# Der Aufbau einer „Entity Collection“ der Forschungsleistung der TU Dortmund

**Hans-Georg Becker[1], Kathrin Höhner[2]**

1,2 Universitätsbibliothek der Technischen Universität Dortmund

**Zusammenfassung.** Mit einer systematischen Kuratierung der lokalen Daten will die Universitätsbibliothek das Ziel erreichen, die Forschungsleistung der Technischen Universität Dortmund möglichst vollständig sichtbar und nachnutzbar zu machen. Um Data Curation effizient zu gestalten, verfolgt die Universitätsbibliothek Dortmund ein umfassendes Gesamtkonzept. So werden ein Metadaten-Managementsystem (MMS) und eine Datenplattform entwickelt, die auf die Nachnutzung möglichst vieler bereits existierender und verlinkbarer Identifikatoren zielen. Zudem bietet das Metadaten-Managementsystem die Möglichkeit, unterschiedlichste Datentypen mit verschiedenen Relationen zu erfassen. Es ist daher für das Monitoring der an der Technischen Universität Dortmund produzierten und nachzuweisenden Forschungsdaten besonders geeignet. Auch wird es eine Synchronisation mit dem Katalog plus der Universitätsbibliothek geben.
Für die Wissenschaftlerinnen und Wissenschaftler wird es für die unterschiedlichen Anwendungsszenarien eine einzige Oberfläche als Einstieg geben. Hierüber können sie nicht nur Publikationen für die Hochschulbibliographie melden, Volltexte in das Repositorium hochladen oder Publikationslisten für Webseiten generieren, sondern auch finanzielle Unterstützung für Open Access-Publikationen beantragen.

**Schlagwörter**. Datenkuratierung, Metadatenmanagement

## Aufbau einer “Entity collection“

Im Laufe des gesamten Forschungsprozesses werden Daten generiert. Deren Analyse und Interpretation erfolgt einerseits im Rahmen von klassischen Publikationen in wissenschaftlichen Zeitschriften oder Monographien, andererseits im Rahmen von (Zwischen)-Berichten, Software, Kunstwerken oder anderem. All dies stellt ebenso wie die erhaltenen Primärdaten den Forschungsoutput einer Universität dar.

Die zugehörigen Metadaten in strukturierter Form abzulegen, zu präsentieren und durchsuchbar zu machen, ist eine Herausforderung für Bibliotheken, die sie mit ihrer Kompetenz für die Strukturierung von Daten bereits jetzt annehmen. Klassischerweise wird das Publikationsaufkommen einer Universität in Fachzeitschriften und Monographien ermittelt und z. B. in Form einer Hochschulbibliographie oder als Publikationslisten für einzelne Wissenschaftlerinnen und Wissenschaftler präsentiert. Jahn und Horstmann dagegen haben bereits 2010 eine „disziplinsensitive Strukturierung der bibliographischen Information“ gefordert. So ist es auch das Anliegen der Universitätsbibliothek (UB) der TU Dortmund, die gesamten Forschungsergebnisse ihrer Wissenschaftlerinnen und Wissenschaftler bibliothekarisch aufzubereiten, sie zu präsentieren und dabei auch Hierarchien und Vernetzungen darstellen zu können. Ein Forschungsinformationssystem ist

an der TU Dortmund noch nicht vorhanden, weswegen Metadaten zu Projekten etc. nur unzureichend bekannt sind.

So entstand die Idee, eine Entity Collection aufzubauen, in der alle zur Darstellung des Forschungsoutputs der TU Dortmund benötigten Metadaten erfasst sind.

An der UB wurden bereits 2015 folgende Prinzipien des Metadatenmanagements aufgestellt, die als Leitlinien für den Aufbau der Entitäten-Kollektion dienen.

- Jeder Datensatz hat ein Mastersystem und ein Masterformat.
- Datensätze werden automatisch verteilt und nicht mehrfach erfasst.
- Linked Data ist ein Kernkonzept.
- Es werden so wenig Daten wie möglich und so viele wie nötig erfasst.
- Es gibt ein Repositorium für digitale Objekte.

Basierend auf diesen Prinzipien wurde im Laufe der letzten beiden Jahre zunächst gemeinsam mit der UB der Ruhr-Universität Bochum, später nur in Dortmund, ein Metadatenmanagementsystem (MMS) entwickelt, mit dessen Hilfe die Entitäten-Kollektion aufgebaut, angepasst aber auch deren Daten weiterverwendet werden können. Als erster Schritt werden die beiden bisher vollkommen voneinander unabhängigen Systeme Hochschulbibliographie und Repositorium für Volltexte und andere digitale Objekte durch das MMS verknüpft.

## Datenmodell und Datenstruktur

Bei der Konzeption des Datenmodells mussten zwei unterschiedliche Ansätze berücksichtigt werden: einerseits müssen Metadaten zu Publikationen unterschiedlichster Art erfasst werden können, um verschiedene Zielsysteme von Webseiten bis zu (Fach)Repositorien bedienen zu können, andererseits waren die Daten der Personen und Organisationseinheiten einschließlich temporärer Projekte der Universitäten zu berücksichtigen, um Relationen darstellen zu können.

Dieses Datenmodell ist zugleich so flexibel, dass in Zukunft weitere benötigte Kategorien (z. B. Projektangaben) eingebaut werden können. Der u.a. vom Wissenschaftsrat und von der DFG für Forschungsinformationssysteme (FIS) empfohlene Metadatenstandard Cerif findet dabei Berücksichtigung.

So basieren die Entitätstypen auf denen des Cerif-Modells und auch die Relationen unter ihnen orientieren sich an diesem Modell. Für die optimale Darstellung des Forschungsoutputs ist es notwendig, Organisationseinheiten, Personen, Arbeitsgruppen und Projekte sowie unterschiedliche Publikationsformen abzubilden. Letztere unterscheiden sich im Cerif-Modell in Publikationen, Patente und Produkte. Dabei versteht man unter Produkten insbesondere nicht-textuelle Objekte.

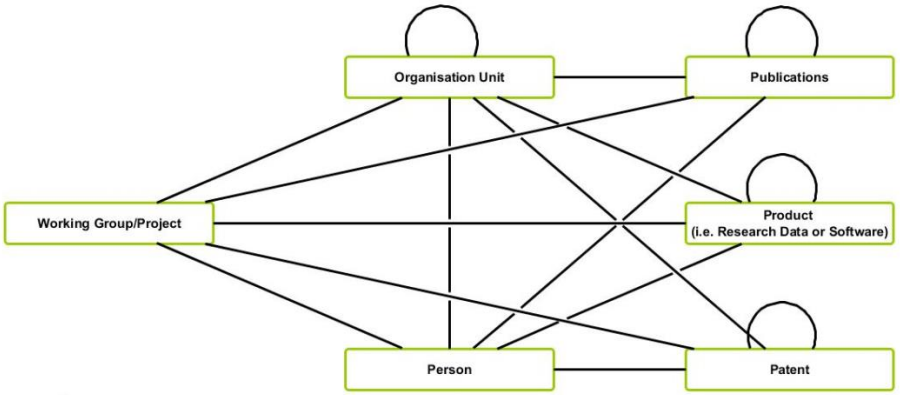




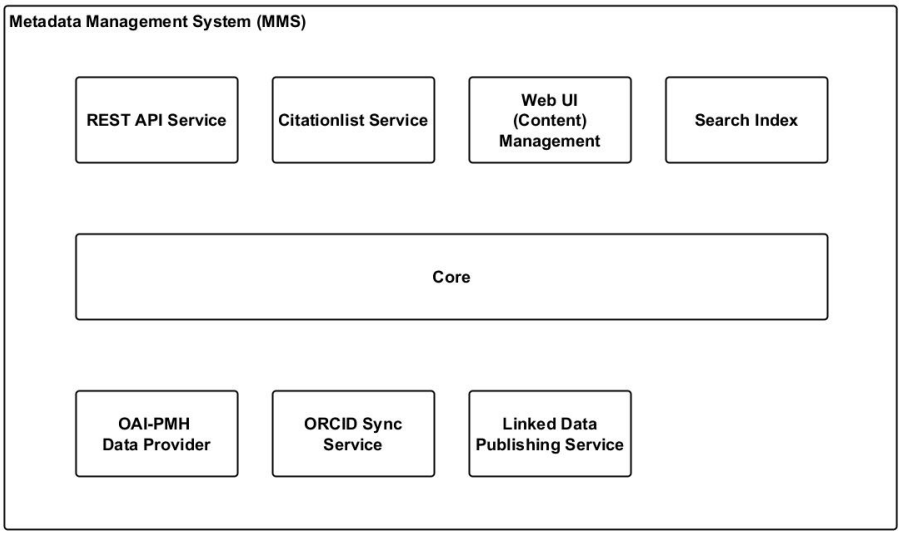

**Abbildung 1.1.** An CERIF orientiertes Datenmodell des MMS (Becker 2017)

Im MMS werden die drei Entitäten Publikationen, Patente und Produkte unter „Works“ zusammengefasst und mittels Typisierung durch kontrolliertes Vokabular unterscheidbar und Cerif-kompatibel gemacht.

Alle Entitäten bekommen im Datenmodell der UB Dortmund eindeutige Identifikatoren (IDs), wozu soweit wie möglich bereits vorhandene IDs nachgenutzt und möglichst wenige selbst generiert werden sollen. Für Personen werden z. B. die Open Researcher and Contributor ID (ORCID iD), für Personen, Organisationen und deren Untereinheiten die Gemeinsame Normdatei ID (GND ID) sowie DOIs oder andere persistente Identifikatoren als IDs für Werke unterschiedlicher Art verwendet. Insbesondere die Verwendung von Normdaten und kontrollierten Vokabularen erleichtert die Umsetzung des linked data-Prinzips.

## Software-Architektur

Das MMS besteht aktuell aus sechs Services, die auf eine mit einem application programming interface (API) versehenen Persistenzschicht aufsetzen.






**Abiildung 1.2** Architektur des Metadatenmanagementsystems (Becker 2017)

Der Service „Web UI (Content) Management“ liefert die Werkzeuge, um die unterschiedlichen Entitäten durch Metadaten zu beschreiben und mit anderen Entitäten zu verknüpfen. Überall dort, wo Verknüpfungen erzeugt werden, gibt es Assistenten, die aus den bereits im System befindlichen Entitäten Vorschläge unterbreiten. Eine Auswahl erzeugt dann eine Verknüpfung der IDs der Entitäten.

Auf Basis der im System erfassten Metadaten wird im Hintergrund ein Suchindex (Search Index) erzeugt. Dabei werden die durchsuchbaren Felder befüllt sowie der entstandene Datensatz als JavaScript Object Notation (JSON) abgelegt. Ferner werden hier die „einfache Suche“ sowie die Vorschlagsassistenten mit Daten befüllt. Auf den Suchindex kann direkt lesend zugegriffen werden; so lassen sich unterschiedliche Arten von Präsentationsanwendungen entwickeln. Auch die Webanwendung der Hochschulbibliographie der TU Dortmund verwendet diesen Index. Zusätzlich werden die Ergebnisse innerhalb der Hochschulbibliographie mittels schema.org im JSON-LD-Format (JSON for Linked Data) angereichert, was der Erhöhung der Sichtbarkeit dient.

Das MMS verfügt über ein Representational State Transfer application programming interface (REST API), mit dem die unterschiedlichen Entitätstypen angelegt, aktualisiert, gelöscht und ausgelesen werden können. Weitere Funktionen sind geplant, die es ermöglichen sollen, dass „Web UI (Content) Management“ nur noch über die REST-Schnittstelle auf die Daten zugreifen kann. Außerdem können damit ohne direkten Zugriff auf den Suchindex „Retrieval“-Anwendungen implementiert werden Die REST-Schnittstelle ist auch der Service, der die Integration in weitere Systemarchitekturen ermöglichen soll.

Der Publikationslisten-Service (Citationlist Service) liefert Zitationslisten von Personen, Organisationseinheiten oder Arbeitsgruppen und Projekten in frei wählbaren Zitationsstilen und Konfigurationen z. B. für die Verwendung in Webseiten.

Für die Nachnutzung der Daten wird neben dem Index auch ein Open Archives-Initiative Protocol for Metadata Harvesting (OAI-PMH) Data Provider zur Verfügung gestellt. Über diesen Service werden z. B. die für OpenAIRE notwendigen Daten zum Harvesting bereitgestellt, so dass

die Publikationen sichtbarer werden, die unter Verantwortung von Angehörigen der TU Dortmund entstanden sind.

Auch die Synchronisation der Daten mit der ORCID-Plattform erhöht die Sichtbarkeit der Daten. Die Synchronisation läuft hierbei in beide Richtungen: vorausgesetzt, eine Wissenschaftlerin / ein Wissenschaftler hat den Zugriff erlaubt, werden auf der einen Seite Daten aus der ORCID-Plattform in die Hochschulbibliographie übernommen, auf der anderen Seite neue oder angereicherte Daten aus dem MMS in die ORCID-Plattform übertragen.. Dabei werden verschiedene Mechanismen zur Vermeidung von Duplikaten angewendet.

Der Linked Data Publishing Service wird die Daten aus dem MMS als an die Cerif-Ontologie angelehnte Resource Description Framework (RDF)-Daten in die Linked Open Data Plattform der UB Dortmund publizieren.

Implementiert ist das System in der Programmiersprache Python unter Verwendung des Webframeworks Flask. Für den Service „Web UI (Content) Management“ wird zusätzlich das WTForms-Framework für die Definition von Webformularen verwendet. Dieses erlaubt es, objektorientiert passende Formulare generieren zu lassen, sowie deren Inhalte ohne weiteren Aufwand als JSON-Daten nachzunutzen.

Für die Persistenz wird ein Apache Solr-Index verwendet, in dem neben den suchbaren Feldern auch der vollständige Datensatz als JSON enthalten ist.

## Zielsysteme

Das im ersten Abschnitt beschriebene MMS dient als zentrale „Datendrehscheibe“ für unterschiedliche Zielsysteme, wie in Abb. 2.1 gezeigt wird. Gerade vor dem Hintergrund, dass eine Hochschulbibliographie oder ein universitäres Repositorium von Wissenschaftlerinnen und Wissenschaftlern wenig als primäre Rechercheinstrumente verwendet werden, wird die Sichtbarkeit der Metadaten insbesondere durch Implementierung verschiedener Schnittstellen zur möglichst breiten Bedienung unterschiedlicher Zielsysteme erhöht.

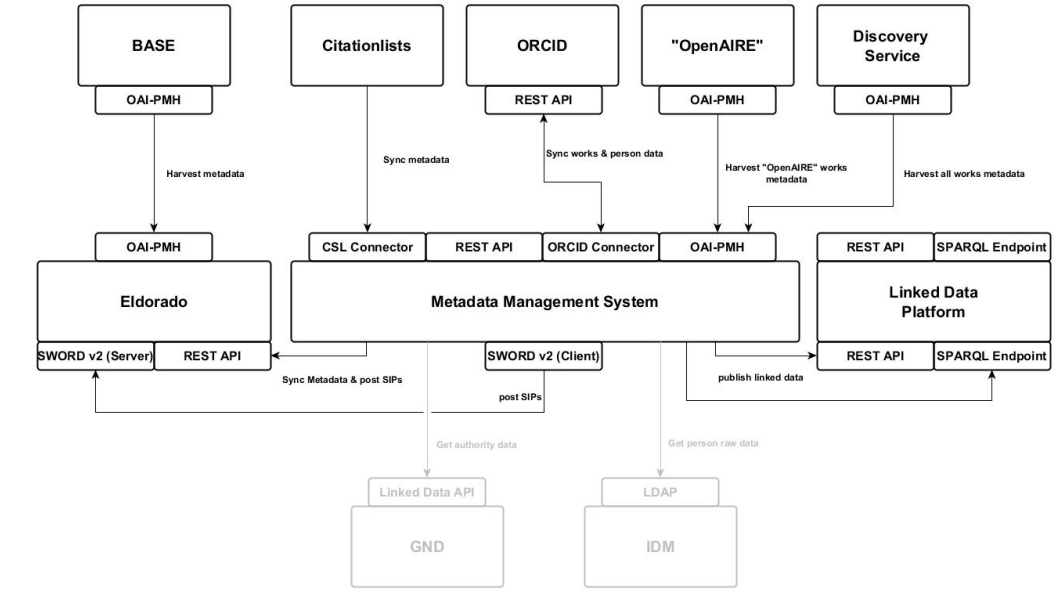





**Abbildung 2.1.** Datenfluss zwischen den Systemen an der UB Dortmund (Becker 2017)

Quellen für das MMS sind ORCID, die GND und für Personendaten das Identity-Managementsystem der TU Dortmund. Für die initiale Verknüpfung des TU-eigenen Repositoriums „Eldorado" mit dem MMS dient dieses ebenfalls als Quelle.

„Eldorado" und die ORCID-Plattform sind gleichzeitig auch Zielsysteme des MMS. Weitere Zielsysteme sind die Webseiten von Angehörigen der TU Dortmund, der Katalog plus der TU Dortmund (in der Grafik als Discovery Service bezeichnet) sowie OpenAIRE als Portal für die Forschungsergebnisse EU-geförderter Projekte.

Die meisten Zielsysteme bedienen die etablierten Schnittstellen REST, OAI-PMH oder SWORD (Simple Web-service Offering Repository Deposit). Für ORCID dagegen musste ein „Connector" geschrieben werden. Da das MMS unterschiedliche Schnittstellen bedient, ist ein Szenario der Datenlieferung an Fach-Repositorien ebenfalls umsetzbar.

Die verschiedenen Zielsysteme werden im Folgenden detaillierter beschrieben. Dabei werden BASE und OpenAIRE nicht weiter berücksichtigt, da die Datenlieferung an diese beiden Plattformen ausschließlich der Erhöhung der Sichtbarkeit der Forschungsergebnisse der TU Dortmund dient und keine TU-Dortmund- spezifischen Anwendungen darstellen.

## Publikationslisten

Die Darstellung der Forschungsleistung der Angehörigen der TU Dortmund soll auf unterschiedlichen Wegen erfolgen. Ein wichtiges Ziel ist die Generierung strukturierter Publikationslisten („Citation lists") für die Webseiten der einzelnen Forscherinnen und Forscher, von Lehrstühlen, Instituten oder auch ganzen Fakultäten. Die individuelle Strukturierung der Listen ermöglicht eine „hochgradig kondensierte Darstellungsweise wissenschaftlicher Erkenntnis" und trägt damit zur Optimierung der Visualisierung der eigenen Forschungsergebnisse bei. Denn nur wenn Forschende einen Mehrwert in der Meldung der Publikationen an die Universitätsbibliothek erkennen,

werden sie bereit sein, ihre Verlags-Publikationen ebenso wie ihre Forschungsarbeiten und grauen Publikationen der UB zu übermitteln. Für einzelne Projekte oder Arbeitsgruppen ist es ebenso notwendig, spezifische Publikationslisten generieren zu können - sei es für die Präsentation auf der Webseite oder für Förderanträge und Berichte zu Förderanträgen, wie z. B. für Sonderforschungsbereiche. Aus diesem Grunde erfasst die UB im Rahmen ihres MMS auch Arbeitsgruppen und Projekte als eigene Entitäten.

## Hochschulbibliographie

Die Hochschulbibliographie dient als zentrales Nachweisinstrument für die an der TU Dortmund entstandenen Publikationen. Darüber hinaus werden hier auch andere Forschungsleistungen, wie z. B. Patente, Vorträge oder auch Projektberichte erfasst. Die Datenlieferung an die Hochschulbibliographie beruht auf freiwilliger Basis. Um von Wissenschaftlerinnen und Wissenschaftlern akzeptiert zu werden, müssen die Inhalte attraktiv dargestellt und ein niedrigschwelliger Einstieg der Publikationsmeldung ermöglicht werden, der sich an unterschiedlichen Bedürfnissen der Forschenden orientiert. Zum einen können Forschende ihre Publikationen selbst manuell eintragen, zum anderen wird es möglich sein, Listen aus Literaturverwaltungsprogrammen im BibTex- oder RIS-Format zu importieren oder auch ganz klassisch Word- oder Excel-Listen an die UB melden. Letztere werden von den Mitarbeitern der UB manuell in die Hochschulbibliographie eingetragen.

Gerade für Wissenschaftlerinnen und Wissenschaftler der Geistes-, Kunst- und Kulturwissenschaften, die in den klassischen Abstract- und Zitationsdatenbanken häufig unterrepräsentiert sind, bieten die Hochschulbibliographie oder auch die unter 2.1 beschriebenen Publikationslisten auf ihren Webseiten eine gute Möglichkeit, ihre Forschungsleistung zu präsentieren.

Für Forschende aller Disziplinen bietet die Hochschulbibliographie die Möglichkeit, all jene Forschungsergebnisse zu präsentieren, die nicht klassische Text-Publikationen sind, z. B Software oder auf Repositorien publizierte Forschungsdaten.

## Repositorium - Eldorado

Mit dem seit Mitte 1997 betriebenen Repositorium „Eldorado“, das auf der Software DSpace basiert, bietet die UB Dortmund eine Plattform, um Dokumente sowohl primär als auch als Zweitveröffentlichung zu publizieren. Dabei spielt „Eldorado“ nicht nur für die digitale Veröffentlichung von Dissertationen und klassischer grauer Literatur, wie z. B. Zwischen-, Jahres- oder Abschlussberichte, sondern auch für Dokumente der Universitätsverwaltung eine wichtige Rolle. Auch Periodika und Schriftenreihen werden genuin hier veröffentlicht, z. B. die Projektberichte der Fakultät für Raumplanung oder die Preprints der Fakultät für Mathematik. „Eldorado“ ist von der Deutschen Initiative für Netzwerkinformation zertifiziert worden und bietet alle notwendigen Schnittstellen, um dort abgelegte Dokumente dauerhaft zur Verfügung zu stellen und eine Indexierung in wissenschaftlich relevanten Suchmaschinen zu ermöglichen. Zudem werden als persistente Identifikatoren DOIs vergeben. Die Anbindung des Repositoriums an das MMS ermöglicht zum einen eine intuitivere Bedienung für Wissenschaftlerinnen und Wissenschaftler und dient zum anderen der Verbesserung der Sichtbarkeit der Publikationen des Repositoriums, da das MMS umfassendere Metadaten erlaubt und insbesondere Verlinkungen, z. B. mit der ORCID-iD

ermöglicht. Die Verwendung von schema.org im MMS und damit auch für die Repositoriums-Daten trägt ebenso zur Erhöhung der Sichtbarkeit und damit zur Optimierung des Repositoriums bei. In den Empfehlungen der „Repositories Early Adopters Expert Group“, werden fünf „required“ Aspekte genannt, von denen die TU Dortmund vier erfüllt; einzig die Forderung 4 nach unterschiedlichen Granularitätslevels für persiste Identifikatoren kann bisher nicht erfüllt werden. Mit der Anbindung des Repositoriums an das MMS werden zudem die drei „recommended“ Aspekte erfüllt - so liefert das MMS hinreichende Metadaten für eine gute Zitation, die Sichtbarkeit wird mittels schema.org erhöht und HTML Meta Data Tags werden bereits verwendet.

### ORCID

ORCID ist für die Bibliothek der TU Dortmund ein attraktives System, weil es über die eindeutige Verknüpfung von Personen mit ihren Publikationen vollständigere Aussagen über die Publikationstätigkeiten der Angehörigen der eigenen Universität ermöglicht. Gleichzeitig strebt die TU Dortmund an, dass ihre Wissenschaftlerinnen und Wissenschaftler die Institution „Technische Universität Dortmund“ in normierter, international anerkannter Form bezeichnen. Dazu hat die TU Dortmund bereits 2016 eine Mitgliedschaft bei ORCID abgeschlossen, um sich von ihren Wissenschaftlerinnen und Wissenschaftlern sowohl das Recht einräumen zu lassen, Daten aus dem Profil zu lesen, als auch Publikationsdaten sowie die Bezeichnung der TU Dortmund in deren Profile einzuspielen. Hieraus ergeben sich Vorteile alle Beteiligten: Daten, die aus ORCID in die Hochschulbibliographie eingespielt werden, fließen hierüber in die Publikationslisten für die Webseiten ein. Damit wird den Wissenschaftlerinnen und Wissenschaftlern das Melden ihrer Publikationen erspart, sofern sie beim Einreichen jeder einzelnen Publikation ihre ORCID ID mit angeben. Die meisten internationalen Verlage verknüpfen die ORCID ID mit dem DOI der Publikation. Wurde zuvor den Organisationen DataCite und Crossref ein „auto-update“ erlaubt, werden alle Publikationen, die einen DOI als Identifikator bekommen, automatisch in das ORCID-Profil eingespielt. Für die Bibliothek ergibt sich der Vorteil einer vollständigeren Hochschulbibliographie. Die Mitarbeiter der UB bereiten die von ORCID gelieferten Daten bibliothekarisch auf und spielen sie in das ORCID-Profil zurück.

### Katalog plus (Discovery Service)

Der Katalog plus der UB Dortmund ist das primäre Suchinstrument der Bibliothek. Neben dem Nachweis des eigenen analogen und digitalen Bestandes sollen alle Publikationen von Angehörigen der TU Dortmund hier zu finden sein - unabhängig davon, ob sie analog, digital oder gar nicht bereitgestellt werden können. Daher ist geplant, die Daten des MMS mittels OAI-PMH in den Katalog plus zu transferieren.

## Ausblick

Das Potential des MMS, unterschiedliche Datentypen mit verschiedenen Relationen zu erfassen, ist insbesondere für das Monitoring der an der TU Dortmund produzierten und nachzuweisenden

Forschungsdaten geeignet - unabhängig von deren Publikation und gegebenenfalls deren Publikationsort.

An der TU Dortmund wurde im Frühjahr 2016 unter den Forschenden der TU Dortmund abgefragt, ob und in welcher Form Unterstützung beim FDM gewünscht wird. Basierend auf den Ergebnissen dieser Bedarfsabfrage wurde eine OAIS-konforme (Open Archival Information System) Architektur entwickelt, die sich im Wesentlichen aus bereits existierenden Softwarekomponenten zusammensetzt und die die maximale Integration und Nachnutzung bereits etablierter Systeme zum Ziel hat.

Das hier vorgestellte MMS kann als Content-Management-System eingesetzt werden, wenn Forschungsdaten, die in beliebigen Arbeitsumgebungen der Forschenden verwaltet wurde, zur Archivierung in Form eines submission information package (SIP) übergeben werden. Das MMS stellt in diesem Kontext unter anderem ein Werkzeug dar, mit dessen Hilfe leichtgewichtig spezielle Formulare mit FDM-spezifischen Metadaten angeboten werden können. Dies erlaubt eine hohe Flexibilität bei der Erfassung von individuellen oder projektbezogenen Metadaten für Forschungsdaten, die archiviert werden sollen.

Gleichzeitig dient das MMS als Verknüpfung zwischen FDM und einem einzuführenden CRIS. Im Datenmodell wurde bewusst der CERIF-Standard berücksichtigt, um Daten automatisiert in ein CRIS zu übertragen.

Die Modularität des MMS erlaubt es, über die bestehenden Anwendungen an der TU Dortmund und der Ruhr-Universität Bochum hinaus, Szenarien des Metadatenmanagements zu unterstützen. Im Sinne der Openness sind Code und Dokumentation über GitHub verfügbar.

## Literaturangaben

Depping, Ralf. 2014. „Publikationsservices im Dienstleistungsportfolio von Hochschulbibliotheken. Eine (Neu-)Verortung in der wissenschaftlichen Publikationskette“, *o-bib - Das offene Bibliotheksjournal* 1: 71-91. Online verfügbar unter http://dx.doi.org/10.5282/o-bib/2014H1S71-91. Zuletzt geprüft am 13.03.2017.

Fenner et al. 2016. „A Data Citation Roadmap for Scholarly Data Repositories”, bioRxiv preprint first posted online Dec. 28, 2016. Online verfügbar unter http://dx.doi.org/10.1101/097196. Zuletzt geprüft am 08.03.2017.

Horstmann, Wolfram und Najko Jahn. 2010. „Persönliche Publikationslisten als hochschulweiter Dienst - eine Bestandsaufnahme“, *Bibliothek, Forschung & Praxis*, 34 (1): 185-193. Online verfügbar unter http://doi.org/10.1515/bfup.2010.032. Zuletzt geprüft am 13.03.2017.