Ein blinder Fleck in der FDM-Versorgungslandschaft?

Dark und Cold Archiving Services am Data Center for the Humanities

Patrick Helling und Felix Rau

Abstract

Ein zentrales Ergebnis von Forschungsdatenmanagement (FDM) ist die Publikation und Bereitstellung von Forschungsdaten im Sinne der FAIR-Prinzipien durch Repositorien. Allerdings können aufgrund unterschiedlicher Schranken und Vorgaben nicht immer alle Forschungsergebnisse ohne weiteres offen zugänglich gemacht werden. Dennoch gilt es auch, solche Ergebnisse langfristig zu sichern. Mit diesem Beitrag stellen wir den Archivierungsworkflow des Data Center for the Humanities (DCH) der Universität zu Köln aus einer technischen und organisatorischen Perspektive vor. Er ermöglicht einen niederschwelligen Service zur langfristigen Datensicherung und zum Datennachweis und füllt eine Leerstelle in der Forschungsdatenmanagement-Servicelandschaft für den Umgang mit nicht publizierbaren Forschungsdaten und -ergebnissen.

One central outcome of research data management (RDM) is the publication and accessibility of research data in compliance with the FAIR Principles through repositories. Nevertheless, there may be limitations that hinder the publication of research data. However, it is essential to store this research data for the long term. In this article, we present the archiving workflow of the Data Center for the Humanities (DCH) at the University of Cologne from both technical and organizational perspectives. This workflow enables us to provide a user-friendly service for archiving research data. Through this service, which allows for the management of research data that cannot be published on one hand and should be handled in accordance with the FAIR Principles on the other, the DCH addresses a gap within the RDM service landscape.

1 Einleitung und Problemstellung

„Grundsätzlich bringen Wissenschaftlerinnen und Wissenschaftler alle Ergebnisse in den wissenschaftlichen Diskurs ein. Im Einzelfall kann es aber Gründe geben, Ergebnisse nicht öffentlich zugänglich (im engeren Sinne in Form von Publikationen, aber auch im weiteren Sinne über andere Kommunikationswege) zu machen."[1]

Die erste Aussage des vorangestellten Zitats aus Leitlinie 13 (*Herstellung von öffentlichem Zugang zu Forschungsergebnissen*) des Kodex Leitlinien zur Sicherung guter wissenschaftlicher Praxis[2] der Deutschen Forschungsgemeinschaft (DFG)[3] repräsentiert eines der zentralen Ziele des Managements von Forschungsdaten: Forschung als selbstreferentielles System lebt insbesondere davon, dass Forschungsergebnisse veröffentlicht und nachgenutzt werden, um darauf aufbauend neue Forschungsfragen zu entwickeln und zu beantworten. Insbesondere digitale Forschungsdaten stellen in diesem Zusammenhang einen Motor für wissenschaftlichen Fortschritt dar.[4] Die standortunabhängige Verfügbarkeit von dokumentierten, digitalen Forschungsergebnissen durch das Internet ermöglicht ihre Nutzung über Institutions- und Ländergrenzen hinweg und befördert die Kooperation und den Austausch zwischen Forschenden weltweit sowie die Reproduzierbarkeit eben jener Forschungsergebnisse. Entsprechend sollte die Publikation von Forschungsdaten und -ergebnissen am Ende jedes Forschungsprozesses stehen.

Die Realisierung dieses Ziels ist unter anderem in den FAIR-Prinzipien[5] als zentrale Empfehlungen im Forschungsdatenmanagement (FDM) beschrieben. FAIR definiert Maßnahmen zur Sicherstellung der Auffindbarkeit (Findability), Zugänglichkeit (Accessibility), Interoperabilität (Interoperability) und Nachnutzbarkeit (Re-Usability) von Forschungsdaten und -ergebnissen. Generische und fach- oder domänenspezifische Repositorien, die bspw. von Infrastruktureinrichtungen oder Forschungs- und Datenzentren betrieben werden, stellen in diesem Zusammenhang einen zentralen Service dar, um Forschungsdaten und -ergebnisse langfristig für andere Forschende nachnutzbar bereitzustellen.[6] Sie verfügen in der Regel über Dokumentations- und Metadatenstandards sowie Kurationspolicies und idealerweise über eine entsprechende personelle, infrastrukturelle und finanzielle Aus-

1 Deutsche Forschungsgemeinschaft. „Guidelines for Safeguarding Good Research Practice. Code of Conduct". (2022) S. 18. doi: https://doi.org/10.5281/ZENODO.6472827.

2 Deutsche Forschungsgemeinschaft. „Guidelines for Safeguarding Good Research Practice. Code of Conduct". (2022). doi: https://doi.org/10.5281/ZENODO.6472827.

3 https://www.dfg.de/ (07. November 2023).

4 Bryant, Rebecca/ Lavoie, Brian/ Malpas, Constance: „The Realities of Research Data Management Part One: A Tour of the Research Data Management (RDM) Service Space". in: OCLC Research (2017). doi: https://doi.org/10.25333/C3PG8J.

5 Wilkinson, Mark D./ Dumontier, Michel/ Aalbersberg, IJsbrand Jan/ Appleton, Gabrielle/ Axton, Myles/ Baak, Arie/ Blomberg, Niklas et al.: „The FAIR Guiding Principles for Scientific Data Management and Stewardship". in: Scientific Data 3 (1): 160018 (2016). doi: https://doi.org/10.1038/sdata.2016.18.

6 Mathiak, Brigitte/ Metzmacher, Katja/ Helling, Patrick/ Blumtritt, Jonathan: The Role Of Data Archives In The Humanities At The University Of Cologne, in: N.N.: DH2019 Book of Abstracts. Utrecht 2019. doi: https://doi.org/10.34894/GEQEKO.

stattung, mit der eine dauerhafte Publikation und Verfügbarkeit von Forschungsdaten und -ergebnissen gewährleistet werden kann.

Neben dieser Idealvorstellung von Wissenschaft und Forschung, der sowohl intrinsische als auch extrinsische Motivationsfaktoren zugrunde liegen, adressiert der zweite Teil des vorangestellten Zitats hingegen die Tatsache, dass es auch gute Gründe geben kann, die (insbesondere in den Geistes- und Sozialwissenschaften) der Veröffentlichung und Nachnutzung von Forschungsdaten und -ergebnissen entgegenstehen. So gibt es zwar bspw. innerhalb der digitalen Literaturwissenschaften bereits Ansätze zur Veröffentlichung von Forschungsergebnissen, die aus der Analyse urheberrechtlich geschützter Texte entstanden sind,[7] allerdings können urheberrechtliche Bestimmungen die Veröffentlichung eines literarischen Korpus (oder sogar der Analyseergebnisse) verhindern.

Gleichzeitig können es Anonymisierungs- und Pseudonymisierungsprozesse sowie informierte Einverständniserklärungen, wie sie bspw. in den Sozialwissenschaften, der Bildungsforschung oder auch in der Ethnologie angewandt werden, ermöglichen personenbezogene oder personenbeziehbare Daten, die im Rahmen von Forschungsvorhaben erhoben und verarbeitet wurden, zu veröffentlichen und anderen Forschenden zur Nachnutzung zugänglich zu machen. Allerdings steht auch häufig der Datenschutz, in diesem Zusammenhang vor allem die Datenschutzgrundverordnung (DSGVO),[8] der Veröffentlichung von personenbezogenen oder personenbeziehbaren Daten entgegen. Zwar sind insbesondere gut anonymisierte, personenbezogene Daten von den Regelungen der Datenschutzgrundverordnung ausgeschlossen, allerdings können nicht alle erhobenen, personenbezogenen Daten immer auch sinnvoll anonymisiert werden. Neben der Tatsache, dass bspw. die Anonymisierung von Videoaufzeichnungen von Schülerinnen und Schülern im Rahmen eines Forschungsvorhabens aus der Bildungsforschung ein aufwendiger Prozess ist, gilt häufig, dass solche Daten in der Regel, um sie veröffentlichen zu können, auf eine Art und Weise anonymisiert werden müssten, die ihre Nachnutzbarkeit nicht mehr gewährleisten kann.

Während es bereits viele generische und fach- oder domänenspezifische Infrastrukturangebote wie bspw. Repositorien gibt, die institutions- und länderabhängig genutzt werden können und die Ablage und Publikation von Forschungsdaten und -ergebnissen ermöglichen, scheint es reine Archivierungsservices in der Breite noch nicht zu geben. Um jedoch Empfehlungen wie den FAIR-Prinzipien und Guidelines wie dem Kodex Leitlinien zur Sicherung guter wissenschaftlicher Praxis der DFG vollständig zu entsprechen, bedarf es auch entsprechender Angebote für die langfristige Ablage von Forschungsdaten und -ergebnissen, die nicht auffindbar und zugänglich sein können, dennoch aber zum Zweck der Nachvollziehbarkeit und Reproduzierbarkeit sowie zur Vermeidung von Datenverlusten langfristig gesichert werden sollten.

Dabei ist ein Bedarf an solchen infrastrukturellen Angeboten insbesondere (allerdings nicht ausschließlich) in Fächern der Geistes- und Sozialwissenschaften zu beobachten. Qualitative und quantitative Umfragen, Aufnahmen von kulturellem Verhalten, persönlicher Dialoge oder Schul- und Lehrveranstaltungen (insbesondere mit Minderjährigen), die Digitalisierung und Analyse von (urheberrechtlich geschützten) Texten oder die Sammlung von Daten und Informationen über politische oder soziale Phänomene stellen grundsätzliche geistes- und sozialwissenschaftliche Methoden dar. Entsprechend sind der verantwortungsvolle Umgang mit diesen (sensiblen) Daten auf der einen Seite und die möglichst vollumfängliche Umsetzung von Empfehlungen wie den FAIR-Prinzipien auf der anderen Seite zentrale Herausforderungen für Datenzentren und Data-Manager in diesem Fachbereich.

In diesem Beitrag stellen wir den Archivierungsworkflow des Data Center for the Humanities (DCH),[9] einem geisteswissenschaftlich-fachspezifischen Datenzentrum an der Philosophischen Fakultät[10] der Universität zu Köln,[11] vor. Dieser Workflow ermöglicht es, Forschungsdaten und -ergebnisse auf institutionell betriebenem Bandspeicher zu archivieren. Dieser Archivierungsservice unterteilt sich in zwei Varianten: Während in der Cold-Archiving-Variante die Forschungsdaten und -ergebnisse durch die Publikation der dazugehörigen Metadaten zumindest auffindbar sind, werden sie in der Dark-Archiving-Variante lediglich archiviert, ohne dass Dritte sie finden können. Der Archivierungsservice des DCH wird aktuell für Forschungsdaten und -ergebnisse angeboten, die nicht über eine andere Infrastruktur veröffentlicht werden können, sowie zur zusätzlichen Sicherung und Wahrung einer Ausfallsicherheit. Innerhalb des Workflows werden alle Datentypen und -formate übernommen, die von Forschenden an der Philosophischen Fakultät der Universität zu Köln erzeugt wurden.

7 Schöch, Christof/ Döhl, Frédéric/ Rettinger, Achim/ Gius, Evelyn/ Trilcke, Peer/ Leinen, Peter/ Jannidis, Fotis/ Hinzmann, Maria/ Röpke, Jörg: „Abgeleitete Textformate: Text und Data Mining mit urheberrechtlich geschützten Textbeständen". in: Zeitschrift für digitale Geisteswissenschaften, Heft 5 (2020). doi: http://dx.doi.org/10.17175/2020_006.

8 https://dsgvo-gesetz.de/ (07. November 2023).

9 https://dch.phil-fak.uni-koeln.de/ (07. November 2023).

10 https://phil-fak.uni-koeln.de/ (07. November 2023).

11 https://www.uni-koeln.de/ (07. November 2023).

2 Infrastrukturelle (Ausgangs-)situation an der Universität zu Köln

Die Universität zu Köln gehört zu den forschungsstärksten Wissenschaftsstandorten in Deutschland. Mit gut 3.900 wissenschaftlichen Mitarbeiterinnen und Mitarbeitern und 621 Professorinnen und Professoren (Stand 2021) gehört sie auch personell zu den größten Forschungseinrichtungen in der Bundesrepublik.[12] Seit 2018 verfügt die Universität mit dem Cologne Competence Center for Research Data Management (C³RDM)[13] über eine zentrale Einrichtung, die Forschende an der Universität zu generischen Fragen zum Management von Forschungsdaten berät. Das C³RDM setzt sich aus Vertreterinnen und Vertretern des Regionalen Rechenzentrums (RRZK),[14] der Universitäts- und Stadtbibliothek (USB)[15] sowie dem Dezernat für Forschungsmanagement (Dezernat 7)[16] zusammen.
Insbesondere das RRZK aber auch die USB der Universität zu Köln stellen in diesem Zusammenhang zentrale Basisinfrastruktur wie bspw. virtuelle Maschinen und Arbeitsumgebungen, Speicherinfrastrukturen aber auch Backup- und Archivierungsinfrastrukturen für Forschende an der Universität zur Verfügung.[17]

2.1 Das Data Center for the Humanities (DCH) an der Philosophischen Fakultät

Das Data Center for the Humanities (DCH) berät und unterstützt als geisteswissenschaftlich-fachspezifisches Datenzentrum bereits seit 2013 Forschende an der Philosophischen Fakultät der Universität zu Köln bei Fragen zum Management von Forschungsdaten.[18] Das Zentrum ist am Dekanat der Fakultät angesiedelt und besteht aus einer dreiköpfigen Leitungsebene[19] und zwei ½ Vollzeitäquivalenten. Das DCH führt regelmäßig Drittmittelprojekte durch, mit denen Versorgungslücken in der FDM-Versorgungslandschaft adressiert und das Management von Forschungsdaten selbst beforscht wird[20] und ist an allen vier geisteswissenschaftlichen Konsortien der Nationalen Forschungsdateninfrastruktur (NFDI)[21] – NFDI4Culture,[22] Text+,[23] NFDI4Memory[24] und NFDI4Objects[25] – beteiligt. Entsprechend wird das DCH-Team durch mehrere Projektmitarbeiterinnen und -mitarbeiter sowie Hilfskräfte komplettiert.[26] Als beratende Instanz verfügt das DCH darüber hinaus über einen Beirat,[27] der sich aus Vertreterinnen und Vertretern unterschiedlicher lokaler, regionaler und nationaler Infrastruktureinrichtungen und Partnerinstitutionen zusammensetzt.
Zum Kerngeschäft des Datenzentrums gehört die persönliche Beratung und langfristige Unterstützung von Forschenden bei Fragen zum Forschungsdatenmanagement. Dabei unterscheidet das DCH grob in vier verschiedene Beratungskategorien:

- Allgemeine FDM-Beratungen,
- Antragsberatungen/-beteiligungen *(ab ovo)*,
- Begleitendes Forschungsdatenmanagement *(in vitae)*, bspw. in laufenden Großprojekten,
- Hilfestellung bei endenden bzw. abgeschlossenen Projekten (*post mortem*).

Das Datenzentrum berät und unterstützt über den gesamten Forschungsdatenlebenszyklus hinweg. Der Beratungsprozess[28] des DCH ist formalisiert und mündet in strukturierte Beratungsprotokolle, die nicht nur Beratungsvorgänge dokumentieren, sondern auch die Basis für die Analyse und Auswertung der FDM-Bedarfslandschaft an der Philosophischen Fakultät darstellen.[29]
Das Zentrum übernimmt im Rahmen seiner Möglichkeiten dauerhaft digitale Ressourcen und Anwendungen und verfügt über einen dezidierten Serviceschwerpunkt

12 https://strategy.uni-koeln.de/strategisches_controlling__informationsmanagement/zahlen_i_daten_i_fakten/index_ger.html (07. November 2023).
13 https://fdm.uni-koeln.de/home (07. November 2023).
14 https://rrzk.uni-koeln.de/ (07. November 2023).
15 https://ub.uni-koeln.de/ (07. November 2023).
16 https://verwaltung.uni-koeln.de/forschungsmanagement/content/index_ger.html (07. November 2023).
17 https://rrzk.uni-koeln.de/daten-speichern-teilen/projekt-und-datenmanagement-am-rrzk (07. November 2023).
18 Blumtritt Jonathan/ Helling, Patrick/ Mathiak, Brigitte/ Rau, Felix/ Witt, Andreas: „Forschungsdatenmanagement in den Geisteswissenschaften an der Universität zu Köln". in: o-bib, Das offene Bibliotheksjournal / Herausgeber VDB (2018) S. 104-117. doi: https://doi.org/10.5282/o-bib/2018H3S104-117.
19 https://dch.phil-fak.uni-koeln.de/ueber-das-dch/leitung (07. November 2023).
20 https://dch.phil-fak.uni-koeln.de/forschung (07. November 2023).
21 https://www.nfdi.de/ (07. November 2023).
22 https://nfdi4culture.de/index.html (07. November 2023).
23 https://text-plus.org/ (07. November 2023).
24 https://4memory.de/ (07. November 2023).
25 https://www.nfdi4objects.net/ (07. November 2023).
26 https://dch.phil-fak.uni-koeln.de/ueber-das-dch/team (07. November 2023).
27 https://dch.phil-fak.uni-koeln.de/ueber-das-dch/satzung-und-beirat (07. November 2023).
28 Helling, Patrick/ Blumtritt, Jonathan/ Mathiak, Brigitte: „Der Beratungsworkflow des Data Center for the Humanities (DCH) an der Universität zu Köln". in: o-bib. Das offene Bibliotheksjournal / Herausgeber VDB (2018) S. 248-261. doi: https://doi.org/10.5282/O-BIB/2018H4S248-261.
29 Helling, Patrick: „Wie Geht Bedarfsorientiertes Forschungsdatenmanagement?: Durchführung, Protokollierung Und Analyse von Beratungsvorgängen Im Geisteswissenschaftlichen Forschungsdatenmanagement Am Beispiel Des Data Center for the Humanities (DCH)". in: ABI Technik 42 4 (2022) S. 242-57. doi: https://doi.org/10.1515/abitech-2022-0044.

im Bereich audiovisuelle Sprachdaten. Es betreibt gemeinsam mit dem Regionalen Rechenzentrum der Universität zu Köln das Language Archive Cologne (LAC),[30] ein fachspezifisches Repositorium für audiovisuelle Sprachdaten. Darüber hinaus bietet das DCH zielgruppenorientierte FDM-Lehr- und Weiterbildungsangebote[31] und stellt Handreichungen und Empfehlungen für den Umgang mit Forschungsdaten in den Geisteswissenschaften zur Verfügung.[32] Zusätzlich wurden am Zentrum Publikationsworkflows entwickelt und etabliert, die bspw. für die Veröffentlichung großer Datenmengen auf der generischen Infrastruktur Zenodo,[33] u.a. für die Publikation von Konferenzbeiträgen wie zu den Jahreskonferenzen des Verbands Digital Humanities im deutschsprachigen Raum e.V.,[34,35] sowie für archäologische Forschungsergebnisse, bspw. in der iDAIWorld,[36,37] regelmäßig genutzt werden.

Für seine Services betreibt das Datenzentrum selbst keine eigene Infrastruktur, es werden ausschließlich bestehende Infrastrukturkomponenten, i.d.R. durch das Regionale Rechenzentrum an der Universität angeboten, nachgenutzt. Entsprechend wird bspw. für die Workflows des Dark und Cold Archiving Services des DCH u.a. der TSM-Bandspeicher des RRZK genutzt.[38]

3 Der Archivierungsservice des DCH

Im Folgenden sollen nun die technischen und organisatorischen Aspekte des Archivierungsservices des DCH vorgestellt und beschrieben werden. Ziel ist es dabei, nicht eine umfassende detaillierte Dokumentation des Services zu liefern, sondern vielmehr eine Beschreibung des Services sowie der Motivation für zentrale Entwicklungsentscheidungen.

3.1 Grundsätzliches

Der Archivierungsservice ist bewusst einfach aufgebaut. Aus diesem Grundprinzip der Einfachheit folgen Wartbarkeit und Verlässlichkeit als wichtigste Eigenschaften. Entsprechend der Ausrichtung des DCH werden – wo möglich – infrastrukturelle Serviceangebote von Partnerinstitutionen[39] nachgenutzt. Dadurch basiert der Archivierungsservice auf bestehenden allgemeinen Infrastrukturkomponenten, die unabhängig vom Archivierungsservice betrieben werden und von einer größeren Gruppe an Services und Institutionen genutzt werden.

Grundsätzlich gilt, dass die Bedarfe der Forscherinnen und Forscher der Philosophischen Fakultät und die technischen, organisatorischen und personell-finanziellen Möglichkeiten des DCH den Archivierungsservice entscheidend in seinem Umfang und Ausprägung definieren.

3.2 Technische Perspektive

Aus technischer Perspektive besteht der Archivierungsservice aus mehreren lose miteinander gekoppelten Servicekomponenten, bei denen an entscheidenden Stellen Services von Partnerinstitutionen wie dem RRZK, der USB und dem CCeH nachgenutzt werden. Der Arbeitsablauf beinhaltet sowohl automatisierte als auch manuelle Arbeitsschritte sowie einige organisatorische Komponenten, die im Detail erst im nächsten Abschnitt beschrieben werden.

Technisch basiert der Archivierungsservice in seinem Kern auf einem Archive Information Package (AIP) im BagIt-Format[40] und einem kleinen Set beschreibender, technischer und administrativer Metadaten, die zusammen mit den zu archivierenden Forschungsdaten und -ergebnissen in einem BagIt-Container abgelegt werden. Das AIP, bestehend aus Datenpaketformat und den in den Metadaten enthaltenen Informationen, stellt den Kern des Archivierungsservices aus technischer Perspektive dar. Alle anderen technischen (und organisatorischen) Aspekte sind modular und austauschbar, ohne dass sich der Archivierungsservice aus Nutzerinnen-/Nutzer-Sicht essentiell verändern würde.

Aktuell wird der Archivierungsservice auf dem Bandspeichersystem des RRZK und einer mit diesem verbundenen virtuellen Maschine (VM) ausgeführt. Diese technische Infrastruktur, bestehend aus Bandspeicher und VM, bildet das infrastrukturelle Grundsystem des Services.

Für den Cold Archiving Service wird darüber hinaus der Registrierungsservice für Digital Object Identifier (DOI) bei DataCite der Universitäts- und Stadtbibliothek genutzt. Eine vom Cologne Center for eHumanities betrie-

30 https://lac.uni-koeln.de/ (07. November 2023).

31 https://dch.phil-fak.uni-koeln.de/sichtbarkeit-und-lehre/lehrveranstaltungen (07. November 2023).

32 https://dch.phil-fak.uni-koeln.de/guidelines-fuer-forschende (07. November 2023).

33 https://zenodo.org/ (07. November 2023).

34 https://dig-hum.de/ (07. November 2023).

35 Helling, Patrick/ Debbeler, Anke/ Borges Rebekka: „Konferenzbeiträge strategisch publizieren". in: o-bib. Das offene Bibliotheksjournal / Herausgeber VDB (2022) S. 1-17. doi: https://doi.org/10.5282/O-BIB/5835.

36 https://idai.world/ (letzter Zugriff: 07. November 2023).

37 Lammers, Lukas/ Reinke, Eva/ Fäder, Eymard: „Archäologisches Forschungsdatenmanagement in der Praxis: Das Projekt FAIR.rdm im SPP2143 ‚Entangled Africa'". in: Archäologische Informationen Bd. 45 (2023) S. 9-14. doi: https://doi.org/10.11588/AI.2022.1.95251.

38 https://rrzk.uni-koeln.de/daten-speichern-teilen/backup-system-tsm (07. November 2023).

39 hierzu zählen insbesondere das Regionale Rechenzentrum (RRZK), die Universitäts- und Stadtbibliothek (USB) und das Cologne Center for eHumanities (CCeH), https://cceh.uni-koeln.de/ (07. November 2023).

40 Kunze, John/ Littman, Justin/ Madden, Liz/ Scancella, John/ Adams, Chris: „The BagIt File Packaging Format (V1.0)". in: Internet Engineering Task Force (2018). Online: https://tools.ietf.org/html/rfc8493.

bene Gitlab-Instanz dient als System zum Management der Software und der organisatorischen und technischen Abläufe des Workflows.

Das AIP folgt dem BagIt-Format und enthält die drei obligatorischen Komponenten eines BagIt-Objektes: (1) die Bag-Declaration-Datei *bagit.txt*, die diese Datenstruktur als Instanz des BagIt-Formats deklariert, (2) einen Ordner *data*, in dem die eigentlichen Forschungsdaten und -ergebnisse (der sogenannte payload) abgelegt sind und (3) mindestens eine Manifestdatei mit Prüfsummen für alle im Datenpaket enthaltenen Dateien.

Innerhalb des Archivierungsworkflows des DCH werden insgesamt drei Manifestdateien für die drei gängigen Algorithmen md5, sha1, sha512 hinzugefügt. Darüber hinaus werden Metadaten zum BagIt-Objekt in der Datei *bag-info.txt* erfasst. Die Metadaten zu den zu archivierenden Forschungsdaten und -ergebnissen werden als Markdown-Dokument unter dem Namen README.md gespeichert. Ein sehr einfaches AIP (Abb. 1) mit nur einer zu archivierenden Datei (*20090427b_1015.WAV*) sähe entsprechend wie folgt aus:

Abbildung 1: beispielhaftes Archive Information Package im Archivierungsservice des DCH

Diese Datenstruktur bildet, wie bereits beschrieben, aus technischer Perspektive den eigentlichen Kern des Archivierungsservices. Alle weiteren technischen Systeme sind dafür da, die Erstellung, Überprüfung, Speicherung und das Retrieval dieser Datenstruktur zu ermöglichen. Sie bestehen dabei aus mehreren Komponenten: Wie bereits beschrieben, wird im Grundsystem des Archivierungsservices aktuell eine sogenannte (1) Managed VM des RRZK genutzt, deren Betriebssystem (Red Hat Enterprise Linux) und installierte Softwarepakete durch einen Mitarbeiter des RRZK administriert werden. Auf der einen Seite entfallen für die Unterhaltung des Archivierungsworkflows dadurch jegliche Systemadministratorarbeiten für das DCH. Auf der anderen Seite sind jedoch die zur Verfügung stehenden Softwarepakete für die VM eingeschränkt und neue Pakete können nicht ohne Weiteres nachinstalliert werden. Letzteres hatte auf alle weiteren Technologieentscheidungen einen entscheidenden Einfluss. Die zweite Komponente des technischen Grundsystems ist das (2) Bandspeichersystem des RRZK, das aktuell über den IBM Tivoli Storage Manager (TSM)[41] angesprochen wird. Auf jeder Managed VM ist der TSM-Client *Distributed Storage Manager Client* als Kommandozeilenprogramm (dsmc) vorinstalliert und wird entsprechend innerhalb des Archivierungsworkflows des DCH genutzt. Folglich basiert das Grundsystem des Archivierungsservice ausschließlich auf Standardinfrastrukturangeboten des RRZK, weitere Service-spezifische Administration wird nicht benötigt.

Auf diesem Grundsystem setzen die für den Archivierungsservice spezifischen Programme auf. Aufgrund der technischen Einschränkungen von Managed VMs in Bezug auf Softwarepakete werden im Archivierungsworkflow des DCH keine externen Bibliotheken genutzt und für die Umsetzung ausschließlich auf die Bash-Skriptsprache zurückgegriffen. Da die Shell Bash und die auf ihr ausführbaren Bash-Skripte essentieller Teil eines jeden Red Hat Enterprise Linux sind, ergeben sich hier keine Abhängigkeiten von zusätzlich installierten Softwarepaketen. Das Grundsystem und die Bash-Skripte setzen den gesamten Archivierungsworkflow von den Vorbearbeitungsschritten zur Erstellung des AIP, über das Schreiben des AIPs ins Bandarchiv, der Überprüfung des AIPs bis zur Erstellung des internen Berichts um.

Die bisher beschriebenen Strukturen und Systeme sind für die Dark und Cold Archiving Variante identisch. Beim Dark Archiving Workflow erhalten die Datengeberinnen/-geber am Ende eine Archivierungsdokumentation als Nachweis für den abgeschlossenen Archivierungsprozess. Beim Cold Archiving Workflow werden im Anschluss zusätzlich ein Subset der Metadaten veröffentlicht und eine DOI registriert. Dieser Abschnitt des Archivierungsworkflows funktioniert unabhängig von der bereits beschriebenen technischen Infrastruktur und wird aktuell noch weitestgehend händisch realisiert: Auf Grundlage der in der Markdown-Datei README.md gebündelten Metadaten werden eine HTML-Datei, ein als JSON-LD serialisierte Schema.org Objekt, und ein dem COinS-Format entsprechendes *span*-Element erstellt, welches unter Zuhilfenahme der Software Zotero produziert wird. Zusätzlich wird eine DataCite-XML-Datei erzeugt. Die HTML-Datei

41 https://de.wikipedia.org/wiki/Tivoli_Storage_Manager (07. November 2023).

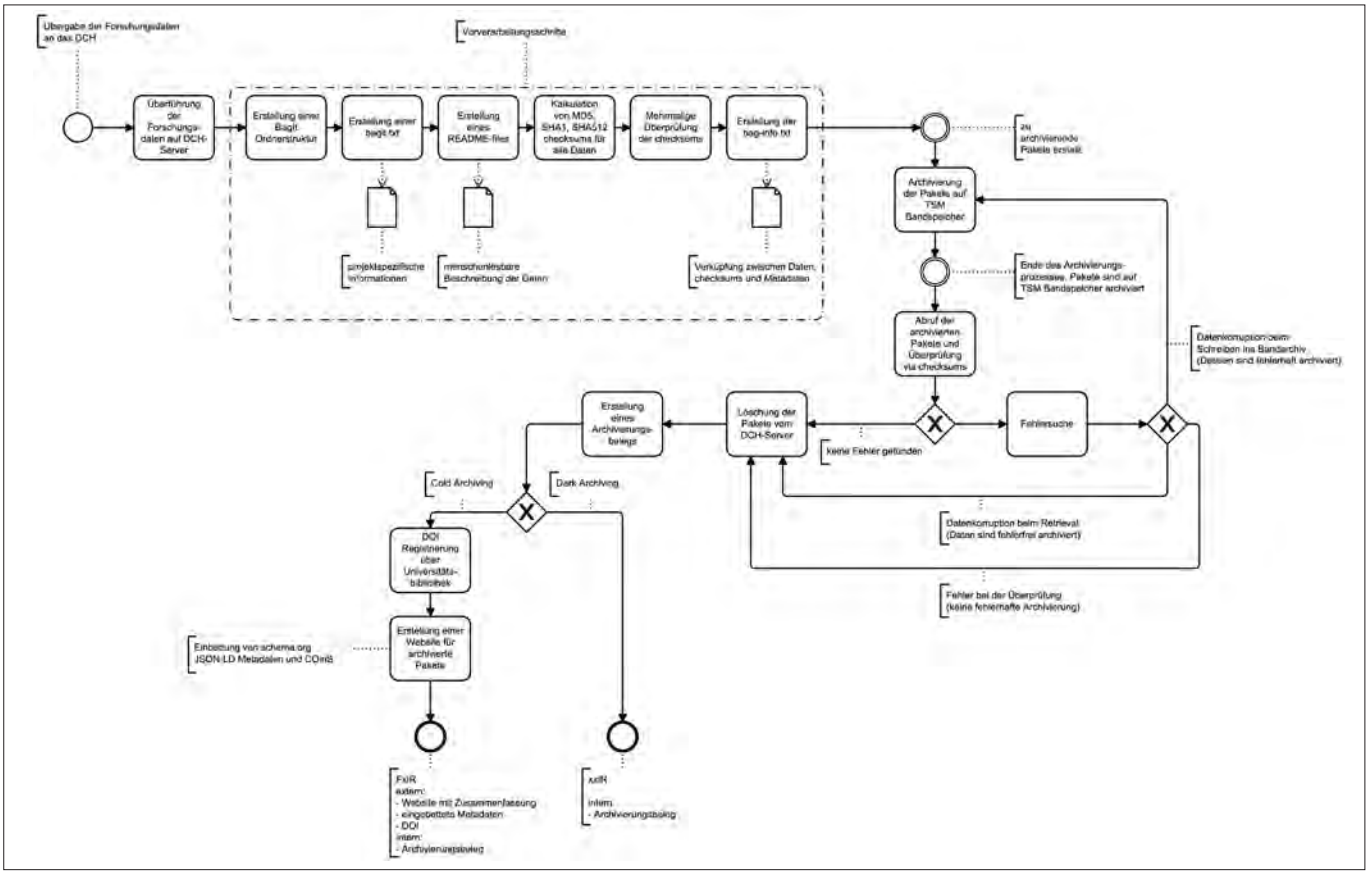

Abbildung 2: Der Archivierungsworkflow des DCH

wird über das Content Management System Typo3, das vom RRZK für die gesamte Universität zu Köln betrieben wird, veröffentlicht.[42] Die Schema.org- und COinS-Metadaten werden in den Quelltext der Website eingebettet. Die DataCite-XML-Datei wird mit zusätzlichen Informationen an den DOI-Registrierungsdienst der USB Köln übergeben.

Der gesamte Archivierungsworkflow (Abb. 2) stellt sich formal modelliert schließlich wie folgt dar:

3.3 Organisatorische Perspektive

Aus organisatorischer Perspektive besteht der Archivierungsservice aus einem definierten Archivierungsworkflow, für den innerhalb des DCH ein designiertes Team verantwortlich ist.

Er besteht zunächst aus (1) einem initialen Gespräch mit der/dem Datengeberin/Datengeber, (2) der Übergabe der zu archivierenden Forschungsdaten und -ergebnisse sowie (3) des Unterzeichnens eines Depositor Agreements.[43] Im Anschluss beginnt der in Abbildung 1 visualisierte Workflow mit entsprechender Qualitätskontrolle des Submission Information Package (SIP) (z.B. keine Symlinks, korrektes README), der Erstellung des AIP, der Archivierung, Ergebniskontrolle, Metadatenerstellung und -veröffentlichung sowie dem finalen Report. Der komplette Workflow von der ersten Kontaktaufnahme bis zum Endbericht an die/den Datengebende/Datengebenden wird über ein Git-Repository in Gitlab organisiert. Jedes Deposit wird als Issue in Gitlab abgebildet und ein Issue-Board mit elf definierten Arbeitsschritten zeigt den Zustand jedes Archivierungsvorgangs für jedes Deposit individuell an. Die elf Arbeitsschritte sind in der internen Dokumentation wie folgt definiert:

1. Request
   a. Anfrage von Datengeberin/Datengeber
   b. Antwort von DCH mit Übersendung von Readme und Übergabeprotokoll
   c. Rücksendung Readme, zu archivierender Forschungsdaten und Übergabeprotokoll durch Datengeberin/Datengeber
2. SIP Creation
   a. Alle Dateien in einem Ordner abgelegt
   b. Qualitätskontrolle (keine Symlinks, Kontrolle des Übergabeprotokolls und des Readme)
3. Data staged
   a. Alle Daten befinden sich qualitätskontrolliert in einem Ordner auf dem Server
   b. Forschungsdaten bereit zur Archivierung
   c. Übergabe an eine/-n andere/-n Mitarbeiterin/Mitarbeiter möglich
4. Archiving
   a. Siehe Dokumentation des Archiving-Workflows (Core-Workflow)
   b. Größe und Archivierungsdatum in der Issue im Board eingetragen

42 bspw. https://dch.phil-fak.uni-koeln.de/bestaende/datensicherung/toboliu-project-first-funding-period-data (08. November 2023).

43 Lammers, Lukas: „Übergabeprotokoll zur Langzeitarchivierung des DCH der Universität zu Köln (1.0)". (2023). doi: https://doi.org/10.5281/zenodo.10083628.

HAN

Die Komplettlösung für das Management von Online-Ressourcen

Direkter Zugriff weltweit

Jedes Endgerät

Zentrales Management

Statistische Auswertung

Transparenz bei den Zugriffen

Sichere Lizenzierung

www.hh-han.com

Ein Produkt der H+H Software GmbH

5. Archived
   a. Forschungsdaten sind archiviert
   b. Übergabe an eine/-n andere/-n Mitarbeiterin/Mitarbeiter möglich
6. Metadata creation
   a. HTML, JSON-LD, COinS erstellt
   b. HTML-Seite erstellt, URL ist angelegt. DOI (nach interner DOI-Zählung) ist blockiert
   c. DOI im Namen der Issue eintragen
7. Metadata published
   a. Ressource hat DOI und URL erhalten, über die die Metadaten auffindbar sind
8. DOI registration
   a. Data Cite XML Metadaten erzeugt
   b. Auswahl von Library of Congress Subject Headings
   c. E-Mail an DOI-Service der USB Köln verschickt
9. Final Testing
   a. Mit rekursivem Link der Landing-Page überprüft, ob DOI vergeben wurde
10. Notification
    a. Prozess abgeschlossen
    b. Versendung einer E-Mail mit der Zitierempfehlung des Datensatzes an Datengeberin/Datengeber
    c. Nach dem Versenden der E-Mail:
11. Closed

Datengeberinnen und Datengeber nehmen dabei in der Regel initial per E-Mail Kontakt mit dem Archivierungsteam des DCH über eine dezidierte Kontaktadresse auf. Wenn der Prozess in einem normalen Forschungsdatenmanagement-Beratungsvorgang angestoßen wird, wird der Kontakt mit dem Archivierungsteam über eine gemeinsame E-Mail an die/den potenzielle/-n Datengeberin/Datengeber und die Kontaktadresse hergestellt.

Der Archivierungsservice des DCH wird aktuell durch zwei Mitarbeiterinnen und Mitarbeiter und eine wissenschaftliche Hilfskraft betreut. Ein/-e Mitarbeiter/Mitarbeiterin fungiert dabei als zentrale Ansprechperson für den Service. Sie/Er ist für die konzeptuelle und technische Weiterentwicklung des Services verantwortlich, übernimmt den Großteil des Erstkontaktes mit neuen Datengeberinnen und Datengebern und aktuell auch die Erstellung der Metadaten sowie die Registrierung der DOI. Die/Der andere Mitarbeiterin/Mitarbeiter und die wissenschaftliche Hilfskraft sind für die Durchführung der SIP- und AIP-Erstellung und die Archivierung verantwortlich. Die Koordination der Arbeitsschritte und gegebenenfalls der Übergaben des Vorgangs von einer/-m Mitarbeiterin/Mitarbeiter an die/den Andere/-n erfolgt über Zuweisungen in Gitlab und Kommunikation über einen abteilungsinternen Mattermost-Chat.

4 Fazit

Die Dark und Cold Archiving Services am Data Center for the Humanities (DCH) bieten einen niederschwelligen Service zur Datensicherung und zum Datennachweis und füllen eine Leerstelle in der Forschungsdatenmanagement-Servicelandschaft für die Philosophische Fakultät der Universität zu Köln. Dabei ist der Archivierungsservice so angelegt, dass er mit wenig personellem Aufwand und vor allem mit wenig technischer Administration und Wartung betrieben werden kann. Dies wird dadurch erreicht, dass viele technische Komponenten des Services existierende Infrastrukturangebote von zentralen Einrichtungen der Universität nutzen und die spezifisch für diesen Service entwickelten Komponenten möglichst einfach und wartungsarm gestaltet wurden. Das Ergebnis ist ein schlanker Service, der den personellen Gegebenheiten des DCH entspricht und trotzdem die Bedarfe der Forscherinnen und Forscher erfüllt. Entsprechend konnten über den Archivierungsservice des DCH allein in den Jahren 2022 und 2023 bereits über 2.500 GB Forschungsdaten und -ergebnisse langfristig archiviert werden.[44]

Patrick Helling
ist Datenmanager und Koordinator am Data Center for the Humanities (DCH) an der Universität zu Köln. Er unterstützt Geisteswissenschaftlerinnen und -wissenschaftler bei Fragen zum Management von Forschungsdaten und bietet fachspezifische FDM-Services an. Darüber hinaus ist er Mitarbeiter im FDM-Team des DFG-Schwerpunktprogramms 2207 „Computational Literary Studies". Als Data Steward des Verbands Digital Humanities im deutschsprachigen Raum e.V. entwickelt er Datenmanagementstrategien für den wissenschaftlichen Output der DHd-Community. Er promoviert an der Universität zu Köln und entwickelt in diesem Rahmen ein formales Beschreibungsmodell für fachspezifische Forschungsdatenmanagement-Bedarfe und -Services in den Geisteswissenschaften.
patrick.helling@uni-koeln.de


© Marcello Perathoner

Felix Rau
ist Geschäftsführer des Data Center for the Humanities (DCH) an der Universität zu Köln. Am DCH ist er u.a. verantwortlich für Betrieb und Weiterentwicklung der Datenarchivierungs- und Datenpublikationsservices sowie den Kompetenzbereich audiovisuelle Daten. Er arbeitet im NFDI-Konsortium Text+ und beteiligt sich auch an anderen NFDI-Konsortien und an der europäischen Infrastruktur CLARIN ERIC.
f.rau@uni-koeln.de

44 Übersicht über die dabei bisher ins Cold Archiving überführten Forschungsdaten und -ergebnisse: https://dch.phil-fak.uni-koeln.de/bestaende/datensicherung (08. November 2023).