Das institutionelle Forschungsdatenrepositorium FDAT der Universität Tübingen

Steve Kaminski, Universität Tübingen
Olaf Brandt, Universität Tübingen

Zusammenfassung:
Das eScience-Center[1] der Universität Tübingen bietet mit dem Forschungsdatenrepositorium FDAT[2] lokalen Forschungsprojekten und Forschenden diverse Dienstleistungen sowie die nötige technische Ausstattung für die Langzeitarchivierung und Nachnutzung ihrer Forschungsdaten an. Dabei folgt FDAT den Richtlinien eines offenen Archivinformationssystems OAIS[3] und wurde von unabhängiger Stelle zertifiziert.[4] Ziel ist es, wissenschaftliche Daten sicher aufzubewahren und der breiten Öffentlichkeit nach Möglichkeit Open Access[5] zur Verfügung zu stellen. Darüber hinaus sollen Wissenschaftlerinnen und Wissenschaftler in allen Phasen des Lebenszyklus ihrer Forschungsdaten durch die Betreiber des Repositoriums beraten und technisch unterstützt werden. Das Repositorium wird legitimiert durch die von der Universität Tübingen verabschiedeten Leitlinien zum Umgang mit Forschungsdaten[6], und neben dem eScience-Center konkret betreut durch die Universitätsbibliothek Tübingen und das ansässige Zentrum für Datenverarbeitung. Das Repositorium hat am 01.01.2017 seinen produktiven Betrieb aufgenommen und hält derzeit (15.08.2018) 6741 digitale Objekte vor.

Summary:
The eScience Center of Tübingen University offers various services and, with the FDAT research data repository, the necessary technical infrastructure for the long-term archiving and reuse of research data for local research projects and the scientists. Following the guidelines for an Open Archival Information System OAIS, FDAT aims at preserving scientific data as well as making the data available to the general public in open access. In addition, scientists are advised and technically supported by the operators of the repository during all steps of the research data life cycle. The infrastructure of the repository is supported by the guidelines for research data handling as adopted by Tübingen University and, in addition to the eScience Center, specifically supervised by Tübingen University Library and the local center for data processing.The repository commenced production on 01.01.2017 and currently (15.08.2018) holds 6741 digital objects.

Zitierfähiger Link (DOI): https://doi.org/10.5282/o-bib/2018H3S61-75
Autorenidentifikation: Kaminski, Steve: GND 1131102592,ORCID: https://orcid.org/0000-0002-0047-881; Brandt, Olaf: GND 1110210620, ORCID: https://orcid.org/0000-0002-3379-1190

1 eScience-Center, Universität Tübingen, <http://www.escience.uni-tuebingen.de/>, Stand: 10.08.2018.
2 Research Data Portal FDAT, re3data.org, <http://www.re3data.org/repository/r3d100012296>, Stand: 10.08.2018.
3 ISO 14721:2012, <http://www.iso.org/standard/57284.html>, Stand: 10.08.2018.
4 The CoreTrustSeal Borad, Implementation of theCoreTrustSeal, <https://www.coretrustseal.org/wp-content/uploads/2018/04/FDAT.pdf>, Stand: 10.08.2018.
5 open-access.net, <https://open-access.net>, Stand: 10.08.2018.
6 Leitlinien zum Forschungsdatenmanagement, Universität Tübingen,<https://www.uni-tuebingen.de/forschung/service/materialien-und-dokumente/leitlinien-zum-forschungsdatenmanagement.html>, Stand: 10.08.2018.

Schlagwörter: Forschungsdaten, Repositorien, Langzeitarchivierung, Forschungsdatenmanagement, FDAT, Open Research Data Portal

1. Das FDAT-Repositorium: Zielsetzung und Organisation

Das FDAT-Repositorium wurde ins Leben gerufen, um vor allem die Geistes- und Sozialwissenschaften in Tübingen dabei zu unterstützen ihre Forschungsdaten sicher und nachnutzbar für mindestens 10 Jahre aufzubewahren, wie es die DFG in ihrem Papier zur „Sicherung guter wissenschaftlicher Praxis“[7] fordert und im Jahre 2015 mit den Leitlinien[8] zum Umgang mit Forschungsdaten nochmals präzisiert. Gerade in den Einrichtungen der genannten Fachbereiche herrscht zumeist ein Mangel an geeigneten technischen Infrastrukturen und Kenntnissen, um diesem Anspruch gerecht zu werden. Daher muss dieses Problem, auch vor dem Hintergrund stetig wachsender Datenmengen, übergeordnet und institutionell gelöst werden.

Die erklärte Schwerpunktsetzung des Repositoriums bedeutet jedoch nicht Ausschließlichkeit; das FDAT-Repositorium steht grundsätzlich allen Fachbereichen der Universität offen, die keinen Zugang zu fachspezifischen Repositorien innerhalb und außerhalb der Universität haben.

Neben der sicheren Verwahrung der Daten ist ihre Verfügbarmachung und Verbreitung über das Internet das zweite wesentliche Ziel des Repositoriums.

Um die Sichtbarkeit der Daten in FDAT zu erhöhen wird neben der Publikation der Daten über ein eigenes Web-Portal[9] grundsätzlich ein Austausch mit anderen Archiven und Nachweissystemen angestrebt. Als Beispiel sei hier die Zusammenarbeit mit dem Datenportal[10] des Forschungsdatenzentrums für Archäologie & Altertumswissenschaften in Berlin genannt.

Um Qualitätsstandards für das Repositorium sicherzustellen und zu dokumentieren, konnte für FDAT erfolgreich ein Prüfsiegel für digitale Repositorien (Core Trust Seal[11]) erlangt werden.

Dem Repositorium stehen insgesamt vier Einrichtungen der Universität Tübingen mit jeweils unterschiedlichen personellen Kompetenzen und Ressourcen (siehe Abbildung 1) zur Verfügung.

7 DFG: Sicherung guter wissenschaftlicher Praxis. Empfehlungen der Kommission „Selbstkontrolle in der Wissenschaft“, Weinheim 2013. Online: <http://doi.org/10.1002/9783527679188.oth1>, Stand: 10.08.2018.

8 DFG: Leitlinien zum Umgang mit Forschungsdaten, 2015, <www.dfg.de/download/pdf/foerderung/antragstellung/forschungsdaten/richtlinien_forschungsdaten.pdf>, Stand: 10.08.2018.

9 Forschungsdatenportal FDAT, Universität Tübingen, <https://fdat.escience.uni-tuebingen.de/portal>, Stand: 10.08.2018.

10 Troja Projekt, Türkei (Universität Tübingen), IANUS Datenportal, <http://datenportal.ianus-fdz.de/pages/collectionView.jsp?dipId=1673203>, Stand: 10.08.2018.

11 Core Certified Repositories, <https://www.coretrustseal.org/why-certification/certified-repositories/>, Stand: 10.08.2018.

Im Zentrum steht das Informations-, Kommunikations- und Medienzentrum (IKM)[12] als eine zentrale Einrichtung der Universität, bestehend aus den beiden Geschäftsbereichen der Universitätsbibliothek und dem Zentrum für Datenverarbeitung (ZDV).

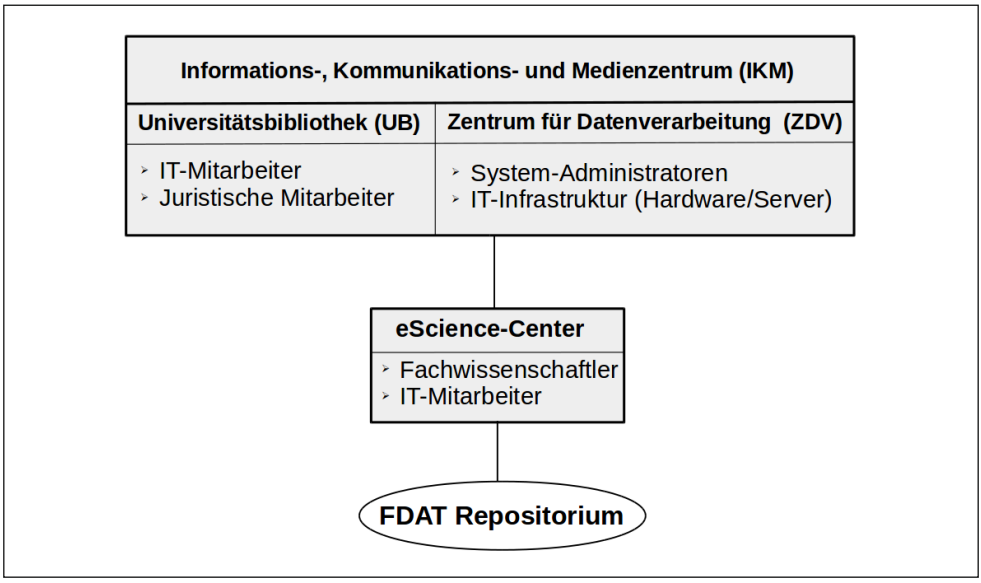

Abb. 1: Organisationsstruktur für das FDAT-Repositorium mit den beteiligten Einrichtungen der Universität Tübingen.

Das ZDV ist für Anschaffung, Betrieb und Wartung benötigter Hardware und Server für das Repositorium verantwortlich, während die Universitätsbibliothek unter anderem rechtliche Unterstützung durch juristisches Fachpersonal zur Verfügung stellt, welches vor allem bei der Ausarbeitung von Archivierungs- und Nachnutzungsvereinbarungen im Rahmen des Aufbaus von FDAT benötigt wurde, und weiterhin zu den Themen Urheberrecht, Datenschutz und Persönlichkeitsrechte benötigt wird.

Unterhalb des IKM steht das eScience-Center mit dem Schwerpunkt Digital Humanities.[13] Eine der vorrangigen Aufgaben des eScience-Center als Core Facility[14] der Universität Tübingen stellt die Bereitstellung von IT-Dienstleistungen für Forschende aus dem Spektrum der Geistes- und Sozialwissenschaften dar. Die konkrete technische Entwicklung des FDAT-Repositoriums sowie die angebotenen Dienstleistungen für Forschende werden hierbei überwiegend ebenfalls von IT-Mitarbeitenden und Fachwissenschaftler/inne/n des eScience-Center erbracht.

12 Informations- Kommunikations- und Medienzentrum (IKM), Universität Tübingen, <www.uni-tuebingen.de/einrichtungen/zentrale-einrichtungen/informations-kommunikations-und-medienzentrum-ikm.html>, Stand: 10.08.2018.

13 Digital Humanities in Tübingen, Universität Tübingen, <www.escience.uni-tuebingen.de/digital-humanities.html>, Stand: 10.08.2018.

14 Core Facilities, Universität Tübingen, <https://www.uni-tuebingen.de/exzellenzinitiative/forschung/core-facilities.html>, Stand: 10.08.2018.

2. Das FDAT-Repositorium: Dienstleistungen und technische Umsetzung

Das Repositorium folgt bei der technischen Umsetzung seiner Dienste den Richtlinien eines offenen Archivinformationssystems OAIS[15], indem es wohl definierte Datenstrukturen (Einlieferungs-, Archiv- und Auslieferungspakete) verwendet, standardisierte Prozesse (Datenübernahme, Zugriff, etc.) umsetzt und klare Zuständigkeiten (Administration, Management) zur Regelung des Systems festlegt. Das OAIS ist ein akzeptierter Standard für die Umsetzung eines offenen digitalen Archivs zur langfristigen Aufbewahrung digitaler Objekte.

Bei der Realisierung des Systems wurde Wert auf die Verwendung von etablierten Open-Source-Komponenten gelegt und darauf, Eigenentwicklungen und damit verbundene Insellösungen nach Möglichkeit zu vermeiden. Auf die verwendeten Softwarelösungen wird in der Beschreibung der einzelnen Abschnitte des Datenlebenszyklus genauer eingegangen.

Das Repositorium begleitet Forschungsprojekte nach Möglichkeit in allen wesentlichen Lebensphasen ihrer Forschungsdaten, von der Planung und Erstellung, bis zur Nachnutzung und Erhaltung. Im Folgenden sollen die Dienstleistungen des Repositoriums anhand der einzelnen Abschnitte des Datenlebenszyklus (siehe Abbildung 2) beschrieben werden.

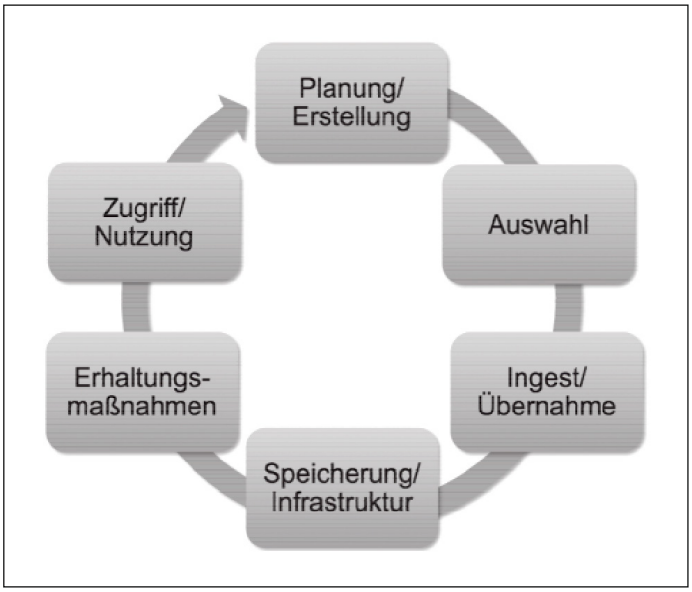

Abb. 2: Schematische Darstellung des Datenlebenszyklus (entnommen WissGrid[16]).

15 ISO 14721:2012, <http://www.iso.org/standard/57284.html>, Stand: 10.08.2018.

16 Ludwig, Jens; Enke, Harry (Hg.): Leitfaden zum Forschungsdaten-Management. Handreichungen aus dem WissGrid-Projekt, Glückstadt 2013. Online: <http://www.forschungsdaten.org/index.php/Datei:Leitfaden_Data-Management-WissGrid.pdf>, Stand: 10.08.2018.

2.1. Planung / Erstellung

In diesem ersten Abschnitt des Datenlebenszyklus (siehe Abbildung 2) sind vielfältige Aspekte des Datenmanagements vom Forschenden in Bezug auf seine Daten zu berücksichtigen. Immer mehr Forschungsförderer erwarten bereits bei der Beantragung von Projekten Aussagen zum Datenmanagement, sowie zur langfristigen Verfügbarkeit und Nachnutzbarkeit der Forschungsdaten. Entsprechende Aussagen hierzu werden in sogenannten Datenmanagementplänen (DMP) zusammengefasst.

Um Wissenschaftlerinnen und Wissenschaftler zu Beginn ihres Forschungsvorhabens darin zu unterstützen, bietet FDAT, neben entsprechender Beratung, derzeit ein webbasiertes Tool[17] zur Generierung eines DMP an.

Parallel dazu arbeitet das eScience-Center derzeit an der Bereitstellung einer Instanz des technisch deutlich ausgereifteren Research Data Management Organiser (RDMO)[18], einem von der DFG und weiteren Institutionen geförderten Projektes. Dieser erlaubt das Erfassen aller relevanten Planungsinformationen eines Forschungsprojektes in DMP und die Verwaltung aller Datenmanagementaufgaben über den gesamten Datenlebenszyklus hinweg. Bei der Zusammenstellung entsprechender Fragenkataloge für einen DMP wird sich das Repositorium an den Guidelines on FAIR Data Management[19] aus dem EU-Rahmenprogramm Horizont2020[20] orientieren und diese bei Bedarf an das jeweilige Forschungsprojekt anpassen.

2.2. Auswahl

Aus fachlichwissenschaftlicher Sicht ist die Auswahl archivwürdiger Daten zur langfristigen Aufbewahrung in einem Forschungsprojekt eine komplexe Fragestellung, die ein tieferes fachliches Verständnis voraussetzt und damit in aller Regel nicht von den Betreibern des FDAT-Repositoriums ausreichend unterstützt werden kann. Das Repositorium versteht sich in erster Linie als technischer Dienstleister für die Bitstream Preservation[21] der übergebenen Forschungsdaten. Eine Ausnahme bilden Forschungsdaten aus dem Schwerpunktbereich der Geistes- und Sozialwissenschaften, hierfür stehen Fachwissenschaftler/innen am eScience-Center für eine Beurteilung der Daten in gewissem Umfang zur Verfügung. Das Angebot einer Datenkuration für alle Fachgebiete ist momentan nicht vorgesehen.

Über die fachlich-wissenschaftlichen Kriterien der Datenauswahl hinaus existieren Grundsätze[22] zur Datenerhebung und Datenübername ins FDAT-Repositorium, welche die Auswahl der zu archivierenden Daten beeinflussen können.

17 Generator für Datenmanagementpläne, Forschungsdatenportal FDAT, <https://fdat.escience.uni-tuebingen.de/portal/#/service_downloads>, Stand: 10.08.2018.
18 Research Data Management Organiser RDMO, <https://rdmorganiser.github.io/>, Stand: 10.08.2018.
19 European Commission: H2020 Programme. Guidelines on FAIR Data Management in Horizon 2020, Version 3.0., 26.07.2016, <https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf>, Stand: 10.08.2018.
20 Horizont 2020, <www.horizont2020.de/>, Stand: 10.08.2018.
21 Bitstream Preservation, forschungsdaten.org, <www.forschungsdaten.org/index.php/Bitstream_Preservation>, Stand: 10.08.2018.
22 Leitlinien und Grundsätze für Forschungsdaten, Forschungsdatenportal FDAT, <https://fdat.escience.uni-tuebingen.de/portal/#/policies>, Stand: 10.08.2018.

Es können nur solche Daten ins Archiv übernommen werden, für die vom Datengeber eine minimale Erschließung mit Metadaten vorgenommen wurde.

Sensible Daten mit Personenbezug müssen vom Datengeber ausreichend anonymisiert bzw. pseudonymisiert übergeben werden, um das Offenlegungsrisiko zu minimieren.

Die zur Archivierung vorgesehenen Daten müssen in elektronischer Form in nicht proprietären Dateiformaten vorliegen und sich zum Zeitpunkt der Übergabe an das Repositorium in einem finalen Zustand befinden, der zur Publikation grundsätzlich geeignet ist.

Der Datengeber muss über die ausschließlichen Urheber- und Verwertungsrechte am zu übergebenden Datenbestand verfügen und darüber hinaus alle für diesen Datenbestand relevanten datenschutzrechtlichen Anforderungen eingehalten haben.

2.3. Übernahme / Ingest

Die Übernahme von Forschungsdaten ins Repositorium erfolgt in Form einer wohldefinierten Datenstruktur (SIP), die die Forschenden mit Hilfe der Open-Source Software Docuteam Packer[23] erstellen. Die von diesem Tool erzeugte Struktur des Einlieferungspakets folgt dem XML-Standard METS[24] und beinhaltet sowohl die Daten selbst als auch deskriptive (EAD[25]) und technische (PREMIS[26]) Metadaten. Das Tool ist konfigurierbar und wird von den Betreibern des Repositoriums an die Bedürfnisse des jeweiligen Forschungsprojekts angepasst. Hierzu zählen die Definition von Wertebereichen, Datentypen und anderen Validierungskriterien für die einzelnen Metadatenfelder sowie die Erweiterung um fachspezifische Metadatenschema.

Die übergebenen Forschungsdaten werden vor dem Einspielen ins Archiv standardmäßig systematisch auf Viren und Schadsoftware geprüft, auf Vollständigkeit der Metadaten sowie auf Validität der technischen Formate. Bezüglich letztgenanntem ist hier das Open-Source-Tool VeraPDF[27] hervorzuheben, welches die Validität archivfähiger PDF/A Dokumente ausführlich gegen den Standard (ISO 19005-1) prüft und dokumentiert.

Forschende erhalten zudem konkrete Unterstützung bei der systematischen Konvertierung ihrer Dateien in archivfähige Formate sowie eine Dokumentation[28] zum Thema über das Web-Portal des Repositoriums. Für die spätere Nachvollziehbarkeit der durchgeführten Aktionen werden Protokolle der Formatkonvertierung nach Möglichkeit mitarchiviert. Grundsätzlich werden neben den archivfähigen Dateien immer auch alle Originaldokumente (unkonvertiert) im Archiv abgelegt, jedoch ohne direkte Zugriffsmöglichkeit durch den Endnutzer.

23 docuteam packer, <https://wiki.docuteam.ch/doku.php?id=docuteam:packer>, Stand: 10.08.2018.
24 METS Schema & Documentaion, METS Metadata Encoding & Transmission Standard, <www.loc.gov/standards/mets/mets-schemadocs.html>, Stand: 10.08.2018.
25 EAD Encoded Archival Description, <http://www.loc.gov/ead/>, Stand: 10.08.2018.
26 PREMIS, <www.loc.gov/standards/premis/>, Stand: 10.08.2018.
27 VeraPDF, <http://verapdf.org/home/>, Stand: 10.08.2018.
28 Archivierung, Erhaltung und Nachnutzung ihrer Forschungsdaten, Forschungsdatenportal FDAT, <https://fdat.escience.uni-tuebingen.de/portal/#/deposit>, Stand: 10.08.2018, siehe Reiter Datenformate/Validierung.

Das Repositorium fordert bei Übernahme eines Datensatzes einen minimalen Satz an mitgelieferten Erschließungsinformationen. Gerade für das Feld des Autoren (Creator) wird anstelle eines Namens als Literal ein Authority Record/File gewünscht, welches über die Open Researcher and Contributor ID[29] (ORCID) erzeugt werden kann. Darüber hinaus unterstützt FDAT derzeit die Virtual International Authority File[30] (VIAF) sowie Normdateien[31] (GND) für Personen, die ebenfalls Bestandteil der VIAF sind.

Beim Ingest ins Repositorium werden automatisch weitere technische Metadaten erzeugt, zumeist aus den Schemata der Fedora Commons Repository Ontology[32] und PREMIS[33], von denen die Prüfsumme (Checksum) der elektronischen Datei als Merkmal der Datenintegrität hervorzuheben ist.

Eine Beschreibung aller im FDAT-Repositorium definierten Metadatenfelder ist online über das FDAT-Portal zugänglich.

Ein wichtiges Merkmal des Repositoriums ist die Erweiterungsfähigkeit auf beliebige fachspezifische Metadatenschemata, welche eine ausreichende Beschreibung von Forschungsdaten aller Disziplinen ermöglichen soll.

Der Prozess der Einlieferung (Ingest) ins Repositorium selbst erfolgt über eine eigens am eScience-Center entwickelte Software, welche alle durchgeführten Aktionen detailliert protokolliert und anschließend mit archiviert. Eine Eigenentwicklung an dieser Stelle erwies sich als notwendig, da es keine frei verfügbare Ingest-Software gab, die mit dem im Repositorium verwendeten Komponenten (Fedora Commons, Docuteam Packer, Apache Solr/Lucene) interagieren konnte.

2.4. Speicherung / Infrastruktur

Über die Exzellenzinitiative von Bund und Ländern[34] stellt die Universität Tübingen im Rahmen ihres Zukunftskonzepts[35] dem FDAT-Repositorium eine moderne technische Infrastruktur sowie exzellente Dienstleistungen zur Verfügung. Das Rechenzentrum der Universität betreibt und wartet diese Infrastruktur (Core-Facility-Cluster) und erweitert sie gegebenenfalls nach den Erfordernissen des Repositoriums.

Die Open-Source-Archivsoftware Fedora Commons 4.x[36] bildet softwareseitig die Kernkomponente des Repositoriums zur Speicherung von Forschungsdaten. Wichtige Entscheidungskriterien für Fedora

29 ORCID, <https://orcid.org/>, Stand: 10.08.2018.
30 Virtual International Authority File (VIAF), <https://viaf.org/>, Stand: 10.08.2018.
31 Gemeinsame Normdatei (GND), <www.dnb.de/DE/Standardisierung/GND/gnd_node.html>, Stand: 10.08.2018.
32 Fedora Commons Repository Ontology, <https://fedora.info/definitions/v4/2016/10/18/repository>, Stand: 10.08.2018.
33 Preservation Metadata, Implementation Strategies (PREMIS) Ontology, <http://id.loc.gov/ontologies/premis.html>, Stand: 10.08.2018.
34 Exzellenzinitiative des Bundes und der Länder (2005-2017), DFG, <www.dfg.de/foerderung/programme/exzellenzinitiative>, Stand: 10.08.2018.
35 Exzellenzinitiative, Universität Tübingen, <www.uni-tuebingen.de/exzellenzinitiative.html>, Stand: 10.08.2018.
36 Fedora, Duraspace, <http://fedorarepository.org>, Stand: 10.08.2018.

als Speichersystem waren zum einen die hohe Flexibilität bezüglich unterstützter Dateiformate, da FDAT aufgrund seines generischen Anspruchs mit heterogenen Daten umgehen muss. Fedora skaliert sehr gut für große Datenmengen und bietet zudem flexible Speichermöglichkeiten wie Datenbanken und Filesysteme an. Letzteres ermöglicht auch eine sinnvolle hierarchische Strukturierung der Daten.

2.5. Erhaltungsmaßnahmen

Das FDAT-Repositorium nimmt im Rahmen seiner technischen Möglichkeiten die notwendigen Maßnahmen zur dauerhaften Erhaltung aller archivierten Daten wahr. Als mögliches Maß für das Niveau der Datenerhaltungsaktivitäten orientiert sich FDAT an den Bewertungskriterien, die durch die National Digital Stewardship Alliance (NDSA)[37] gegeben sind. Hier werden fünf entscheidende Aspekte eines Repositoriums definiert und einer von vier verschiedenen Ebenen von Erhaltungsmaßnahmen zugeordnet. Diese Aspekte umfassen Dateiformate, Metadaten, Datenintegrität, Informationssicherheit, Speicherung und geografische Lage.[38] Das derzeitige Level an Erhaltungsmaßnahmen für Forschungsdaten in FDAT ist über das Web-Portal[39] einsehbar.

Um den Zugang und die Nachnutzbarkeit der Daten im Archiv auch zukünftig zu sichern, geben FDAT-Mitarbeiter/innen den Forschenden jeweils Empfehlungen darüber, ob und wann aufgrund neuer Standards eine Migration digitaler Objekte im Archiv in andere Formate notwendig wird. Ebenfalls wird die Entwicklung sämtlicher für das Archivsystem eingesetzten Softwarekomponenten verfolgt und Entscheidungen über den Wechsel zu etablierteren oder stabileren Softwarekomponenten erwogen.

2.6. Zugriff / Nutzung

Das FDAT-Repositorium ist generell bestrebt, Forschungsdaten online frei verfügbar zu machen. Forschende können sich jedoch dazu entschließen, ihre Daten durch Festlegung von Zugriffsbeschränkungen vorübergehend, maximal für fünf Jahre, nur einem eingeschränkten Personenkreis über das Web-Portal zugänglich zu machen. Jedem digitalen Objekt im FDAT-Repositorium muss zwingend eine Lizenz zur Regelung seiner Nutzungsbedingungen durch Dritte vom Datengeber zugeordnet werden, vorzugsweise nach dem Lizenzierungssystem der Creative Commons. Hierbei folgt das Repositorium dem Appell der DFG zur Nutzung offener Lizenzen[40] für Forschungsdaten.

Um die Auffindbarkeit und den Nachweis archivierter Forschungsdaten stets zu gewährleisten, werden Metadaten in FDAT zwingend Open Access unter der Lizenz CC0 1.0 Universal (CC0 1.0) Public Domain[41] publiziert.

Die Lizenz wird als Teil der Metadateninformation eines Datensatzes behandelt und daher zusammen mit dem Datensatz selbst dauerhaft im Archiv gespeichert. Darüber hinaus werden

37 National Digital Stewardship Alliance (NDSA), <http://ndsa.org/>, Stand: 10.08.2018.
38 Zur Erläuterung des Bewertungssystem vgl. NDSA Levels of Preservation, <www.digitalpreservation.gov:8081/ndsa/activities/levels.html>, Stand: 10.08.2018.
39 Archivierung, Erhaltung und Nachnutzung ihrer Forschungsdaten, Forschungsdatenportal FDAT, <https://fdat.escience.uni-tuebingen.de/portal/#/deposit>, Stand: 10.08.2018, siehe Reiter Erhaltungsplanung.
40 DFG: Appell zur Nutzung offener Lizenzen in der Wissenschaft, Information für die Wissenschaft Nr. 68, 20.11.2014, <www.dfg.de/foerderung/info_wissenschaft/2014/info_wissenschaft_14_68/>, Stand: 10.08.2018.
41 CC0 1.0 Universal, <https://creativecommons.org/publicdomain/zero/1.0>, Stand: 10.08.2018.

Nutzungsbedingungen über das Webportal visuell hervorgehoben und auf ausführliche rechtliche Informationen verlinkt.

Die Visualisierung und der Download der Forschungsdaten sowie die Navigation im Datenbestand erfolgen über ein eigens entwickeltes Portal/Frontend. Die schnelle Auffindbarkeit der Ressourcen wird durch das zusätzliche Vorhalten ihrer Metadaten in einer leistungsfähigen Suchmaschine[42] gewährleistet. Diese ermöglicht sowohl die Freitextsuche im Datenbestand als auch eine facettenbasierte Suche über geeignete vorausgewählte Kategorien, wobei letztere überwiegend durch kontrollierte Vokabulare gestützt ist.

Durch das Web-Portal werden die Benutzer/innen zudem über Nachnutzungsbestimmungen für alle verfügbaren Ressourcen informiert.

Sämtliche im FDAT-Repositorium vorhandenen Metadateninformationen können systematisch mit Hilfe des etablierten Protokolls OAI-PMH[43] gesammelt werden. Derzeit werden als Ausgabeformate Dublin Core[44] in Form des Standards oai_dc[45] sowie MARC 21[46] unterstützt. FDAT ist zu diesem Zweck als Data-Provider[47] bei der Open Archives Initiative[48] (OAI) registriert und validiert.

Die Authentifizierung der Benutzer/innen am Portal wird über die vom Deutschen Forschungsnetz DFN[49] bereitgestellte verteilte Authentifizierungs- und Autorisierungsinfrastruktur DFN-AAI[50] ermöglicht, wobei FDAT ein registrierter Service-Provider in diesem System ist. Dieses baut technisch auf dem etablierten Standard SAML[51] und seiner Weiterentwicklung Shibboleth[52] auf. Über die Anbindung an DFN-AAI haben Personen aller registrierten Forschungseinrichtungen in Deutschland die Möglichkeit ein Benutzerkonto am FDAT-Repositorium einzurichten.

Um neben dem institutionellen auch einen personalisierten Zugang zu FDAT zu ermöglichen, ist die Anbindung an ORCID geplant.

Das Zugriffsrecht eines Benutzers auf Forschungsdatenobjekte kann über das Portal feingranular durch die Administratoren am System festgelegt werden. Alle zugriffsbeschränkten Ressourcen werden zudem explizit im Archivierungs- und Nachnutzungsvertrag aufgeführt, welcher zwischen den Datengebern und dem Repositorium abgeschlossen wird.

42 Solr, <http://lucene.apache.org/solr>, Stand: 10.08.2018.
43 The Open Archives Initiative Protocol for Metadata Harvesting, <www.openarchives.org/OAI/openarchivesprotocol.html>, Stand: 10.08.2018.
44 Dublin Core Metadata Initiative, <http://dublincore.org/>, Stand: 10.08.2018.
45 oai_dc, <www.openarchives.org/OAI/2.0/oai_dc.xsd>, Stand: 10.08.2018.
46 MARC 21, <www.loc.gov/marc/marcdocz.html>, Stand: 10.08.2018.
47 OAI-PMH Registration Record, <www.openarchives.org/Register/BrowseSites?viewRecord=http://fdat.escience.uni-tuebingen.de/portal/rest/oai>, Stand: 10.08.2018.
48 Open Archives Initiative, <www.openarchives.org/>, Stand: 10.08.2018.
49 Deutsches Forschungsnetz, <www.dfn.de/>, Stand: 10.08.2018.
50 DFN-AAI - Authentifikations- und Autorisierungs-Infrastruktur, <www.aai.dfn.de/>, Stand: 10.08.2018.
51 SAML XML.org, <http://saml.xml.org/>, Stand: 10.08.2018.
52 Shibboleth, <www.shibboleth.net>, Stand: 10.08.2018.

3. Übergeordnete Aspekte des Datenlebenszyklus

Neben den Aspekten des Forschungsdatenmanagements, die eindeutig einem Bereich im Datenlebenszyklus zugeordnet werden können, sollen an dieser Stelle übergeordnete Themen diskutiert werden sowie ihre konkrete Umsetzung in FDAT.

3.1. Kosten und Finanzierung

Das FDAT-Repositorium verfolgt eine einfache Kostenpolitik bezüglich der aufzubewahrenden Forschungsdaten, mit einem grundsätzlichen Freikontingent von 5 GB für jedes Forschungsprojekt. Über dieses Kontingent hinaus stellt das Rechenzentrum der Universität den Datengebern die Kosten für die Anschaffung von Speicherhardware (Storage) in Rechnung. Dabei übernimmt das Rechenzentrum die laufenden Personalkosten für Installation, Betrieb und Wartung sowie die laufenden Energiekosten für Storage in Eigenleistung.

3.2. Metadaten

Die Mitarbeiter/innen des FDAT-Repositoriums unterstützen den Datengeber beim Auffinden eines für sein Fachgebiet etablierten und geeigneten Metadatenschemas, idealerweise noch vor der Erhebung der Daten selbst.

Da solche Schemata für viele Forschungsgebiete jedoch nicht existieren, müssen Metadatenfelder und ihre Bedeutung oftmals auf Projektebene definiert werden, wobei ihre Wiederverwendbarkeit und Akzeptanz durch andere Forschungsprojekte, aus angrenzenden oder selbst demselben Fachgebiet, nicht garantiert werden können. Dieser Sachverhalt stellt ein generelles Problem für Forschungsdatenrepositorien und das Datenmanagement im Allgemeinen dar.

Einen interessanten technischen Lösungsansatz für dieses Problem, entwickelt für Sprachressourcen, bietet die Component Metadata Initiative (CMDI)[53], ein Framework zum Erstellen und Registrieren selbstdefinierter Metadatenschemata. Alle Schemata werden hierbei in der CMDI-Datenbank vorgehalten und sind zitierfähig.

Ein weiterer wichtiger Aspekt von Metadaten ist ihre Verwendung im Zusammenspiel mit kontrollierten Vokabularen. Die Vermeidung von Synonymen und die festgelegte Benennung bestimmter Sachverhalte erleichtern sowohl die Erschließung als auch das Auffinden von Forschungsdaten durch reglementierte Suchbegriffe, wodurch eine umfassende Suche nach Informationen in großen Datenmengen überhaupt erst effizient wird.

Für insgesamt 6 der in FDAT definierten Metadatenfelder konnten geeignete kontrollierte Vokabulare gefunden werden (siehe Web-Portal[54]), deren Einhaltung bei der Datenübernahme systematisch

53 Component Metadata, CLARIN, <www.clarin.eu/content/component-metadata>, Stand: 10.08.2018.
54 Archivierung, Erhaltung und Nachnutzung ihrer Forschungsdaten, Forschungsdatenportal FDAT, <https://fdat.escience.uni-tuebingen.de/portal/#/deposit>, Stand: 10.08.2018, siehe Reiter Metadaten/Vokabulare.

geprüft wird. Hervorgehoben werden soll an dieser Stelle das Vokabular JACS[55] und sein Nachfolger HECoS[56], die die Zuordnung eines Datensatzes zu einer bestimmten Forschungsdisziplin ermöglichen.

3.3. Identifikatoren

Die systematische und eindeutige Identifikation von digitalen Objekten ist ein nicht zu unterschätzender, über den gesamten Lebenszyklus der Forschungsdaten hinweg wichtiger Bestandteil des Datenmanagements. Spätestens jedoch zum Zeitpunkt der Publikation digitaler Objekte im Internet muss die Auffindbarkeit durch eindeutige und persistente Identifikatoren (PID) gegeben sein.

Sämtliche im Repositorium befindlichen Datenpakete besitzen eine automatisch zugewiesene PID nach dem Handle-System[57], worüber die dauerhafte Zitierbarkeit der Forschungsdaten gewährleistet ist. Die Zahl der benötigten Identifikatoren hängt dabei direkt mit der Frage nach der Granularität, also der kleinsten zu speichernden Einheit der Forschungsdaten im Projekt zusammen. Diese muss konkret festgelegt werden und ist wesentlicher Bestandteil des Datenmanagements.

Durch die Vergabe von PIDs und den lokalen Betrieb der dafür notwendigen Infrastruktur garantiert das FDAT-Repositorium die dauerhafte Zitierbarkeit archivierter Forschungsdaten für die vereinbarte Dauer der Archivierung. Datensätze im FDAT-Repositorium besitzen den Identifikator-Präfix 10900.1 und können darüber eindeutig zugeordnet werden.

Aufgrund der hohen Bekanntheit und Akzeptanz in der Forschungslandschaft wird neben den Handle-Identifikatoren zusätzlich die Bereitstellung von DOIs[58] für digitale Objekte in FDAT erwogen.

3.4. Recht

Vereinbarungen zur Datenüberlassung zwischen dem Datengeber und dem Repositorium als Dienstleister bilden die rechtliche Grundlage für die Langzeitarchivierung von Forschungsdaten. Die Vorlage einer solchen Vereinbarung stellt das Repositorium auf seinem Web Portal frei zum Download[59] zur Verfügung.

Das FDAT-Repositorium benötigt vom Datengeber die Einräumung eines nicht ausschließlichen Nutzungsrechts für die Archivierung und Verbreitung seiner Forschungsdaten, ohne zeitliche Beschränkung. Das Nutzungsrecht umfasst hierbei insbesondere die folgenden Aspekte.

- Technische, nicht inhaltsbezogene Bearbeitung des Datenbestandes, insbesondere Konvertierungen in archivfähige Datenformate
- Archivierung des Datenbestandes

55 JACS 3.0, HESA, <www.hesa.ac.uk/support/documentation/jacs/jacs3-detailed>, Stand: 10.08.2018.
56 HECoS, HESA, <www.hesa.ac.uk/innovation/hecos>, Stand: 10.08.2018.
57 Handle, <www.handle.net>, Stand: 10.08.2018.
58 DOI, <www.doi.org/>, Stand: 10.08.2018.
59 Archivierung, Erhaltung und Nachnutzung ihrer Forschungsdaten, Forschungsdatenportal FDAT, <https://fdat.escience.uni-tuebingen.de/portal/#/deposit>, Stand: 10.08.2018, siehe Reiter Rechtliche Aspekte.

- Migration des Datenbestandes in technische Nachfolgesysteme des IKM, oder andere Archivsysteme
- Veröffentlichung des Datenbestandes über eine Onlineplattform, einschließlich der Möglichkeit zum Download
- Weitergabe von Erschließungsinformationen (Metadaten) zum Datenbestand an andere elektronische Nachweis- und Archivsysteme

4. Verknüpfung zwischen Forschungsdaten und Publikationen

In Zusammenarbeit mit dem Publikationsdienst der Universitätsbibliothek hat das FDAT-Repositorium eine für Forschende niederschwelligen Lösung für die Verknüpfung ihrer Publikationen mit den dazugehörigen Forschungsdaten bereitgestellt. Das Standardszenario sieht Doktorandinnen und Doktoranden vor, die ihre Dissertation der Bibliothek übergeben und dazu die Möglichkeit erhalten, in einem konzertierten Prozess ihre Forschungsdaten als einzelnes gepacktes Dokument mit abzuliefern, ohne sich dafür zusätzlich an das eScience-Center als Repositoriumsbetreiber wenden zu müssen.

Der hierfür überarbeitete Veröffentlichungsvertrag für Publikationen deckt die Forschungsdaten in einer Zusatzklausel mit ab. Der Publikationsdienst erfragt zudem vom Datengeber einen minimalen Satz an Metadaten für die Forschungsdaten und übergibt in einem automatisierten Prozess diese Informationen dem eScience-Center, zusammen mit dem für die Publikation vergebenen Identifikator (DOI, Handle oder URN). Das eScience-Center gibt einen seinerseits generierten PID (Handle) für die Forschungsdaten zurück an das Publikationssystem, um somit eine direkte Verlinkung zwischen beiden Datensätzen zu erhalten.

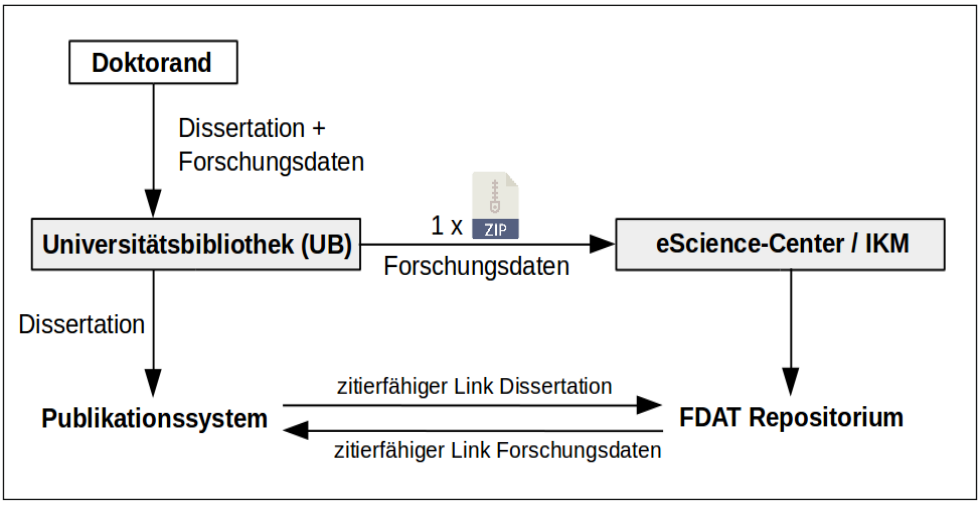

Abb. 3: Aktionsschema für die Verknüpfung von Publikationen und dazugehörigen Forschungsdaten in der Zusammenarbeit zwischen dem Publikationsdienst der Universitätsbibliothek und dem Forschungsdatenrepositorium FDAT.

5. Aktuelle Entwicklungen und Ausblick

Neben dem institutionellen Repositorium FDAT gibt es weitere fachspezifische Repositorien an der Universität Tübingen. Um die Nachhaltigkeit dieser meist getrennt voneinander entstehenden Systeme signifikant zu erhöhen, wurde das Infrastrukturprojekt Open Research Data Portal (ORDP)[60], gefördert vom MWK Baden-Württemberg, ins Leben gerufen.

Entwickelt wird ORDP von den Tübinger Core-Facilities eScience-Center und Zentrum für Quantitative Biologie (QBIC)[61] mit dem Ziel, eine für beide Systeme übergeordnete Portalinfrastruktur zur Verwaltung und Publikation von Forschungsdaten aufzubauen, welche zugleich die Komplexität und Redundanz beider Systeme reduzieren und damit ihre Nachhaltigkeit erhöhen soll.

Letztlich soll mit der ORDP-Infrastruktur in Tübingen allen existierenden und zukünftigen Forschungsdatenrepositorien der Universität eine gemeinsam nutzbare und nachhaltige Portalplattform zur Verfügung gestellt werden, welche die Entwicklungszeit künftiger Repositorien signifikant verkürzen soll.

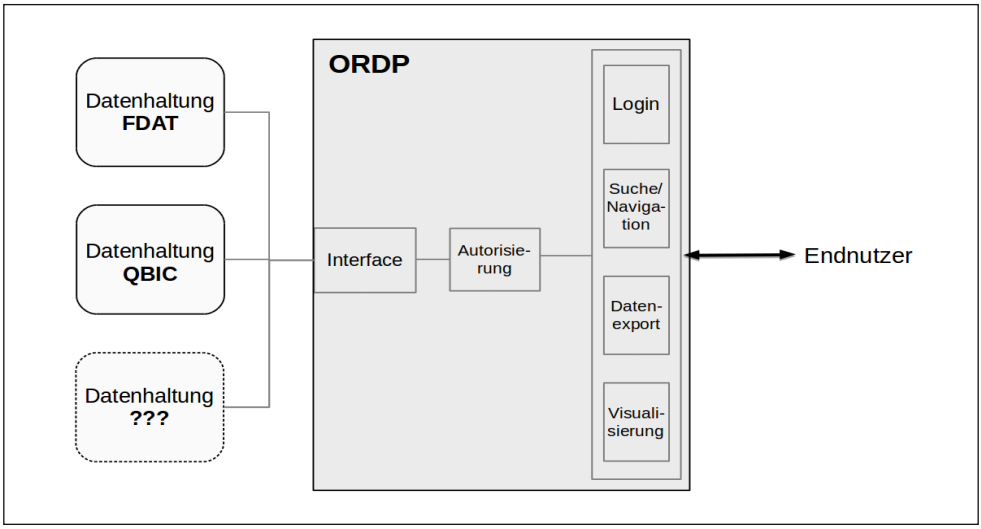

Abb. 4: Schematische Darstellung der Portalinfrastruktur ORDP und der daran angebundenen Datenhaltungssysteme der Tübinger Core-Facilities eScience-Center (FDAT) und QBIC. Das System soll darüber hinaus die Anbindung der Datenhaltung weiterer Repositorien ermöglichen.

60 Projekt ORDP Open Data Research Portal, Forschungs-Information Tübingen (FIT), <https://fit.uni-tuebingen.de/Project/Details?id=4667>, Stand: 10.08.2018.

61 Zentrum für Quantitative Biologie, Universität Tübingen, <www.qbic.uni-tuebingen.de/>, Stand: 10.08.2018.

6. Aktuelle Herausforderungen für das FDAT-Repositorium

An dieser Stelle sollen die wichtigsten aktuellen Herausforderungen an das Repositorium und seine Betreiber diskutiert werden, die seit dem Bestehen des Systems entstanden sind.

In Forschungsdisziplinen mit starkem Raumbezug wie der Archäologie und den Geowissenschaften haben in den letzten Jahren neue technische Verfahren Einzug gehalten und die erzeugten Datenmengen enorm steigen lassen. Als Beispiele seien hier aufwendig erzeugte 3D Modelle[62] von archäologischen Fundstücken und Grabungsanlagen genannt, sowie hochauflösende Luftbildaufnahmen[63]. Diese Daten werden mit moderner Messtechnik zum Teil auch am eScience-Center der Universität Tübingen erzeugt. Diese Daten in geeignete Langzeitformate zu überführen und in ihrem Umfang zu handhaben, ist eine Herausforderung für Repositorien, mit der sich derzeit spezielle Einrichtungen befassen.[64]

Um auch sensible Forschungsdaten mit Personenbezug nachnutzen zu können, benötigen Forschende gerade aus den Fachrichtungen der Sozialwissenschaften zum Teil konkrete technische Unterstützung bei der Anonymisierung bzw. Pseudonymisierung ihrer Daten. An diesen Bedarf muss das FDAT-Repositorium seine Dienstleistungen weiter anpassen und prüft derzeit geeignete technische Hilfsmittel, wie das von OpenAire[65] bereitgestellte Tool Amnesia[66].

Über alle Fachbereiche hinweg besteht der ausdrückliche Bedarf der Forschenden nach zumindest zeitlich begrenztem Embargo für ihre Daten, vorrangig um Publikationsinteressen zu schützen. Dieses Bedürfnis steht der Zielsetzung des Repositoriums nach freiem Zugang zu Forschungsdaten zwar entgegen, als Kompromisslösung erlauben die Richtlinien für FDAT dennoch eine Zugriffsbeschränkung für archivierte Daten bis zu maximal 5 Jahre. Als Ergebnis unterliegt der überwiegende Teil der Daten im Repositorium momentan einer Zugriffsbeschränkung.

Deskriptive Metadaten existieren für Forschungsdaten häufig in geringerem Umfang und zum Teil weniger gut organisiert, als für eine Übernahme ins Repositorium und eine gute Nachnutzbarkeit notwendig wäre. Es werden oft individuelle Lösungen benötigt, um alle Metadateninformationen zu den Forschungsdaten zusammenzutragen und fehlende Informationen zu ergänzen (z.B. durch Vorbelegung mit Standardwerten), um somit ein valides Einlieferungspaket (SIP) zu erzeugen. Genügend flexible Schnittstellen in Open-Source-Datenerfassungsprogrammen stehen unserer Erfahrung derzeit nicht zur Verfügung.

62 3D-Museum, Universität Tübingen, <www.unimuseum.uni-tuebingen.de/de/sammlungen/3d-museum.html>, Stand: 10.08.2018.
63 Dokumentation historischer Denkmäler im Oman, Univerität Tübingen, <www.escience.uni-tuebingen.de/projekte/dokumentation-historischer-denkmaeler-im-oman.html>, Stand: 10.08.2018.
64 3D und Virtual Reality, IANUS, <www.ianus-fdz.de/it-empfehlungen/3d>, Stand: 10.08.2018.
65 Open AIRE, <www.openaire.eu/>, Stand: 10.08.2018.
66 Amnesia, <https://amnesia.openaire.eu/>, Stand: 10.08.2018.

Die Akzeptanz des Angebotes, publikationsbegleitend Forschungsdaten beim Publikationsdienst der Universitätsbibliothek abliefern zu können, ist derzeit sehr gering. Bestehende technische und rechtliche Hürden für die Forschenden müssen überprüft und – wo möglich – weiter abgesenkt werden. Darüber hinaus wollen die beteiligten Einrichtungen des Repositoriums verstärkt Aufklärungsarbeit über diesen Dienst leisten, bis in die forschenden Arbeitskreise hinein.

Literaturverzeichnis

- DFG: Appell zur Nutzung offener Lizenzen in der Wissenschaft, Information für die Wissenschaft Nr. 68, 20.11.2014, <www.dfg.de/foerderung/info_wissenschaft/2014/info_wissenschaft_14_68/>, Stand: 10.08.2018.

- DFG: Leitlinien zum Umgang mit Forschungsdaten, 2015, <www.dfg.de/download/pdf/foerderung/antragstellung/forschungsdaten/richtlinien_forschungsdaten.pdf>, Stand: 10.08.2018.

- DFG: Sicherung guter wissenschaftlicher Praxis. Empfehlungen der Kommission „Selbstkontrolle in der Wissenschaft“, Weinheim 2013. Online: <http://doi.org/10.1002/9783527679188.oth1>, Stand: 10.08.2018.

- European Commission: H2020 Programme. Guidelines on FAIR Data Management in Horizon 2020, Version 3.0., 26.07.2016, <https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf>, Stand: 10.08.2018.

- Ludwig, Jens; Enke, Harry (Hg.): Leitfaden zum Forschungsdaten-Management. Handreichungen aus dem WissGrid-Projekt, Glückstadt 2013. Online: <http://www.forschungsdaten.org/index.php/Datei:Leitfaden_Data-Management-WissGrid.pdf>, Stand: 10.08.2018.