Offene Forschungsdaten an der Universität Heidelberg: von generischen institutionellen Repositorien zu fach- und projektspezifischen Diensten

Jochen Apel, Universitätsbibliothek Heidelberg
Fabian Gebhart, Universitätsrechenzentrum Heidelberg
Leonhard Maylein, Universitätsbibliothek Heidelberg
Martin Wlotzka, Universitätsrechenzentrum Heidelberg

Zusammenfassung:
Die Universität Heidelberg hat 2014 das Kompetenzzentrum Forschungsdaten als gemeinsame Serviceeinrichtung der Universitätsbibliothek und des Universitätsrechenzentrums eingerichtet. Der vorliegende Beitrag stellt die Angebote des Kompetenzzentrums zur Publikation von Forschungsdaten vor, fasst bisherige Erfahrungen zusammen und diskutiert auf dieser Grundlage exemplarisch die Rolle von institutionellen Veröffentlichungsplattformen für Open Research Data. Im Einzelnen werden dabei das institutionelle Datenrepositorium heiDATA, die Bild- und Multimediadatenbank heidICON sowie die derzeitige Weiterentwicklung des Dienstleistungsportfolios des Kompetenzzentrums im Rahmen des Projekts „Community-spezifische Forschungsdatenpublikation (CS-FDP)“ vorgestellt.

Summary:
In 2014 Heidelberg University established the Competence Centre for Research Data as a joint facility of the University Library and the university's Computing Centre. This article describes the Competence Centre's services for publishing research data and examines on that basis the role of institutional publication platforms for open research data. In particular the paper discusses the institutional research data repository heiDATA, the image and multimedia database heidICON and the current refinement of the Competence Centre's service portfolio within the project "Community Specific Research Data Publication (CS-FDP)".

Zitierfähiger Link (DOI): https://doi.org/10.5282/o-bib/2018H2S61-71
Autorenidentifikation: Apel, Jochen: GND 1025737180, ORCID: https://orcid.org/0000-0002-0395-4120; Gebhart, Fabian: GND 1148481745, ORCID: https://orcid.org/0000-0003-2466-8120; Maylein, Leonhard: GND 1019546514, ORCID: https://orcid.org/0000-0002-9423-489X; Wlotzka, Martin: GND 1031120645, ORCID: https://orcid.org/0000-0003-1794-4799
Schlagwörter: Forschungsdaten; Forschungsdatenmanagement; Forschungsdatenrepositorium; Open Research Data

1. Forschungsdaten- und Publikationsservices an der Universität Heidelberg

Das Kompetenzzentrum Forschungsdaten der Universität Heidelberg (KFD) ist eine gemeinsame Serviceeinrichtung der Universitätsbibliothek und des Universitätsrechenzentrums, die Wissenschaftlerinnen und Wissenschaftler beim strukturierten und nachhaltigen Management ihrer Forschungsdaten

unterstützt.[1] Der vorliegende Beitrag stellt die Angebote des KFD zur Publikation von Forschungsdaten vor, fasst bisherige Erfahrungen zusammen und diskutiert auf dieser Grundlage exemplarisch die Rolle von institutionellen Veröffentlichungsplattformen für Open Research Data. Im Einzelnen werden dabei das institutionelle Datenrepositorium heiDATA[2], die Bild- und Multimediadatenbank heidICON[3] sowie die derzeitige Weiterentwicklung der KFD-Dienstleistungen im Rahmen des MWK-geförderten Projekts Community-spezifische Forschungsdatenpublikation (CS-FDP)[4] vorgestellt.

Diese Dienste zur Forschungsdatenpublikation sind zum einen in das umfassendere Serviceportfolio des KFD zum Forschungsdatenmanagement (FDM) eingebettet. Das KFD bietet darüber hinaus fachliche Beratung zu technischen und organisatorischen Themen sowie Informationen über rechtliche Aspekte. Dies umfasst auch Unterstützung bei der Erstellung von Projektanträgen und Datenmanagementplänen. Hinzu kommen generische IT-Services zur aktiven Arbeit mit Daten im Projekt, z.B. Dateiablage und –Sharing mit dem Dienst heiBOX, die zentrale Speicherung wissenschaftlicher Daten durch den Service SDS@HD oder der Zugang zu Ressourcen für das Hochleistungsrechnen.[5] Zum anderen ergänzen die Datenpublikationsdienste die breite Palette von universitären Angeboten zur Open-Access-Veröffentlichung, die derzeit folgende Publikationsmöglichkeiten umfasst: [6]

- heiUP: Heidelberg University Publishing
- heiBOOKS: Heidelberger E-Books
- heiJOURNALS: Heidelberger Open-Access-Zeitschriften
- heiDOK: Heidelberger Dokumentenserver
- heidICON: Heidelberger Bild- und Multimediadatenbank
- heiDATA: Heidelberger Forschungsdatenrepositorium
- Publikationsplattformen für Altertumswissenschaften, Kunstgeschichte und Asienwissenschaft im Kontext der Fachinformationsdienste (FID)
- Digitale Editionen für Textcorpora

2. heiDATA – das Forschungsdatenrepositorium der Universität Heidelberg

Mit der Einrichtung des KFD wurde 2014 zugleich die Einführung eines institutionellen Forschungsdatenrepositoriums beschlossen. Das Repositorium soll insbesondere für diejenigen Forschenden der Universität ein Angebot zur Datenpublikation zur Verfügung stellen, in deren Forschungsbereichen es keine geeigneten fachlichen Repositorien gibt. Darüber hinaus soll es eine Möglichkeit zur Veröffentlichung für Datensupplemente schaffen, die Textpublikationen auf den oben genannten Heidelberger Open-Access-Plattformen ergänzen. heiDATA, so der Name der Plattform, wurde im Sommer 2014

1 Kompetenzzentrum Forschungsdaten der Universität Heidelberg, http://data.uni-heidelberg.de/.
2 heiDATA, https://heidata.uni-heidelberg.de/.
3 heidICON, http://heidicon.ub.uni-heidelberg.de.
4 CS-FDP, https://www.urz.uni-heidelberg.de/de/cs-fdp.
5 Vgl. „Dienste und Angebote," KFD, zuletzt geprüft am 22.05.2018, http://data.uni-heidelberg.de/dienste.html.
6 Eine Übersicht über diese Plattformen findet sich hier: „Elektronisches Publizieren – Open Access," Universität Heidelberg, zuletzt geprüft am 22.05.2018, www.openaccess.uni-hd.de.

auf Grundlage der Open-Source-Software Dataverse[7] implementiert und bietet folgende wesentliche Eigenschaften, die für nachhaltige Datenpublikationen erforderlich sind:

- Persistente zitierfähige Adressierung von Datensätzen über DOI
- Unterstützung unterschiedlicher Open-Content-Lizenzen
- Verknüpfung von Forschungsdaten mit zugehörigen Publikationen
- Individuelle Berechtigungsmodelle für Zugriffe (inkl. Zugriffsmöglichkeiten für Gutachter vor der Freischaltung via Private URLs)
- Publikation unterschiedlicher Dateiformate
- Unterstützung verschiedener domänenspezifischer Metadatenstandards (z.B. DDI u.a. für sozialwissenschaftliche Daten, ISA-Tab für die Lebenswissenschaften oder das VOResource Schema für die Astronomie sowie weitere)[8]
- Untergliederung in Sektionen, sog. Dataverses, z.B. für Projekte, Forschungsgruppen, Institute, Journals oder auch einzelne Personen
- Kuration der Daten auf Datensatz- und Dateiebene als Dienstleistung des KFD
- Nachweis aller Datensätze im Data-Citation-Index[9] sowie weiteren Nachweissystemen über offene API
- Nachweis von Datenpublikationen in der Hochschulbibliographie heiBIB basierend auf der Katalogisierung der Datenpublikationen im Südwestdeutschen Bibliotheksverbund (möglich für eigene und externe Repositorien durch Verlinkung)[10]

Die Dataverse-Software wird maßgeblich am Harvard Institute for Quantitative Social Science (IQSS) entwickelt. Weltweit gibt es bislang 33 öffentliche Instanzen der Software (Stand: 23.4.2018). Deren größte ist das Harvard Dataverse mit über 75.000 publizierten Datensätzen, das als generisches Angebot Forschenden weltweit die Möglichkeit bietet, kostenfrei Forschungsdaten zu publizieren.[11] Interessant ist zudem das DataverseNL, das als gemeinsames Angebot von 13 niederländischen Forschungseinrichtungen betrieben wird.[12] Die Entwicklung von Dataverse erfolgt als Open-Source-Projekt. Die Community vernetzt sich, neben der offenen Softwareentwicklung und dem zugehörigen Austausch auf GitHub[13], über ein Forum[14], monatliche Community Calls[15] sowie jährliche Community-Meetings.[16]

7 The Dataverse project, https://dataverse.org/.
8 DDI steht für den gleichnamigen Metadatenstandard der Data Documentation Iniative, ISA-Tab ist das sog. Investigation/Study/Assay tab-delimited Format für Metadaten aus dem Bereich der omics-basierten Methoden, das VOResource Schema ist das Metadatenschema der International Virtual Observatory Alliance. Eine Übersicht über sämtliche von Dataverse unterstützten Metadatenstandards inkl. Verlinkungen zu deren Dokumentationen findet sich auf http://guides.dataverse.org/en/latest/user/appendix.html.
9 Der Data Citation Index ist ein separat lizenzierbarer Teilindex des Web of Science, der ausgewählte Forschungsdatenpublikationen sowie Zitierungen dieser Datensätze nachweist. „Data Citation Index," Clarivate Analytics, zuletzt geprüft am 22.05.2018, http://wokinfo.com/products_tools/multidisciplinary/dci/
10 heiBIB, http://heibib.uni-hd.de.
11 Harvard Dataverse, https://dataverse.harvard.edu/
12 DataverseNL, https://dataverse.nl/
13 https://github.com/IQSS/dataverse.
14 https://dataverse.org/forum.
15 https://dataverse.org/community-calls.
16 https://dataverse.org/events.

Wesentliches Charakteristikum des in Heidelberg verfolgten Ansatzes zur institutionellen Datenpublikation ist die Übernahme von Kurationsarbeiten durch das KFD. Ziel ist es hierbei, Datensätze bereits während des Ingest-Prozesses so für das Repositorium aufzubereiten, dass sie soweit als möglich in einer Form vorliegen, die eine Langzeitarchivierung der Daten erlaubt. Dies umfasst die Überprüfung und die Validierung von Dateiformaten unter Einsatz gängiger Tools wie DROID und JHOVE, wobei man jedoch konstatieren muss, dass ein sinnvoller Einsatz dieser Systeme nur für eine verhältnismäßig geringe Zahl von Dateiformaten möglich ist.[17] Unter Umständen werden Datenbestände vor der Publikation in archivfähige Formate konvertiert.[18] Auch die Verbesserung der fachwissenschaftlichen Dokumentation des Datensatzes durch geeignete Metadaten und kontrollierte Vokabulare wird durch das KFD in enger Abstimmung mit den Forschenden unterstützt.

2.1. Erfahrungen nach drei Jahren heiDATA - lohnt sich der Aufwand?

Nach mittlerweile über drei Jahren Produktivbetrieb lässt sich zumindest ein erstes Zwischenfazit für den Betrieb von heiDATA ziehen, anhand dessen sich die grundsätzliche Frage nach der Sinnhaftigkeit von institutionellen Veröffentlichungsplattformen für Forschungsdaten zumindest in Ansätzen diskutieren lässt. Eine naheliegende Haltung hierzu ist, dass es sich bei Forschungsdaten – im Unterschied zu Textpublikationen – um deutlich heterogeneres und interpretationsbedürftigeres Material handelt. Und damit könnten sie zum einen nur durch Fachleute aus der jeweiligen Disziplin angemessen kuratiert werden und zum anderen nur in spezifisch zugeschnittenen Fachrepositorien wohldokumentiert und auffindbar publiziert werden. Diese Allgemeinplätze der Forschungsdatendiskussion sind sicher zutreffend.[19] Auch die bisherige, eher verhaltene Nutzung von heiDATA legt dies nahe. Bislang wurden auf der Plattform 105 Datensätze veröffentlicht (Stand: 23.4.2018). Vor dem Hintergrund, dass die Universität Heidelberg eine der forschungs- und publikationsstärksten Universitäten Europas mit aktuell ca. 5.800 Wissenschaftlerinnen und Wissenschaftlern ist, ist dies auf den ersten Blick keine besonders hohe Zahl. Zieht man den Vergleich zu institutionellen Datenpublikationsplattformen weiterer deutscher Universitäten, so stellt man fest, dass sich dort im Großen und Ganzen ein ähnliches Bild ergibt.[20]

17 DROID, https://digital-preservation.github.io/droid/; JHOVE, http://jhove.sourceforge.net/.

18 Beispielhaft sei hier auf einen Datensatz mit archäologischen Grabungsdaten verwiesen, bei dessen Publikation ca. 1.100 Bilddateien vom proprietären CorelDraw-Format in Vektorgrafiken konvertiert wurden (unter wiederholter Rückkopplung mit dem verantwortlichen Wissenschaftler, um Informationsverlust durch die Konversion möglichst ausschließen zu können): Paul A. Yule, „Pottery Drawings, Zafar, Jemen, Mostly Excavated," *heiDATA*, V3, 6. April 2017, https://doi.org/10.11588/data/10068.

19 Vgl. hierzu auch Louise Corti u. a., *Managing and Sharing Research Data: A Guide to Good Practice* (Los Angeles, Calif. [u.a.]: SAGE, 2014), 197–201.
Nebenbei bemerkt: In der Praxis muss man sich die vermeintlich geeigneteren fachlichen Repositorien freilich immer im Detail anschauen. Für große, lange und gute etablierte Dienste gilt dies sicherlich, aber recherchiert man z.B. im Kontext der Beratung von Forschenden quer durch re3data (https://www.re3data.org/), so stößt man auf etliche Fachrepositorien, die weder in Bezug auf Datendokumentation und Metadatenstandards besondere fachbezogene Tiefe bieten noch im Hinblick auf die nachhaltige Verfügbarkeit des Materials Vertrauen erwecken.

20 Beispielhaft sei hier auf die folgenden universitären Dienste verwiesen, die entweder eigenständige Forschungsdatenrepositorien oder kombinierte Repositorien für Texte, Forschungsdaten und weitere Materialien sind. Bei den kombinierten Repositorien kann man sich in der Regel über die Browsingfunktionalität den Veröffentlichungstyp „Forschungsdaten" oder verwandte Kategorien anzeigen lassen: LMU München, Open Data LMU, https://data.ub.uni-muenchen.de/; TU Berlin, DepositOnce, https://depositonce.tu-berlin.de/; Universität Bielefeld, PUB, https://pub.uni-bielefeld.de/data; Universität Freiburg, FreiDok plus, https://freidok.uni-freiburg.de/; Universität Mannheim, MADATA, https://madata.bib.uni-mannheim.de/; Universität Tübingen, FDAT, https://fdat.escience.uni-tuebingen.de/portal/#/start; MPG, EDMOND, https://edmond.mpdl.mpg.de/imeji/.

Was aber heißt dies? Sind und bleiben generische institutionelle Datenrepositorien ein Nischenangebot, das mittelfristig sogar wieder komplett verschwinden wird? Tatsächlich ist das Publizieren von Forschungsdaten in vielen Disziplinen noch Neuland und diejenigen Disziplinen, in denen dies nicht der Fall ist, verfügen über geeignete und gut genutzte Fachrepositorien. Umgekehrt sind es gerade diejenigen Disziplinen ohne „Open-Data-Kultur" in denen es noch keine fachspezifischen Publikationsplattformen für Forschungsdaten gibt, aber in denen doch zunehmend ein Wissenschaftlerinteresse an offenen Forschungsdaten entsteht. Hier füllen institutionelle Repositorien eine Leerstelle. Dies macht ein Blick auf die Verteilung der auf heiDATA publizierten Datensätze deutlich: Die Veröffentlichungen stammen aus neun der zwölf Fakultäten der Universität und aus vier zentralen wissenschaftlichen Einrichtungen und verteilen sich somit breit über die an der Universität Heidelberg vertretenen Disziplinen. Hinzu kommen Datensupplemente zu Textveröffentlichungen der Heidelberger Publikationsdienste. Veröffentlicht werden dabei ganz unterschiedliche Datentypen von biophysikalischen Simulationsdaten oder Umfragedaten aus dem Bereich der Public Health über Textcorpora aus der Computerlinguistik bis hin zu archäologischen Grabungsdaten oder Ergebnissen von Distant-Reading-Analysen theologischer Texte. Mehr als 9.000 Downloads seit dem Beginn des Betriebs sprechen dabei zudem dafür, dass die publizierten Datensätze durchaus in breiterem Ausmaß von Dritten genutzt oder zumindest gesichtet werden. Selbst wenn also die Nachfrage rein quantitativ (noch?) nicht exorbitant hoch ist, so kann man dennoch feststellen, dass sie sich breit über verschiedene Disziplinen verteilt. Genau an diesen sog. „long tail" richtet sich entsprechend das generische Angebot institutioneller Forschungsdatenrepositorien.[21]

Eine hieran anschließende Frage, die sich angesichts der oben dargestellten Erfahrungen stellt, ist, ob sich der Aufwand lohnt, eigene institutionelle Angebote zur Forschungsdatenveröffentlichung zu schaffen. Die Heidelberger Antwort auf diese Frage lautet: Unser Ziel ist es, dauerhaft Verantwortung für die Forschungsdaten der eigenen Institution zu übernehmen. Eine Datenpublikation auf heiDATA soll mehr sein als das bloße Abladen von Dateien auf einem Server und das Registrieren einer DOI für diese Dateien. Sie sollen vielmehr den FAIR-Data-Principles genügen.[22] Wenn es sich tatsächlich als richtig erweisen sollte, dass Forschungsdaten über sämtliche oder zumindest hinreichend viele Wissenschaftsdisziplinen hinweg so wertvolle Ressourcen sind, dass sie langzeitarchiviert und möglichst breit zugänglich gemacht werden sollten, dann müssen Betreiber von Forschungsdatenrepostorien notwendigerweise die Kuration der Daten und deren Langzeitarchivierung als ihre Aufgabe sehen. Bei zukünftig zunehmendem Bedarf müssen diese Aufgaben dann innerhalb der forschungsinfrastrukturellen Arbeitsteilung auf viele Schultern verteilt werden – in dieser Weise skizziert es auch der Rat für Informationsinfrastrukturen in seinen Überlegungen zur „Nationalen Forschungsdateninfrastruktur".[23]

21 Zur Rolle des „long tail" im Kontext wissenschaftlicher Daten vgl. P. Bryan Heidorn, „Shedding Light on the Dark Data in the Long Tail of Science," *Library Trends* 57, Nr. 2 (2008): 280–299, https://doi.org/10.1353/lib.0.0036; Jillian C. Wallis, Elizabeth Rolando und Christine L. Borgman, „If We Share Data, Will Anyone Use Them? Data Sharing and Reuse in the Long Tail of Science and Technology," *PLOS ONE* 8, Nr. 7 (Juli 2013): e67332, https://doi.org/10.1371/journal.pone.0067332.

22 Mark D. Wilkinson u. a., „The FAIR Guiding Principles for Scientific Data Management and Stewardship," *Scientific Data*, 15. März 2016, https://doi.org/10.1038/sdata.2016.18.

23 Vgl. RfII – Rat für Informationsinfrastrukturen, *Leistung aus Vielfalt. Empfehlungen zu Strukturen, Prozessen und Finanzierung des Forschungsdatenmanagements in Deutschland* (Göttingen, 2016), zuletzt geprüft am 22.05.2018, http://www.rfii.de/?wpdmdl=1998.

Wie diese Arbeitsteilung im Detail aussehen wird, ist dabei weitgehend offen. Es ist keineswegs ausgemacht, dass es über alle Disziplinen hinweg fachliche Datenarchive geben wird (ebenso wenig wie es zum aktuellen Zeitpunkt überhaupt ausgemacht ist, ob es den Bedarf dafür wirklich übergreifend gibt). Im Gegenteil, die Wahrscheinlichkeit, dass hierbei auch institutionelle Dienste eine wichtige Rolle spielen werden, ist durchaus hoch. Dies gilt umso mehr, als man auch im Hinblick auf die Nachhaltigkeit verschiedener fachlicher Repositorien durchaus kritisch sein darf. Insbesondere in den Fächern, für die das Publizieren von Forschungsdaten noch neu ist, werden aktuell auf nationaler und internationaler Ebene mit erfreulich hoher Drittmittelunterstützung etliche Datenrepositorien eingerichtet. Welche von diesen sich aber tatsächlich im Fach etablieren und den Sprung in einen nachhaltigen Regelbetrieb inklusiver gesicherter Finanzierung schaffen werden, ist unklar.

Weiterhin gibt es auch von Seiten der Forschenden durchaus Nachfrage nach institutionellen Angeboten. Lokale FDM-Serviceeinheiten sind nah an den Wissenschaftlerinnen und Wissenschaftlern, Absprachen und persönliche Treffen sind kurzfristig und mit wenig Aufwand zu realisieren, man kennt sich aus anderen Kontexten und arbeitet daher vertrauensvoll zusammen, kleinere und mittlere Datenmengen können ggf. kostenfrei publiziert werden etc. Lokale Repositorien bieten den Datengebern mit geringem Aufwand ein Schaufenster für die Datenprodukte des eigenen Instituts, der eigenen Forschungsgruppe oder anderer Teileinheiten; insbesondere die oben dargestellte Untergliederung von heiDATA in unterschiedliche Dataverses ist hierbei hilfreich.[24]

2.2. Repositorium oder Repositorien?

heiDATA ist ein eigenständiges Forschungsdatenrepositorium, das vor Ort parallel zu vier EPrints-Instanzen (universitärer Dokumentenserver heiDOK plus drei Fachrepositorien für die FIDs), dem mit der Software easydb betriebenen Bild- und Multimediarepositorium heidICON sowie weiteren OJS- und OMP-basierten Publikationsplattformen betrieben wird. Damit stellt sich auch die Frage, ob der Betrieb einer solchen Repositorienlandschaft nicht deutlich mehr Aufwand erzeugt als ein gemeinsames Repositorium für Texte, Forschungsdaten und ggf. weitere Materialien. Und tatsächlich sind entsprechende Lösungen verbreitet, teilweise mit Integration weiterer Funktionalitäten für Universitätsbibliographien oder Forschungsinformationssysteme. Beispielhaft sei hier auf die Systeme PUB an der Uni Bielefeld[25], DepositOnce der TU Berlin[26] oder FreiDok plus an der Universität Freiburg[27] verwiesen.

Dass heiDATA als eigenständiges System betrieben wird, liegt primär in einer pragmatischen Einschätzung begründet: Andere Lösungen wären im Hinblick auf Administration und Betrieb (vermutlich) deutlich arbeitsintensiver gewesen. Nachdem 2014 die Entscheidung zum Angebot eines Services zur Datenpublikation gefallen war, wurden hierfür die Systeme EPrints, DSpace, ESciDoc und Dataverse gesichtet. Eine zentrale Anforderung an die Software war dabei die Möglichkeit der schnellen Implementierung eines Services mit wenig Anpassungsbedarf, z.B. beim Datenschema. Zum damaligen

24 Beispiele: https://heidata.uni-heidelberg.de/dataverse/awiexeco oder https://heidata.uni-heidelberg.de/dataverse/asiaeurope.
25 https://pub.uni-bielefeld.de/
26 https://depositonce.tu-berlin.de/
27 https://freidok.uni-freiburg.de/

Zeitpunkt erfüllte Dataverse insbesondere dieses Kriterium am besten. Darüber hinaus überzeugt Dataverse im laufenden Betrieb bislang auch durch umfassende Funktionalitäten, eine aufgeräumte und strukturierte Benutzeroberfläche, wenig administrativen Aufwand und stabilen Betrieb sowie eine aktive, schnelle und an den Wünschen der Nutzercommunity ausgerichtete Entwicklungsarbeit.

Der tatsächliche Arbeitsaufwand für die Erstinstallation lässt sich dabei grob mit ca. zwei Personenwochen beziffern, der in 2017 durchgeführte größere Versionssprung auf Dataverse Version 4.x inklusive Migration der publizierten Datenbestände mit ca. zwei bis drei Wochen. Die für den laufenden Betrieb anfallenden systemadminstrativen Aufgaben haben aber einen sehr geringen Umfang. Der Hauptaufwand entsteht im Kontext der Beratung der Forschenden und im Zuge der Kurationsaktivitäten beim Daten-Ingest. Und dieser Aufwand entsteht ohnehin plattformunabhängig und auch unabhängig davon, ob man ein eigenständiges Forschungsdatenrepositorium oder ein integriertes System betreibt.

2.3. Und wenn es sich anders entwickelt? Exitstrategie

Wie im bisherigen Verlauf des Beitrags beschrieben, ist nicht klar, wie sich das Thema Open Research Data abseits der ohnehin datenpublizierenden Disziplinen entwickeln wird. Interesse und breite Akzeptanz auf Seiten der Wissenschaftler müssen sich noch entwickeln, und auch ob tatsächlich in substantiellem Maß eine Nachnutzung von publizierten Forschungsdaten stattfindet, welche die Versprechen des „Data Web“[28] und der allgegenwärtigen Data-Science- und Big-Data-Rhetorik einlöst, muss sich noch erweisen. Damit einhergehend ist auch offen, wie sich die Landschaft der fachlichen und institutionellen Datenrepositorien zukünftig entwickeln wird.

Die Heidelberger Dataverse-Instanz wurde und wird daher als Übergangslösung betrachtet, deren Eignung vor dem Hintergrund des sich entwickelnden Felds des Forschungsdatenmanagements kontinuierlich neu bewertet werden muss. Sämtliche organisatorischen und technischen Festlegungen sind so gestaltet, dass die Datenbestände perspektivisch in andere Systeme, seien es übergreifende Dienste oder neue institutionelle Lösungen, überführt werden können. Diese Überlegung war ebenfalls ein wesentliches Kriterium bei der Entscheidung für Dataverse als Softwarebasis für heiDATA: Die Anwendung ist kostenfrei, open source und vor allem ist der Export von Daten und Metadaten in strukturierter Form problemlos möglich, sodass bei einer entsprechenden Neubewertung die Überführung des in heiDATA publizierten Materials in fachspezifische Dienste oder übergreifende Lösungen für Institutionen wie RADAR oder KITopenData (vormals bwDataDiss) prinzipiell möglich ist.[29]

28 Michael Nielsen, *Reinventing Discovery: The New Era of Networked Science* (Princeton: Princeton University Press, 2012), 110–116.

29 RADAR und KITopenData sind zentrale, nicht-kommerzielle Repositoriendienste, die von Institutionen kostenpflichtig in Anspruch genommen werden können, um ihren Forschenden eine Möglichkeit zur Datenpublikation zu bieten. Vgl. RADAR, https://www.radar-service.eu; KITopenData, https://bwdatadiss.kit.edu/.

3. heidICON oder wie zählt man eigentlich Datenpublikationen?

Für ein vollständiges Bild der Heidelberger Publikationsangebote für Forschungsdaten lohnt es sich zusätzlich zu heiDATA auch das seit 2005 an der Universität Heidelberg betriebene Bild- und Multimediaarchiv heidICON in den Blick zu nehmen.[30] heidICON basiert auf der Software easyDB, einem Media-Asset-Management-System der Firma Programmfabrik und bietet umfangreiche Funktionalitäten zur Verwaltung, Bereitstellung und Erschließung von AV-Daten in offenen und geschlossenen Medienpools.[31] Für die Erschließung wird dabei das normierte Vokabular der GND sowie das Metadatenschema LIDO eingesetzt.[32]

In heidICON sind gegenwärtig über 510.000 Objekte abgelegt (Stand: 23.4.2018), die im Schnitt auf ca. 260.000 Einzelobjektaufrufe pro Jahr kommen. Dies ist im Kontext dieses Aufsatzes ein interessantes Faktum, da die in heidICON publizierten AV-Materialen gemäß der einschlägigen Definitionen selbstverständlich auch unter das Lemma „Forschungsdaten" fallen und man somit mit guten Gründen sagen kann, dass in Heidelberg (an vielen anderen Universitäten verhält es sich ähnlich) seit 2005 ein Forschungsdatenrepositorium betrieben wird, in dem bereits über eine halbe Million Forschungsdatensätze publiziert wurden. Setzt man dies zu den oben erwähnten 105 Datensätzen auf heiDATA in Beziehung, wird eine weitere Schwierigkeit bei der quantitativen Beurteilung von Forschungsdatenrepositorien deutlich: Es ist eben nicht ohne weiteres klar, wie man Forschungsdaten zählt. Was ist das Verhältnis von Einzeldateien (und den ihnen zugehörigen Metadaten) zu einem Gesamtdatensatz (und ihm zugehörigen Metadaten)? Wann bildet eine Menge von Dateien gemeinsam einen Datensatz, wann eher eine Kollektion oder einen Pool von Datensätzen? Usw.

4. CS-FDP – ein Projekt zur Entwicklung fach- und communityspezifischer Datenportale

Insbesondere für umfangreichere Datensammlungen ist das Angebot einer individuell zugeschnittenen Präsentationsoberfläche eine Anforderung, die häufig von Seiten der Forschenden formuliert wird. heiDATA bietet dies nicht, sondern stellt letztendlich nur eine Downloadoption für die Daten bereit. Gleichzeitig gibt es in vielen Fällen für solche Datensammlungen keine geeigneten Fachrepositorien, die eine ansprechende Präsentation der Daten ermöglichen würden. Dies führt dazu, dass Forschungsprojekte vielfach projekteigene Datenbanken mit Webpräsentation aufsetzen, deren dauerhafter Betrieb über die Projektlaufzeit hinaus jedoch nicht gesichert ist. Diese Beobachtung bildet die Ausgangslage für die gegenwärtige Weiterentwicklung der Forschungsdatenservices des KFD, die im Rahmen des Projekts „Community-spezifische Forschungsdatenpublikation" (CS-FDP) vorangetrieben wird.[33] CS-FDP wird gefördert vom baden-württembergischen Ministerium für Wissenschaft, Forschung und Kunst (MWK) und ist gemeinsam mit weiteren Projekten aus den Bereichen

30 heidICON, http://heidicon.ub.uni-heidelberg.de.
31 Programmfabrik, https://www.programmfabrik.de/
32 „LIDO," ICOM International Committee for Documentation, zuletzt geprüft am 22.05.2018, http://www.lido-schema.org/.
33 „Community-spezifische Forschungsdatenpublikation," Universitätsrechenzentrum Heidelberg, zuletzt geprüft am 22.05.2018, https://www.urz.uni-heidelberg.de/de/cs-fdp.

Forschungsdatenmanagement und Virtuelle Forschungsumgebungen Teil der E-Science-Initiative des Landes Baden-Württemberg.[34]

Ziel von CS-FDP ist der Aufbau eines überschaubaren Pools von generischen, durch das KFD bereitgestellten Softwarewerkzeugen zur Erstellung dynamischer Forschungsdatenportale, mit denen eine Vielzahl der bei unterschiedlichen Projekten auftretende Anforderungen erfüllt werden können, die sich aber gleichzeitig mit vertretbarem Aufwand zentral administrieren und archivieren lassen. Darüber hinaus ist es Ziel von CS-FDP, ein tragfähiges Konzept für die Langzeitarchivierung der universitären Forschungsdatenbestände zu entwickeln. CS-FDP soll damit die Lücke zwischen generischer, universitärer FDM-Infrastruktur und individuellen, fachcommunity-spezifischen Anforderungen verkleinern. Das KFD kooperiert mit fünf Pilotpartnern, mit denen unterschiedliche technische Lösungen prototypisch umgesetzt und auf ihre Eignung für ein breiteres Serviceangebot geprüft werden.

Im ersten Pilotprojekt wird in Zusammenarbeit mit dem Heidelberger Institut für Semitistik eine zukunftsfähige Lösung für das am Institut mit DFG-Unterstützung aufgebaute „Semitische Tonarchiv" erarbeitet, das für die semitistische Forschung ein zentrales Angebot an Audiodaten bereitstellt.[35] Die dem bisherigen Institutsarchiv zugrundeliegende Software ist mittlerweile stark veraltet und eine adäquate technische Betreuung kann von Seiten des Instituts nicht mehr gewährleistet werden. Im Rahmen von CS-FDP werden die ca. 3.000 dort archivierten Tondokumente nach heidICON überführt. Ziel ist es, heidICON als Backend für die Daten zu verwenden und darauf aufsetzend eine eigenständige Rechercheoberfläche anzubieten, die unabhängig von heidICON ist. Diese Oberfläche soll spezifische Darstellungs- und Rechercheoptionen realisieren, die die semitistische Forschungscommunity zur Arbeit mit den Daten benötigt (z.B. Geovisualisierung und -browsing). Realisiert wird das Frontend mit der Software Vue.js unter Nutzung der easyDB-API.[36]

Im Rahmen des zweiten Pilotprojekts wird ebenfalls die Software easyDB eingesetzt, allerdings für eine reine Metadatenverwaltung (ohne direkten Bezug zu AV-Medien). Mit easyDB können komplexe Datenmodelle schnell, flexibel und ohne Programmieraufwand abgebildet werden. Ein ausgefeiltes Rechte- und Tag-Management ist dabei inbegriffen. Ein Anwendungsfall hierfür wird gemeinsam mit dem DFG-Projekt „Dynastinnen und Bettelorden im spätmittelalterlichen Reich. Weibliche Frömmigkeit zwischen Hof, Stadt und Kloster (1250–1400)", kurz „Gender & Piety" erprobt. Hier geht es darum, komplexe soziale und räumliche Beziehungen in historischen Netzwerken abzubilden.

Im Rahmen des dritten Pilotprojekts „Digitale Papyrussammlung" wird der Aufbau individueller, webbasierter Rechercheoberflächen mit Hilfe der Software Django[37] sowie deren Verknüpfung mit standardisierten Metadaten - in diesem Fall Publikationsmetadaten gemäß des TEI/EpiDoc-Standards

34 Allgemeine Informationen zur E-Science-Initative: „E-Science: Wissenschaft unter neuen Rahmenbedingungen," Ministerium für Wissenschaft, Forschung und Kunst, zuletzt geprüft am 22.05.208, https://mwk.baden-wuerttemberg.de/de/forschung/forschungslandschaft/e-science/; Projektübersicht: https://www.forschungsdaten.info/praxis-kompakt/aktuelle-infrastrukturprojekte/.

35 www.semarch.uni-hd.de/.

36 Vue.js, https://vuejs.org/.

37 Django, https://www.djangoproject.com/.

– entwickelt.[38] Die Präsentation der digitalen Faksimiles sowie der Transkriptionen erfolgt über DWork, dem von der Universitätsbibliothek Heidelberg entwickelten Workflowmanagement- und Präsentationssystem für Retrodigitalisate, dessen Editionsmodul zu diesem Zweck erweitert wurde.[39] Insbesondere die Visualisierung und Normalisierungen von editorischen Eingriffen unter Berücksichtigung von fachspezifischen Eigenheiten wurden überarbeitet und deutlich stärker konfigurierbar gestaltet (sowohl für den Editierenden als auch für den Leser). Diese Entwicklungen können bereits jetzt im Editionsprojekt „Documenta Nepalica" der Heidelberger Akademie der Wissenschaften nachgenutzt werden, in dem ein umfangreiches Korpus von Dokumenten und Texten zur Religions- und Rechtsgeschichte des vormodernen Nepal digital bereitgestellt wird.[40]

Für ein viertes, lebenswissenschaftliches Pilotprojekt zur Zellbiologie wurde die freie Mikroskopiesoftware OMERO in der vom Universitätsrechenzentrum betriebenen Cloudinfrastruktur heiCLOUD eingerichtet.[41] Derzeit wird sie vom Institut für Anatomie und Zellbiologie für die Analyse und das Management von Mikroskopiedaten genutzt. Zuletzt erfolgte die Anbindung großer Speicherressourcen auf der Large Scale Data Facility mit Hilfe des Dienstes SDS@hd.[42] Dies ist ein entscheidender Schritt, um die anfallenden Datenmengen im Multi-Terabyte-Bereich in OMERO importieren, analysieren und publizieren zu können. Im weiteren Verlauf des Projektes ist geplant, eine Präsentationsoberfläche für die Publikation von Mikroskopiedaten basierend auf OMERO zu erstellen.

Der Forschungsbereich „Biostratigraphie und Paläoökologie" am Institut für Geowissenschaften betreibt derzeit eine lokale Instanz der Software Specify.[43] Specify dient der Erfassung, Taxierung und digitalen Präsentation geologischer und biologischer Sammlungen. Eine zentral gehostete Instanz wurde in der heiCLOUD installiert.[44] Derzeit wird im Rahmen des fünften Pilotprojekts die Migration der lokalen Instanz des Instituts für Geowissenschaften in die zentrale Instanz vorbereitet. Anhand der gewonnenen Erfahrungen aus dieser Pilotanwendung soll das Portfolio um Specify-Unterstützung für weitere Anwender erweitert werden.

5. Zusammenfassung

Die Universität Heidelberg treibt den Ausbau universitärer Infrastrukturen für offene Forschungsdaten voran. Das Kompetenzzentrum Forschungsdaten als zuständige Serviceeinrichtung betreibt bereits seit drei Jahren mit heiDATA ein institutionelles Forschungsdatenrepositorium auf der Basis der Software Dataverse. Hinzu kommt das bereits seit 2005 in Betrieb befindliche Repositorium heidICON für AV-Daten sowie ein im Rahmen des MWK-geförderten Projekts CS-FDP im Aufbau

38 „EpiDoc: Epigraphic Documents in TEI XML," Sourceforge, zuletzt geprüft am 22.05.2018, https://sourceforge.net/p/epidoc/wiki/Home/.
39 DWork – Heidelberger Digitalisierungsworkflow, http://www.ub.uni-heidelberg.de/helios/digi/dwork.html.
40 „Religions- und rechtsgeschichtliche Quellen des vormodernen Nepal," Heidelberger Akademie der Wissenschaften, zuletzt geprüft am 22.05.2018, http://www.haw.uni-heidelberg.de/forschung/forschungsstellen/nepal/publ_docs.de.html.
41 OMERO, https://www.openmicroscopy.org/omero/.
42 SDS@hd – Scientific Data Storage, https://www.urz.uni-heidelberg.de/de/sds-hd.
43 Specify, http://www.sustain.specifysoftware.org/.
44 heiCLOUD, https://heicloud.uni-heidelberg.de/heiCLOUD.

befindlicher Service für die Erstellung und Betreuung individueller Forschungsdatenportale. Nachfrage und Nutzung der Angebote dokumentieren zwar eine zunehmende Bedeutung des Themas Open Research Data sowie die Relevanz institutioneller Datenpublikationsplattformen, zeigen aber gleichzeitig die unterschiedliche und sich im Fluss befindende Ausgangssituation in verschiedenen Fachdisziplinen auf, sodass die aufgebaute technische Infrastruktur kontinuierlich auf ihre Eignung und Bedarfsangemessenheit geprüft werden muss.

Literaturverzeichnis

- Corti, Louise, Veerle Van den Eynden, Libby Bishop und Matthew Woollard. *Managing and Sharing Research Data: A Guide to Good Practice*. Los Angeles, Calif. [u.a.]: SAGE, 2014.

- Heidorn, P. Bryan. „Shedding Light on the Dark Data in the Long Tail of Science." *Library Trends* 57, Nr. 2 (2008): 280–299. https://doi.org/10.1353/lib.0.0036.

- Nielsen, Michael. *Reinventing Discovery: The New Era of Networked Science*. Princeton: Princeton University Press, 2012.

- RfII – Rat für Informationsinfrastrukturen. *Leistung aus Vielfalt. Empfehlungen zu Strukturen, Prozessen und Finanzierung des Forschungsdatenmanagements in Deutschland*. Göttingen, 2016. Zuletzt geprüft am 22.05.2018. http://www.rfii.de/?wpdmdl=1998.

- Wallis, Jillian C., Elizabeth Rolando und Christine L. Borgman. „If We Share Data, Will Anyone Use Them? Data Sharing and Reuse in the Long Tail of Science and Technology"l. *PLOS ONE* 8, Nr. 7 (Juli 2013): e67332. https://doi.org/10.1371/journal.pone.0067332.

- Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, u. a. „The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Scientific Data*, 15. März 2016. https://doi.org/10.1038/sdata.2016.18.

- Yule, Paul A. „Pottery Drawings, Zafar, Jemen, Mostly Excavated." *heiDATA*, V3, 6. April 2017. https://doi.org/10.11588/data/10068.