequenzen, Daten von Landmaschinen) oder geräteabhängige Daten (z.B. Röntgengerät, Emissionsmessgerät).

Eine wesentliche Besonderheit in der Medizin, aber auch teils in den Ernährungswissenschaften, sind personenbezogene Daten, die entsprechenden Datenschutzbedingungen unterliegen. Unter anderem führen diese zu folgenden Rahmenbedingungen, die im projektbezogenen Forschungsdatenmanagement berücksichtigt werden müssen:

- Schutz persönlicher Interessen (Personenbezug von Phänotypdaten, *omics-Daten, Biomaterial) / Pflicht zur Anonymisierung,
- gesetzliche Aufbewahrungsfristen (minimal und maximal),
- rechtlicher Rahmen: Geflecht aus MBO-Ä (ärztliche Schweigepflicht), Bundes-/ Landesdatenschutzgesetz, Gesetz zur wirtschaftlichen Sicherung der Krankenhäuser und zur Regelung der Krankenhauspflegesätze,
- ethische Aspekte bzgl. Erhebung und Nutzung,
- komplexes Regelwerk bzgl. klinischer Studien („Gute klinische Praxis“),
- Schutz kommerzieller Interessen (Innovationsschutz),
- proprietäre Formate.

Daher spielen für medizinische Forschungsprojekte Datenschutzkonzepte, die Patienteninformationen, Einwilligungsverfahren, Pseudonymisierungs- und Anonymisierungsverfahren berücksichtigen, eine große Rolle. Diese sind in der Regel Voraussetzung, um überhaupt eine Datenpublikation anstreben zu können.

Aufgrund dieser Rahmenbedingungen stehen beim Teilen von Daten in der Medizin auch nicht unbedingt “offene” Daten im Sinne des Open Access im Vordergrund, sondern es spielen unterschiedliche Zugangsweisen zu Forschungsdaten eine Rolle:

- Teilzugang nach Anfrage,
- Modell „Transferstelle“ (z.B. Transferstelle für Daten und Biomaterialienmanagement, Universitätsmedizin Greifswald),
- Zugang zu faktisch anonymisierten Daten (DeStatis, Statistisches Bundesamt: Gesundheit)
- Zugang “on Screen”,
- Zugang „remote“: „Anfrage einschicken“ (z.B. Informationssystem Versorgungsdaten (Datentransparenz) Daten, DaTraV / DIMDI),
- Zugang „on Site“: Auswertung vor Ort (z.B. DaTraV Forscherplatz).

Dies gilt selbstverständlich auch für sensible Daten anderer lebenswissenschaftlicher Disziplinen wie beispielsweise ökologische Daten für bedrohte Spezies.

In der Regel müssen die genannten Besonderheiten bereits vor einer Datenpublikation berücksichtigt und entsprechende Anonymisierungsverfahren etc. durchgeführt worden sein. Sie spielen für das projektbezogene Forschungsdatenmanagement folglich eine große Rolle. Bei der eigentlichen Publikation von und der Recherche nach lebenswissenschaftlichen Forschungsdaten kommen u.a. folgende fachbezogene Aspekte zum Tragen:

- Berücksichtigung von Fach-Thesauri bei der Verschlagwortung bzw. bei der Suche (Medical Subject Headings MeSH, Multilingual Agricultural Thesauraus AGROVOC, Umweltthesaurus UMTHES),
- fachbezogene Metadaten, z.B. geographischer Ort, Datenerhebungsform
- Auswahl relevanter Fachgruppen aus fachlichen Klassifikationen, z.B. Dewey Decimal Classification DDC,
- Auswahl einer geeigneten “offenen” Lizenz für die Publikation
- Vergabe eines DOI.

Das ZB MED Publikationsportal PUBLISSO

PUBLISSO umfasst verschiedene Publikationsplattformen: Zum einen Plattformen für die html-basierte Open-Access-Erstpublikationen von Journalartikeln, Kongressbeiträgen und Büchern bzw. Buchkapiteln (Publikationsplattform Lebenswissenschaften „gold“) und zum anderen das Fachrepositorium Lebenswissenschaften für Zweitveröffentlichungen als PDF und die Publikation unterschiedlicher Dateien, z.B. Forschungsdaten. Es besteht jeweils die Möglichkeit, zu Publikationen gehörende Forschungsdaten parallel zu veröffentlichen und auf diese zu verweisen, wobei die Datenpublikation im Fachrepositorium Lebenswissenschaften oder in anderen Repositorien erfolgen kann (z.B. durch die Kooperation von ZB MED mit Dryad, einem englischsprachigen Forschungsdaten-Repositorium aus den USA). Die Forschungsdaten stellen eine eigenständige Publikation dar und sind mit den zugehörigen Publikationen verknüpft. Durch die gegenseitige Verlinkung ist es beispielsweise möglich, Forschungsdaten, die zu einem Journalartikel gehören, beim Lesen des Volltexts direkt aufzurufen. Der Volltext kann zudem auch über die publizierten Forschungsdaten gefunden werden.

Die folgende Graphik gibt einen Überblick über die Struktur des Publikationsportals (vgl. Abb. 1).

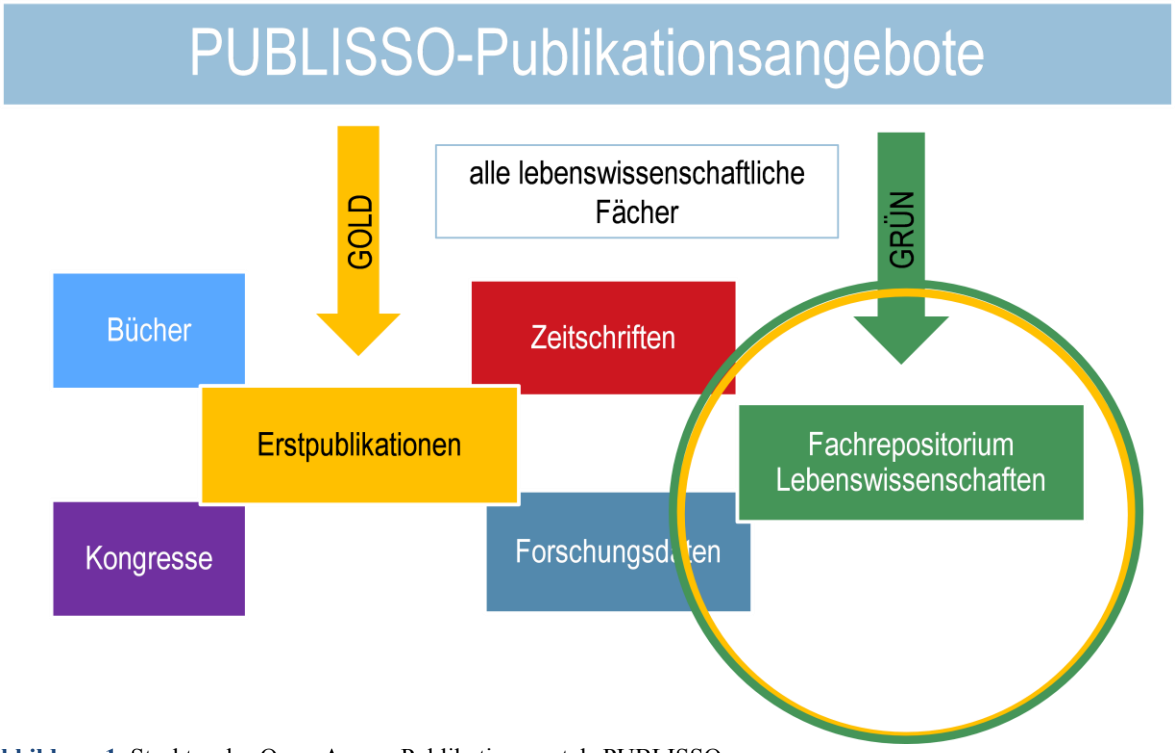

Abbildung 1. Struktur des Open-Access Publikationsportals PUBLISSO

Die Kernangebote der Publikationsplattform sind derzeit die Produkte German Medical Science (GMS -medizinische Open-Access-Zeitschriften, -Kongressabstracts und -Forschungsberichte „gold“) und Living Handbooks (Monografien „gold“). Es ist geplant, die beiden Angebote in den nächsten Jahren zusammenzuführen und auf alle Fächer der Lebenswissenschaften auszuweiten.

Die „Living Handbooks“ erlauben es, Forschungsergebnisse kapitelweise und zeitnah zu publizieren und regelmäßig zu aktualisieren, ohne von einem langwierigen Print-Publikationsprozess und der gleichzeitigen Fertigstellung aller eingereichten Kapitel abhängig zu sein. Dadurch sind

die Bücher weniger statisch, sondern „leben“. Ihre Attraktivität erhalten sie auch durch die Möglichkeit der Einbindung von multimedialen Inhalten, Forschungsdaten etc. Dafür wurde das Content-Management-System Drupal (Open Source) an die Anforderungen des wissenschaftlichen Publizierens angepasst. Mit „Living Textbooks of Hand Surgery“ wurde bereits der Prototyp eines Open-Access-Handbuchs (Living Handbooks) entwickelt.

Künftig wird die Publikationsplattform „gold“, so ausgebaut, dass darauf Zeitschriften, Kongresse und Bücher sowie - mittelfristig --auch zugehörige Forschungsdaten aus einer Hand publiziert werden können, um jedem der Fächer im Zuständigkeitsbereich von ZB MED die gleichen Publikationsmöglichkeiten zu bieten. Die Sichtbarkeit der Publikationen wird zusätzlich gestärkt und Querverweise, auch interdisziplinär, ermöglicht.

Fachrepositorium Lebenswissenschaften

Das Fachrepositorium Lebenswissenschaften wird gemeinsam mit dem technischen Kooperationspartner Hochschulbibliothekszentrum Nordrhein-Westfalen (hbz) aus- und aufgebaut und basiert auf der technischen Grundlage Fedora bzw. Drupal für die Ansicht. Neu entwickelte Erfassungsmasken erlauben es unter anderem, neben Monographien auch weitere Publikationsformate wie selbstständige und unselbstständige Literatur aufzunehmen und zweitzuveröffentlichen. Auch Video-, Bild- und Audiodateien können aufgenommen und über einen integrierten Viewer direkt abgespielt oder dargestellt werden. Die technische Grundlage erlaubt es zudem, weitere Forschungsdaten (singulär sowie in Verbindung mit einer Publikation) zu publizieren. Sofern noch nicht vorhanden, erhalten alle aufgenommenen Publikationen einen persistenten Identifikator (DOI-Digital Object Identifier).

Ziele und Strategien für die Publikation von Forschungsdaten im Fachrepositorium Lebenswissenschaften

Die Publikationsmöglichkeiten für Forschungsdaten im Rahmen von PUBLISSO bauen auf dem strategischen Ziel auf, bereits vorhandene Infrastrukturen zur Datenpublikation in den Lebenswissenschaften aufzuzeigen und an den Stellen eigene Angebote aufzubauen, wo Lücken identifiziert werden. Dies bezieht sich beispielsweise auf den sog. long tail der Forschungsdaten, also Daten, die ein geringes Datenvolumen aufweisen, in verschiedenen Datenformaten vorliegen und somit nur schwer standardisierbar sind, aber auch auf lebenswissenschaftliche Teildisziplinen, in denen Möglichkeiten zur Datenarchivierung und -publikation weitgehend fehlen.

Metadatenschema für die Publikation von Forschungsdaten

Als Orientierungshilfe für die Entwicklung eines Metadatenschema zur Erfassung von Forschungsdaten wurden zunächst Metadatenschemata existierender (Daten-) Repositorien analysiert und auf eine Übertragbarkeit für die vorliegende Aufgabenstellung geprüft.

Eine der wichtigsten Quellen stellte das DataCite-Metadatenschema dar, das aktuell in der Version 4.0 vom September 2016 vorliegt. DataCite ist eine internationale Organisation zur Vergabe von DOIs für Forschungsdaten, bei der ZB MED Mitglied ist. Das bei DataCite verwen-

dete Metadatenschema hat keinen fachlichen Fokus, da die Referenzierung verschiedenster Objekte aus allen Disziplinen angestrebt wird. Es versucht jedoch auch fachspezifische Aspekte einzubinden, z.B. durch das Feld GeoLocation. Außerdem setzt es die Hürde zur Registrierung und damit Publikation von Forschungsdaten durch lediglich sechs verpflichtende Metadaten recht niedrig an und unterscheidet darüber hinaus in „empfohlene“ und „optionale“ Felder.

Um den fachspezifischen Anforderungen lebenswissenschaftlicher Forschungsdaten gerecht zu werden, wurde in der Entwicklung des Metadatenschemas für das Fachrepositorium eine Reihe fachspezifischer Metadaten und Standards berücksichtigt wie z.B. die fachliche Zuordnung, geographische Angaben oder der Thesaurus AGROVOC für die Verschlagwortung. Das Ziel der Nachnutzbarkeit wird durch Kriterien wie „Beschreibung“ oder „Hinweise zur Nutzung“ als verpflichtende Metadatenangaben erreicht, unabhängig davon, ob eine Textpublikation existiert, die ebenfalls Beschreibungen liefert.

Bei einer Forschungsdatenpublikation im Fachrepositorium Lebenswissenschaften kann darüber hinaus ein komplexes Beziehungsnetzwerk abgebildet werden: In den Metadaten ist verzeichnet, ob die Forschungsdaten zu einer Textpublikation gehören, ob sie Teil eines größeren Datensatzes sind oder ob es mehrere Versionen gibt. Bei Forschungsdatensätzen, die aus mehreren Teilen bzw. Dateien oder zugehörigen Dateien wie Beschreibungen bestehen, ist es möglich, alle Bestandteile unter einen Metadateneintrag zu stellen. Alle Verknüpfungen werden möglich, da die Forschungsdaten mit eigenem DOI im Fachrepositorium Lebenswissenschaften abgelegt und auch die zugehörigen Publikationen nach Möglichkeit mit einem persistenten Identifier in die Metadaten aufgenommen werden. Die DOI-Registrierung erfolgt über DataCite.

Auf der Grundlage der genannten Aspekte und insbesondere auf dem DataCite-Schema, aber auch auf anderen Beispielen (wie u.a. PANGAEA Data Publisher for Earth & Environmental Science) sowie den Anforderungen der Forschungsdaten in den lebenswissenschaftlichen Disziplinen fußt das für das Fachrepositorium Lebenswissenschaften entwickelte Metadatenschema:

Tabelle 1. Metadatenschema für Forschungsdaten im Fachrepositorium Lebenswissenschaften (Pflichtfelder).

Metadatum (übergeordneter Begriff)	Feldname	Feldname (untergeordnet)
Titel	Titel	
Urheberschaft	Autor*in (Linked Data: Gemeinsame Normdatei GND)	Nachname
		Vorname
		ORCID (optional)
		Affiliation (optional)
	Körperschaft (wenn kein Autor vorhanden)	
Dateiupload	Hochzuladende Datei	
	Format (xls, jpeg, etc.)	
	Medientyp (Bild, Video, Software, etc.)	
	Größe	
	Zugriffsrechte (open access, Embargo)	Embargofristende
	Copyrightjahr	
	Lizenz	Empfehlung: Open Data Commons Open Database License (ODbL)
	DOI	Neu
		Vorhanden
	Zuletzt hochgeladen	

Erschließung	Abstract	Sprache
	Fachgruppenzuordnung (Medizin, Umwelt, Agrar, Ernährung, interdisziplinär)	
	DDC-Klassifikation (Auswahl lebenswissenschaftlicher Fachgruppen)	
	Sprache (dt.,engl., frz., span., ital.)	

Tabelle 2. Metadatenschema für Forschungsdaten im Fachrepositorium Lebenswissenschaften (optionale Felder).

Metadatum (übergeordneter Begriff)	Feldname	Feldname (untergeordnet)
Beteiligte	Beteiligte Personen (Linked Data: Gemeinsame Normdatei GND)	ORCID
		Affiliation
	Förderer	Förder-ID
Erfassung	Schlagworte (Linked Data: AGROVOC)	Sprache
	Datenerhebungsform (z.B. Probe, Interview, Genomsequenzierung)	
	Erhebungszeit	Zeitpunkt
		Zeitraum
	Erfassungsort	Koordinaten (Point)
		Koordinaten (Box)
Externe Referenzen	Verwendete Publikationen	
	Zugehörige Publikationen	
	Versionen	Vorgänger
		Nachfolger

Dieses Schema stellt die aktuelle Basis für die Publikation von lebenswissenschaftlichen Forschungsdaten dar und soll in Kooperation mit den fachlichen Communities weiterentwickelt werden, um spezifischer auf die disziplinabhängigen Bedarfe eingehen zu können.

Dazu beteiligt sich ZB MED an Forschungsprojekten, wie beispielsweise an dem “Verbundvorhaben Emissionsminderung Nutztierhaltung - Einzelmaßnahmen”, in dem das Bundeslandwirtschaftsministerium technische Maßnahmen in Tierställen zur Emissionsminderung erproben lassen möchte. Hierbei soll ZB MED die Erstellung eines Datenmanagementplans sowie die Publikation der Forschungsdaten im Fachrepositorium Lebenswissenschaften, in der Regel Messergebnisse, übernehmen.

Zur Verbesserung der Sichtbarkeit der Publikationen werden alle Inhalte des Fachrepositoriums Lebenswissenschaften in das ZB MED Suchportal LIVIVO übernommen. Eine OAI-Schnittstelle erlaubt darüber hinaus das harvesten durch andere Systeme wie beispielsweise Bielefeld Academic Search Engine BASE.

Das ZB MED Suchportal LIVIVO

LIVIVO - das ZB MED-Suchportal für Lebenswissenschaften (https://www.livivo.de) bietet umfassende und kostenfreie Recherchewerkzeuge für die interdisziplinäre Literaturversorgung in den Fächern der Lebenswissenschaften.

Die Datengrundlage von LIVIVO bilden qualitätsgeprüfte und kuratierte Datenquellen. Sie umfassen ein großes Spektrum der Literatur in den Lebenswissenschaften. Wichtige Datenquellen sind die ZB MED-Kataloge und Artikeldaten (CCMED, CCGREEN), German Medical Science,

Medline (PubMed) , AGRIS, AGRICOLA, Fachkataloge, Verlagsdaten, der Dissonline-Dienst der Deutschen Nationalbibliothek und ein fachspezifischer Auszug von BASE, der Bielefeld Academic Search Engine.

Zur effektiven Verarbeitung und Anreicherung dieser Daten wurde 2016 das ZB MED-Knowledge Environment (ZB MED-KE) eingeführt, ein universelles Instrument zum Import von Metadaten und Volltexten sowie deren Verarbeitung und Nachnutzung. Es stellt somit die umfassende Datenbasis für LIVIVO dar und dient gleichzeitig als unverzichtbarer Bestandteil der Informationsdienste und der Forschungsaktivitäten an ZB MED (s. Abb. 2).

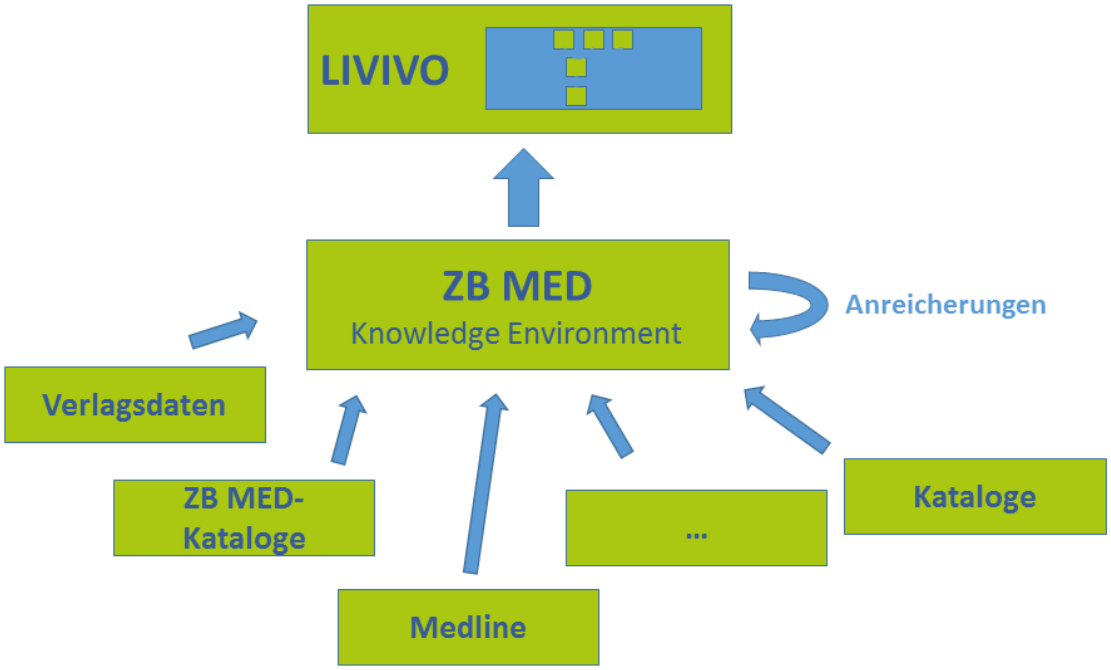

Abbildung 2. Das ZB MED-Suchportal LIVIVO

Die auf den PUBLISSO-Plattformen publizierten oder eingebundenen Forschungsdaten werden über Schnittstellen bzw. Konverter in das ZB MED-KE eingespeist und im Suchportal LIVIVO nachgewiesen. Sie sind somit breit auffindbar. LIVIVO stellt aktuell Such- bzw. Filtermöglichkeiten für mit Textpublikationen verknüpfte Forschungsdaten zur Verfügung.

Zurzeit sind bereits die Datensätze des Repositoriums DRYAD in Verknüpfung mit den korrespondierenden Volltexten eingebunden. Künftig sollen auch singuläre Forschungsdaten suchbar werden.

Das Angebot recherchierbarer Forschungsdaten soll deutlich erhöht werden, indem die Daten weiterer DataCite-Datenzentren sowie anderer qualitätsgeprüfter Forschungsdatenrepositorien (z.B. auf der Grundlage des Meta-Portal für Forschungsdatenrepositorien re3data.org) eingebunden werden. Der über DataCite vergebene DOI und die somit vorhandenen Metadaten sind ein wichtiges Kriterium, Forschungsdaten in LIVIVO nachweisen zu können. Daher stellt das Vorhandensein eines persistenten Identifikators ein wichtiges Qualitätskriterium dar.

Durch die in LIVIVO integrierte semantische Erschließung ist außerdem eine Kontextualisierung der Forschungsdaten möglich. “Recherchen werden durch linguistische Verfahren aufbereitet und semantisch mit sprachunabhängigen Konzepten annotiert. Als Fachthesauri werden für die Fächer Medizin und Gesundheit die Medical Subject Headings, für die Ernährungs-, Umwelt- und

Agrarwissenschaften AGROVOC und UMTHES verwendet. Durch das Abbilden der Fachbegriffe in unterschiedlichen Sprachen auf ihre linguistischen Repräsentationen können sprachübergreifend Suchergebnisse gefunden werden. Gleichzeitig wird die Suche nach Wortvarianten und Synonymen ermöglicht."

Abschlussbemerkung

Deutlich wird, dass die Verfügbarkeit und Publikation von lebenswissenschaftlichen Forschungsdaten in Informationsinfrastrukturen eingebettet sein muss, die sowohl die bibliothekarischen als auch die fachspezifischen Anforderungen an die Daten berücksichtigen. Insofern gilt es künftig die Zusammenarbeit von ZB MED und der lebenswissenschaftlichen Forschung noch enger zu verzahnen und die Publikations- und Suchportale gemeinsam mit den Forschenden weiterzuentwickeln.

Literaturangaben

Arning, Ursula, Birte Lindstädt, und Jasmin Schmitz. 2016. „PUBLISSO: „Das Open-Access-Publikationsportal für die Lebenswissenschaften", *GMS Medizin - Bibliothek - Information* 16 (3). doi: 10.3205/mbi000370.

Bielefeld Academic Search Engine. Online verfügbar unter https://www.base-search.net. zuletzt geprüft am 06.03.2017.

Data Cite: „Meta Data Scheme 4.0". Online verfügbar unter https://schema.datacite.org/meta/kernel-4.0/doc/DataCite-MetadataKernel_v4.0.pdf. zuletzt geprüft am 03.03.2017.

Deutsche Nationalbibliothek: „Katalog der DNB". Online verfügbar unter http://search.dissonline.de/. zuletzt geprüft am 06.03.2017.

Dryad Data Repository. Online verfügbar unter http://datadryad.org/. zuletzt geprüft am 06.03.2017.

Food and Agriculture Organisation of the United Nation: „agris". Online verfügbar unter; http://agris.fao.org/. zuletzt geprüft am 06.03.2017.

German Medical Science. Online verfügbar unter http://www.egms.de/. zuletzt geprüft am 06.03.2017.

German Medical Science: „GMS Books". Online verfügbar unter http://www.gms-books.de. zuletzt geprüft am 06.03.2017.

National Agricultural Library: „NAL Catalogue - agricola“. Online verfügbar unter http://agricola.nal.usda.gov. zuletzt geprüft am 06.03.2017.

NCBI: „PubMed.gov“. Online verfügbar unter https://www.ncbi.nlm.nih.gov/pubmed. zuletzt geprüft am 06.03.2017.

Open Data Commons: “Legal Tools for Open Data”. Online verfügbar unter 06.03.2017, http://opendatacommons.org/licenses/odbl/. zuletzt geprüft am 06.03.2017.

Christioph Poley. 2016. ”LIVIVO - “Neue Herausforderungen an das ZB MED-Suchportal für Lebenswissenschaften”. *GMS Medizin - Bibliothek - Information* 16(3), doi: 10.3205/mbi000376