

**Disziplinspezifisches
Forschungsdatenmanagement
FDM-Bedarfserfassung in den
Computational Literary Studies**

Helling, Patrick

patrick.helling[at]uni-koeln.de

Institut für Deutsche Philologie, Universität Würzburg, Deutschland

ORCID-iD: 0000-0003-4043-165X

Jung, Kerstin

kerstin.jung[at]ims.uni-stuttgart.de

Institut für Deutsche Philologie, Universität Würzburg, Deutschland

ORCID-iD: 0000-0002-9548-8461

Pielström, Steffen

pielstroem[at]biozentrum.uni-wuerzburg.de

Institut für Deutsche Philologie, Universität Würzburg, Deutschland

Zusammenfassung. Die außerordentliche Heterogenität geisteswissenschaftlicher Forschungsfragen und Methoden stellt das Forschungsdatenmanagement (FDM) vor besondere Herausforderungen. Die DFG fördert seit 2020 ein eigenes Schwerpunktprogramm für die Computational Literary Studies (SPP-CLS), ein junges, interdisziplinäres Forschungsfeld unter Beteiligung der Literaturwissenschaften, der Informatik und der Computerlinguistik. Mit zehn geförderten Forschungsprojekten an verschiedenen Universitäten in Deutschland und der Schweiz kann dieses Schwerpunktprogramm als Blaupause aufgefasst werden, an der sich die FDM-Bedarfe dieser noch jungen Fachdisziplin untersuchen lassen. Hierzu wurde mit allen Einzelprojekten eine 2-stufige Befragung basierend auf 47 meist offen gehaltenen Leitfragen durchgeführt. Die Ergebnisse stellen die Basis für eine umfassende FDM-Landschaftsvermessung und Bedarfserfassung dar, die als Ausgangspunkt für die effektive Beschreibung und Entwicklung von FDM-Lösungsstrategien für das Programm, als auch perspektivisch für die gesamte Fachdisziplin der Computational Literary Studies dient.

1 Ausgangssituation

Forschungsdatenmanagement ist ein zentraler Bestandteil ausgezeichneter und nachhaltiger Forschung.¹ Aufgrund der Diversität geisteswissenschaftlicher Forschungsfragen, Methoden und Ziele sowie einer ausdifferenzierten, fachdisziplinären Forschungslandschaft sind geistes- und kulturwissenschaftliche Forschungsdaten typischerweise sehr heterogen.² Entsprechend bedarf es (fach)-spezifischer Lösungsstrategien, die eine passgenaue Bedienung fachlich geprägter Bedarfe im geisteswissenschaftlichen Forschungsdatenmanagement (FDM) im Sinne der FAIR-Prinzipien³ ermöglichen.

Es existieren bereits unterschiedliche Angebote, die auf verschiedene fachliche oder strukturelle Communities spezialisiert sind: FDM-Services für die Breite der Geisteswissenschaften werden u.a. vom Infrastrukturkonsortium CLARIAH entwickelt.⁴ Unterschiedlich angesiedelte Infrastrukturen mit ausdifferenzierteren Angeboten stellen standortunabhängige, fachspezifische Lösungen zur Verfügung.⁵ Darüber hinaus konnten sich standortspezifisch einzelne fach- und disziplinspezifische FDM-Angebote etablieren⁶ und nicht zuletzt sind die Entwicklungen hin zu einer Nationalen Forschungsdateninfrastruktur (NFDI) in ihrer konsortialen Organisationsform stark daten- und fachgetrieben.⁷

Ein zentrales Element bei der Entwicklung und Etablierung von solchen spezifischen und strukturell unterschiedlich angegliederten FDM-Servicestrukturen sind zu Beginn i.d.R. Bedarfserfassungen und -abfragen⁸ sowie die Orientierung an Fachtraditionen.⁹ Die adäquate Adressierung bestimmter (Teil)-Communities stellt sowohl bei der Bedarfserfassung

¹ Bryant, Lavoie & Maplas 2017.

² Pempe 2012.

³ Wilkinson et al. 2016.

⁴ „CLARIAH-DE“, letzter Zugriff 30. März 2021, <https://www.clariah.de/>.

⁵ vgl. „Language Archive Cologne“, letzter Zugriff 29. März 2021, <https://lac.uni-koeln.de/>. „GerDraCor“, letzter Zugriff 29. März 2021, <https://dracor.org/ger>. „Deutsches Textarchiv“, letzter Zugriff 29. März 2021, <https://www.deutschestextarchiv.de/>. „DIPF - Leibniz-Institut für Bildungsforschung und Bildungsinformation“, letzter Zugriff 29. März 2021, <https://www.dipf.de/de/dipf-aktuell>. „CrossAsia“, letzter Zugriff 29. März 2021, <https://crossasia.org/>.

⁶ u.a. Blumtritt et al. 2018; Helling, Moeller & Mathiak 2018.

⁷ „Nationale Forschungsdateninfrastruktur“, letzter Zugriff 29. März 2021, <https://www.nfdi.de/>.

⁸ u.a. Kronenwett 2016; Akers & Doty 2013; Vock 2019.

⁹ „NFDI4Memory - Problem Stories“, letzter Zugriff 29. März 2021, <https://4memory.de/problem-stories-overview/>. „NFDI Text+ - User Stories“, letzter Zugriff 30. März 2021, <https://www.text-plus.org/forschungsdaten/user-stories/>.

als auch bei der Einrichtung und Verbreitung von Services in diesem Zusammenhang häufig eine wesentliche Herausforderung dar.

2 Digitale Literaturwissenschaften in a Nutshell - das DFG Schwerpunktprogramm "Computational Literary Studies"

Unter dem Begriff „Computational Literary Studies“ hat sich in den letzten Jahren ein Forschungsfeld etabliert, das computergestützte, quantitative Forschungsmethoden zur Analyse digitalisierter, literarischer Texte einsetzt, um literaturwissenschaftliche Forschungsfragen zu untersuchen. Es handelt sich damit um ein interdisziplinäres Forschungsfeld zwischen Informatik, Computerlinguistik und Literaturwissenschaft. Die DFG fördert seit 2020 ein eigenes Schwerpunktprogramm "Computational Literary Studies" (SPP-CLS).¹⁰ Dieses setzt sich aus zehn Teilprojekten, einem Zentralprojekt sowie einem assoziierten Projekt an unterschiedlichen Standorten in Deutschland und der Schweiz zusammen.

Wenngleich das SPP-CLS selbst nicht die gesamte Forscher*innen-Landschaft der Fachdisziplin abbildet, so stellt es dennoch einen klar abgesteckten Raum dar, der sich durch eine Konzentration fachlich naher Forschungsprojekte auszeichnet und als Blaupause für die noch junge Fachdisziplin verstanden werden kann.

3 FDM Landschaftsvermessung im SPP-CLS

Aufgrund der Konzentration fachlich naher Forschungsprojekte bietet es sich für das Programm an gezielt spezifische FDM-Empfehlungen und Best Practices zu formulieren und Lösungsstrategien für die im Programm angesiedelten Projekte und modellhaft für die gesamte Fachdisziplin zu identifizieren bzw. zu entwickeln.

Um entsprechende FDM-Bedarfe und -Bedingungen innerhalb des SPP-CLS zu erfassen, wurden qualitative Interviews via Zoom mit Vertreter*innen aller Teilprojekte geführt. Die Durchführung der Interviews orientierte sich an einem Gesprächsleitfaden, der aus 47 meist offenen Fragen besteht (Helling et al. 2020). Die Gespräche wurden

¹⁰ SPP 2207 „Computational Literary Studies“, letzter Zugriff: 27. Juli 2021, <https://dfg-spp-cls.github.io/>.

aufgezeichnet. Zur Steigerung der Qualität der Antworten wurden jedem Projekt seine Ergebnisse inkl. gegebener Antwortmöglichkeiten aus Interviews mit anderen Teilprojekten für eine Überarbeitung zur Verfügung gestellt.

Die Ergebnisse der Interviews stellen die Grundlage zur bedarfsorientierten Entwicklung von Best Practices und Empfehlungen für den Umgang mit Forschungsdaten sowie dem Aufbau von passgenauen und disziplinspezifischen FDM-Lösungsstrategien dar.

3.1 Erste Ergebnisse der Landschaftsvermessung

Um einen grundsätzlichen Überblick über Methoden und disziplinspezifische Strukturen innerhalb des SPP-CLS zu erhalten, wurde unter anderem die Anwendung von Analyseverfahren und methodischen Werkzeugen erfragt (Abb. 1). Dies ermöglicht es Trends und Dynamiken innerhalb des Programms abzulesen, die schließlich als beispielhafte Beschreibung für die Definition eines State-of-the-Art der gesamten Fachdisziplin verwendet werden können.

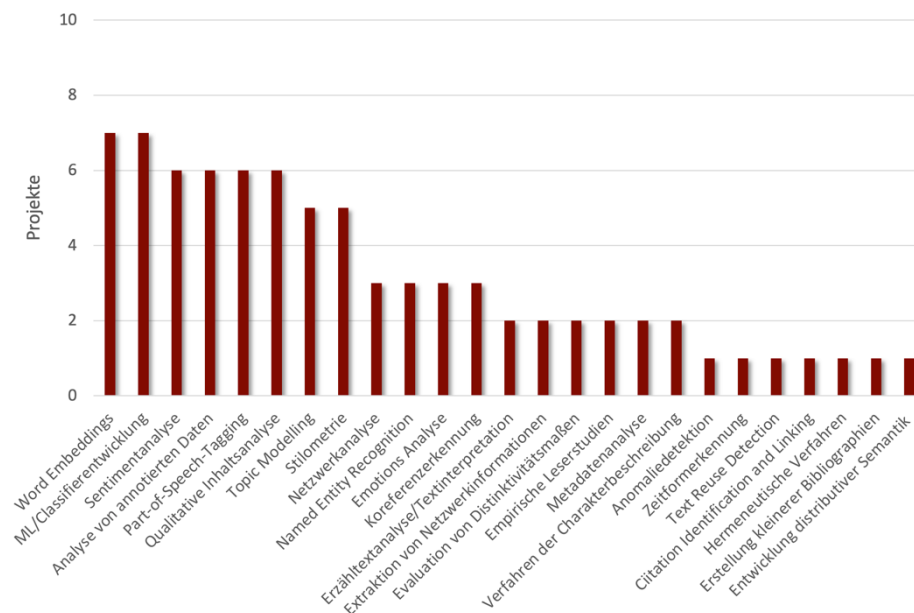


Abb. 1. Analyseverfahren und methodische Werkzeuge im SPP-CLS.

Ähnliches gilt auch in Bezug auf die Verwendung virtueller Umgebungen und technischer Infrastruktur zur Projektverwaltung (Abb. 2) sowie

zentraler Tools, bspw. zur manuellen und automatischen Annotation (Tabelle 1).

Tabelle 1. Genutzte Annotationstools innerhalb des SPP-CLS.

Annotationstool	Projekte
Catma	5
TreeTagger	1
CorefAnnotator	1
Sentiment Analyzer	1

Dabei stellt die erfasste Relevanz einiger Tools wie u.a. Github und Gitlab als Versionsverwaltungssysteme bereits eine wichtige Information in Bezug auf das aktive FDM innerhalb des SPP-CLS dar (Abb. 2).

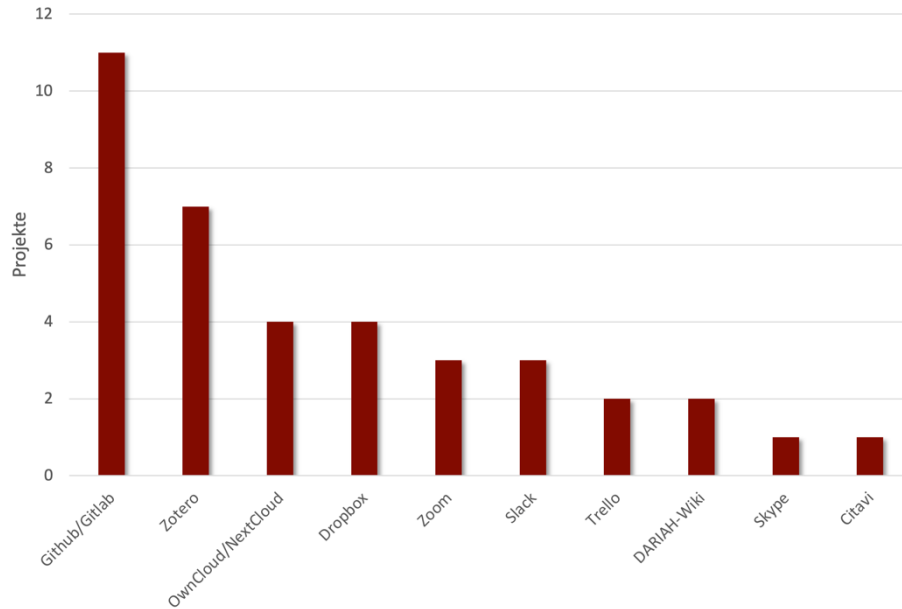


Abb. 2. Virtuelle Umgebungen und technische Infrastruktur im SPP-CLS.

Neben der Verwendung bestimmter Programmiersprachen (Tabelle 2) sowie Bibliotheken (Abb. 3), die auf Technologie-Stacks hinweisen, mit denen am Ende der Projektlaufzeiten umgegangen werden muss, sind auch die Nutzung und Produktion von bestimmten Datentypen (Tabelle 3) und -formaten (Abb. 4) wichtige Indikatoren für ein passgenaues FDM.

Tabelle 2. Programmier- und Skriptsprachen im SPP-CLS.

Programmier-/Skriptsprache	Projekte
Python	9
R	4
Shell-Skripte	4
Java	3
X-Technologien	2
JavaScript	2
HTML	2
CSS	2
SQL	1

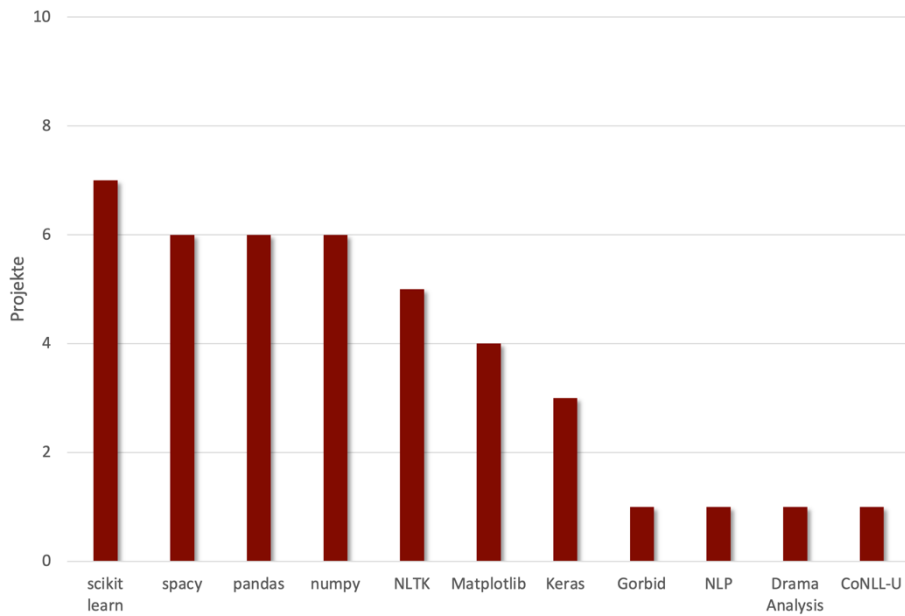


Abb. 3. Genutzte Bibliotheken im SPP-CLS.

Tabelle 3. Genutzte Datentypen innerhalb der Teilprojekte im SPP-CLS.

Datentypen	Projekte
Text	10
Softwarecode	9
Numerische Daten	6
Bilddaten	5
Bibliographische Daten	4
Wörterbücher/Listen	2
Interviewdaten	2
Netzwerkdaten	1

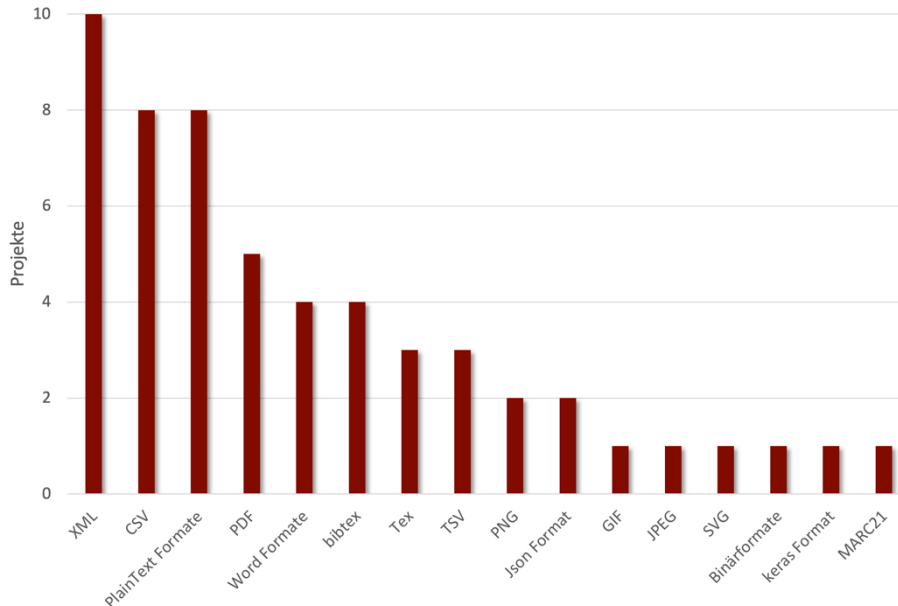


Abb. 4. Datenformate, die in den Teilprojekten eine wichtige Rolle spielen.

Allerdings stellen nicht nur Forschungsdaten Ergebnisse wissenschaftlicher Forschungsprozesse dar, sondern auch lebende Systeme wie bspw. Software, Websites oder dynamische Datenbanken (Tabelle 4).

Tabelle 4. Geplante lebende Systeme im SPP-CLS.

Lebende Systeme	Projekte
Website	7
Tools/Anwendungen	3
Bibliotheken	1
Dashboard	1

Zudem gilt es bei der Entwicklung einer FDM-Gesamtstrategie, insbesondere bei einer verteilten Organisationsstruktur wie im SPP-CLS, bereits in den Einzelprojekten etablierte FDM-Routinen und -strukturen (Abb. 5 und 6) zu integrieren bzw. zu berücksichtigen.

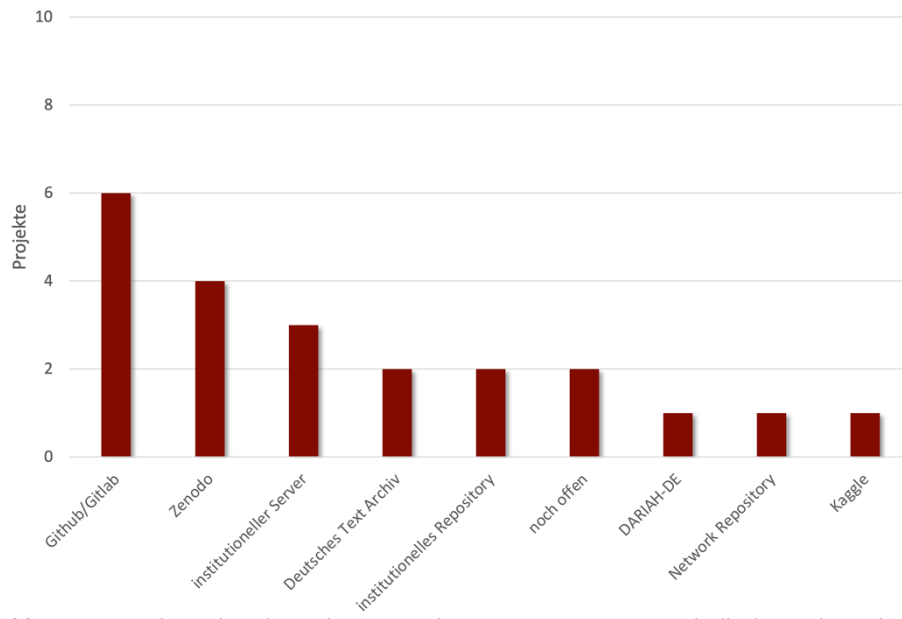


Abb. 5. Bereits bestehende Archivierungslösungen/-ansätze innerhalb der Teilprojekte.

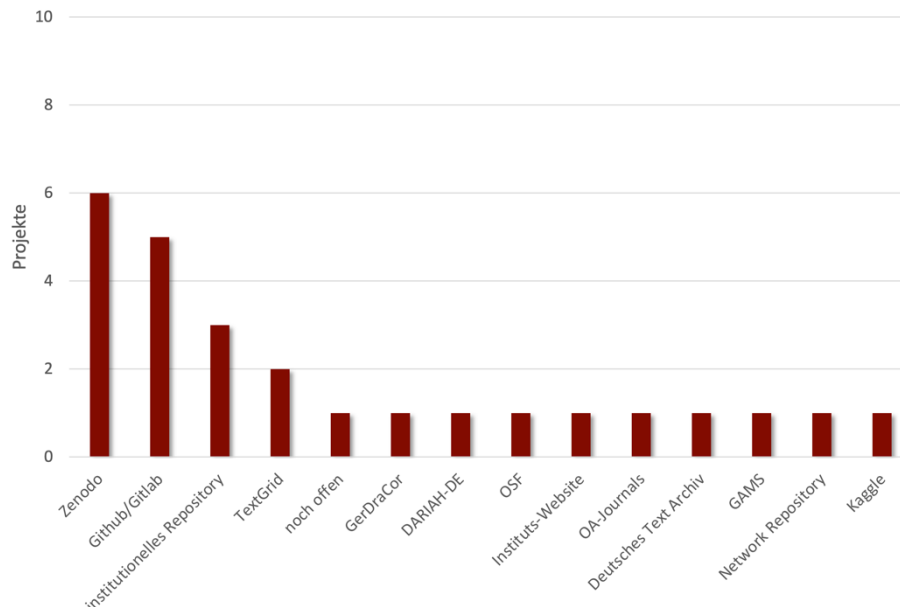


Abb. 6. Bereits bestehende Publikationslösungen/-ansätze innerhalb der Teilprojekte.

3.2 Maßnahmen, Herausforderungen und Perspektiven

Im Rahmen der Auswertungen der Interviewergebnisse konnten bereits einige Bedarfe und Anforderungen an das generelle FDM innerhalb des SPP-CLS und entsprechend perspektivisch für die Fachdisziplin identifiziert werden: Services wie Github und Gitlab stellen zentrale Instrumente der Verwaltung von und Kollaboration in Projekten dar. Aufgrund individueller Kapazitätsgrenzen und Zugriffsregelungen gab es allerdings den Bedarf einer geschlossenen, institutionell betriebenen Lösung für die Projektorganisation und den Transfer größerer Datenmodelle. Hierfür wurde eine gemeinsame Gitlab-Instanz aufgesetzt.

Trotz einiger projektübergreifend verbreiteter Datentypen und -formate weisen die Teilprojekte dennoch eine gewisse Heterogenität auf. Neben einer formatspezifischen Datenstrategie bedarf es am Ende schließlich auch eines gemeinsamen Metadatenmodells, um die technische und inhaltliche Heterogenität der Forschungsergebnisse als gemeinsames Gesamtergebnis des SPP-CLS abzubilden.

Eine besondere Herausforderung stellt der Umgang mit lebenden Systemen dar, die in einzelnen Teilprojekten auf Basis verschiedener Technologie-Stacks und mit unterschiedlichen Funktionalitäten entwickelt werden. Um diesen hochdynamischen Prozess zu kanalisieren, sollen zukünftig entsprechende Handreichungen und Best Practices formuliert und die Teilprojekte aktiv bei der Vorbereitung einer langfristigen Archivierung und Verfügbarmachung unterstützt werden.

4 Ausblick

Das SPP-CLS, die Ergebnisse der FDM-Landschaftsvermessung und die daraus entstehenden FDM-Services, -Maßnahmen und Best Practices haben das Potential als Blaupause für die gesamte Fachdisziplin zu fungieren.

In unserem Vortrag werden wir die beschriebenen Ergebnisse kompletieren und ein umfassenderes Bild der Forschungs- und FDM-Landschaft innerhalb der Computational Literary Studies zeichnen.

Bibliografie

Akers, Katherine G. und Jennifer Doty. 2013. „Disciplinary differences in faculty research data management practices and perspectives.“ *The International Journal of Digital Curation*, Volume 8, Issue 2: 5-26.
<https://doi.org/10.2218/ijdc.v8i2.263>.

Blumtritt, Jonathan, Patrick Helling, Brigitte Mathiak, Felix Rau und Andreas Witt. 2017. „Forschungsdatenmanagement in den Geisteswissenschaften an der Universität zu Köln.“ *o-bib - Das offene Bibliotheksjournal*, Bd. 5 Nr. 3 (2018): 104-117. <https://doi.org/10.5282/o-bib/2018H3S104-117>.

Bryant, Rebecca, Brian Lavoie und Constance Malpas. 2017. “A Tour of the Research Data Management (RDM) Service Space. The Realities of Research Data Management, Part 1.” <https://doi.org/10.25333/C3PG8J>.

- Helling, Patrick, Katrin Moeller und Brigitte Mathiak. 2018. „Forschungsdatenmanagement in den Geisteswissenschaften – der Dienstekatalog der AG-Datenzentren des Verbands Digital Humanities im deutschsprachigen Raum.“ *ABI Technik*, Band 38 Heft 3: 251–261.
<https://doi.org/10.1515/abitech-2018-3006>.
- Helling, Patrick, Kerstin Jung, Nils Reiter und Steffen Pielström. 2020. „Interviewleitfaden zur FDM-Bestandsaufnahme im Schwerpunktprogramm Computational Literary Studies.“ Zenodo. <http://doi.org/10.5281/zenodo.4269639>
- Kronenwett, Simone. 2016. „Forschungsdaten an der Philosophischen Fakultät der Universität zu Köln“. M.A.-Arbeit, Technische Hochschule Köln. <https://publiscologne.th-koeln.de/frontdoor/index/index/docId/925>. Letzter Zugriff: 08.04.2021.
- Pempe, Wolfgang. 2012. „Geisteswissenschaften.“ In *Langzeitarchivierung von Forschungsdaten. Eine Bestandsaufnahme*. Hrsg. Heike Neuroth, Stefan Strathmann, Achim Oßwald, Regine Scheffel, Jens Klump, Jens Ludwig, 137-60. Bielefeld: Verlag Werner Hülsbusch.
- Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. “The FAIR Guiding Principles for scientific data management and stewardship.” *Scientific Data*, 3(1): 160018. <https://doi.org/10.1038/sdata.2016.18>.
- Vock, Rainer. 2019. „Evaluation der FDM-Beratung 2019 – Evaluation des Beratungsangebots der Kontaktstelle Forschungsdatenmanagement (KS FDM) an der Friedrich-Schiller-Universität Jena.“ Bericht 4.2. eeFDM-Projekt (BMBF), Jena. <https://doi.org/10.22032/dbt.40382>.

