Bausteine Forschungsdatenmanagement
Empfehlungen und Erfahrungsberichte für die Praxis von Forschungsdatenmanagerinnen und -managern

# Lösungsansätze zu einer technischen Infrastruktur für Forschungsdatenmanagement

Timo Borst[i]

2018

**Zitiervorschlag**

Borst, Timo. 2018. Lösungsansätze zu einer technischen Infrastruktur für Forschungsdatenmanagement. *Bausteine Forschungsdatenmanagement. Empfehlungen und Erfahrungsberichte für die Praxis von Forschungsdatenmanagerinnen und -managern* Nr. 1/2018: S. 50-56. DOI: 10.17192/bfdm.2018.11.7939.



[i]ZBW Leibniz-Informationszentrum Wirtschaft. ORCID: 0000-0002-2481-029X

# 1 Abstract

Das heutige Management von Forschungsdaten beinhaltet in seinem Kern den Aufbau, die Entwicklung sowie Etablierung einer komplexen technischen Infrastruktur, an der verschiedene Ziel- und Interessengruppen beteiligt sind. Zu den Grundsatzfragen, die im Vorfeld oder begleitend zu einer technischen Lösung seitens einer Institution zu erwägen und zu entscheiden sind, gehören etwa, ob eher eine generische und/oder eine disziplinspezifische Lösung angestrebt werden soll, ob man in erster Linie das Publizieren oder auch das Prozessieren von Forschungsdaten unterstützen will, oder ob man eigenen lokalen Infrastrukturen lieber vertrauen will als kommerziellen Cloud-Diensten und entsprechenden Anbietern. Fragen dieser Art sind auf geeignete Weise zu thematisieren und zu klären, wobei sich Instrumente wie eine Anforderungs- und Umfeldanalyse, eine Stakeholderanalyse, ferner eine Marktsichtung hinsichtlich bestehender Lösungen anbieten.

# 2 Einleitung

Neben den organisatorischen, fachlichen, kulturellen und rechtlichen Aspekten gehören die technischen Aspekte beim Forschungsdatenmanagement sicher zu den wichtigsten Herausforderungen und Aktivitäten. Gerade in einem bereits relativ technisierten Umfeld ermöglicht erst eine (funktionierende) technische Infrastruktur ein operatives und effizientes Forschungsdatenmanagement, das die verschiedenen Beteiligten bestmöglich unterstützt. Dabei ist jederzeit zu berücksichtigen, dass ein solches Forschungsdatenmanagement zum einen allgemeine technisch-infrastrukturelle Aspekte beinhaltet – wie z.B. eine regelmäßige Sicherung der Daten, einen effizienten Zugriffsschutz oder die Verarbeitung von Metadaten –, zum anderen aber auch disziplinären oder Community-spezifischen Praxen und Kulturen Rechnung tragen muss, die zu ganz eigenen, singulären Lösungen führen können. So sind z.B. die Programme und Anwendungen, mit denen Forschungsdaten erzeugt, bearbeitet, analysiert oder auch visualisiert werden, bisweilen hochgradig speziell. Zum anderen sind die Daten selbst und ihre Formate – ob als Labor-, Befragungs- oder Sensordaten – sehr heterogen. Grundsätzlich besteht eine technische Lösung für Forschungsdatenmanagement aus keinem monolithischen System, sondern in einer Aneinanderreihung von über Schnittstellen, Protokolle und Datenformate kommunizierenden Systemkomponenten, die jede für sich einen oder mehrere Aspekte im Forschungsdatenzyklus unterstützen: Beispielsweise Sensoren oder Messwerkzeuge zur Generierung von Daten, virtuelle Forschungsumgebungen zu ihrer Verarbeitung, Analyse und Visualisierung, Repositorien zu ihrer Bereitstellung, zitierfähigen Veröffentlichung und ggf. Replikation, schließlich Archivsysteme zu ihrer dauerhaften Verfügbarmachung.

# 3 Handlungsfelder

Bei der Planung technischer Lösungen für das Forschungsdatenmanagement ist das jeweilige Handlungsfeld und seine Rahmenbedingungen zu berücksichtigen.

*Disziplinäre oder generisch-interdisziplinäre Lösung?*

Von einer disziplinären Lösung werden die stärkere Anbindung an in der jeweiligen Fachcommunity bereits etablierte Systemkomponenten und Arbeitsumgebungen erwartet. Aktuelle Beispiele für disziplinäres Forschungsdatenmanagement sind in der Biotechnologie Datenbanken für Genomsequenzen[1] oder sozialwissenschaftliche Studien[2]. Ein generisch-interdisziplinärer Lösungsansatz wie z.B. das Datenrepository Zenodo[3] hingegen zielt zum einen auf die Integration der Datenbestände von verschiedenen Disziplinen, und zum anderen auf zentralisierte Dienste und interdisziplinäre Zugänge wie z.B. die Suche nach oder die Veröffentlichung von zitierfähigen Datenpublikationen.

*"Make or buy?"*

Neben der Entwicklung und dem Betrieb einer eigenen lokalen Lösung ist auch an die Möglichkeit eines Outsourcing zu denken, wie es z.B. die Firma Amazon mit ihren Web Services bzw. S3 anbietet[4]. Community-spezifische Lösungen können jedoch einen erhöhten Anteil an Anpassung und Weiterentwicklung von Forschungsumgebungen bedeuten – mithin weitergehende Administrationsrechte, die kommerzielle Provider dann nur noch auf dedizierten Root-Servern zulassen, ohne solche Entwicklungen selbst anzubieten oder durchzuführen.

*"Tiefe" des FD-Managements: Publizieren und/oder Infrastruktur?*

Abhängig vom grundsätzlichen Anforderungsprofil kann eine technische Lösung für Forschungsdatenmanagement ganz unterschiedlich ausfallen: Geht es allein oder primär um die formale Veröffentlichung und Verbreitung von Daten, bieten sich hierfür eigene lokale Repository-Instanzen, ferner Datenplattformen wie Zenodo oder Figshare an. Geht es dagegen oder darüber hinaus um die technische Unterstützung des gesamten Datenlebenszyklus von der Erzeugung, Ablage, Bearbeitung, Analyse, Visualisierung bis hin zur Replikation oder allgemein Prozessierung der (Roh-)Daten mit den hierfür vorgesehenen Programmen und Laufzeitumgebungen, können weitergehende Infrastrukturlösungen nötig werden wie z.B. Virtuelle Maschinen oder Container, die in dafür vorgesehenen geschützten Umgebungen laufen.

[1] Genome Data Viewer, https://www.ncbi.nlm.nih.gov/genome/gdv.
[2] Inter-university Consortium for Political and Social Research (ICPSR), https://www.icpsr.umich.edu/icpsrweb.
[3] Zenodo, https://zenodo.org.
[4] AWS Research Cloud Program, https://aws.amazon.com/de/government-education/research-and-technical-computing/reserach-cloud-program.

*Datenvolumen:(Zwischen-)Speicher und Bandbreiten zur Übertragung und Verarbeitung größerer Datenmengen?*

Gerade in den experimentellen Disziplinen, aber auch bei Simulationen und Modellberechnungen fallen häufig große Datenmengen an, die in lokalen Umgebungen – wie dem Arbeitsplatzrechner eines Forschenden, oder Servern des Lehrgebiets oder des örtlichen Rechenzentrums – in vertretbarer Zeit gar nicht verarbeitet werden können. Hier ist an dedizierte hard- und softwaretechnische Infrastrukturkomponenten zu denken, die ggf. Ausschreibungen und Beschaffungen beinhalten können.

# 4 Beteiligte Personen und Einrichtungen

Abgesehen von den oben genannten Interessengruppen sind üblicherweise die folgenden Akteure und Rollen am Aufbau und Betrieb eines operativen Forschungsdatenmanagements beteiligt:

- Forschungsdatenmanagerinnen und -manager
  Neben fachlichen, organisatorischen und kommunikativen Fähigkeiten und Kenntnissen weist dieser Personenkreis auch ein technisches Profil auf: Er muss zumindest vom Grundsatz her wissen, welche technischen Lösungen bereits im Einsatz sind oder sich als Optionen bieten. Dazu zählen z.B. Plattformen und Architekturen, aber auch das prinzipielle Wissen um die IT-betrieblichen Anforderungen an ein nachhaltiges Forschungsdatenmanagement.
- Projektleiter und -leiterinnen
  Die Einführung eines technischen Systems zum Forschungsdatenmanagement verläuft typischerweise im Rahmen eines Projekts, also eines Vorhabens, das „im Wesentlichen durch Einmaligkeit der Bedingungen in ihrer Gesamtheit gekennzeichnet ist.“[5] Entsprechend haben Projektleiter und Projektleiterinnen in erster Linie die Aufgabe, die zeit-, kosten- und qualitätsgerechte technische Lösung für ein Forschungsdatenmanagement zu planen und ggf. (in Zusammenarbeit mit der Forschungsdatenmanagerin) zu konzipieren, zu organisieren und zu steuern. Dabei können Forschungsdatenmanagement und Projektleitung – je nach Situation und Aufgabenstellung – in Personalunion durchgeführt werden, wobei es sich um grundsätzlich verschiedene bzw. komplementäre Rollenprofile handelt.
- Softwareentwickler und -entwicklerinnen
  Der Aufbau eines eigenen lokalen technischen Forschungsdatenmanagements beinhaltet in aller Regel softwaretechnische Entwicklungsaufgaben, auch wenn man sich auf bereits vorhandene Standardkomponenten stützen kann und sollte. Angefangen bei systemtechnischen Installationen und Konfigurationen, über

[5] DIN Normenreihe: DIN 69901-01:2009-01, Berlin: Beuth.

die Anpassung von Benutzerschnittstellen, (Meta-)Datenformaten, Dokumenttypen und Workflows bis hin zu der Entwicklung von Daten- und Programmierschnittstellen (Application Programming Interfaces) impliziert dies typische Arbeitsgänge, die über die initiale Entwicklung und Einrichtung einer softwaretechnischen Lösung hinaus auch als dauerhaft zu betrachten sind.

- Systemadministratoren und -adminstatorinnen
  Wesentlich für eine nachhaltige technische Lösung ist schließlich die (wiederholte) Inbetriebnahme des Gesamtsystems mitsamt seinen softwaretechnischen Anpassungen und Weiterentwicklungen in Form von versionierten Ständen (Releases). Dies beinhaltet regelmäßige Aktivitäten, von der Qualitätssicherung in Form von Funktions-, Nutzer-, Integrations- und Lasttests, über das Aktualisieren und ggf. Konfigurieren der Produktionsumgebung (Deployment) bis hin zur laufenden Überwachung der Anwendungen für das Forschungsdatenmanagement (Monitoring). Darüber hinaus gibt es grundständige Infrastrukturaktivitäten wie z.B. das regelmäßige Sichern von Daten, das Management der Hardwareressourcen, oder die Virtualisierung von Anwendungen.

# 5 Handlungsempfehlungen

Beim Auf- oder Ausbau einer technischen Infrastruktur sollten grundsätzlich die folgenden Aspekte berücksichtigt werden, die je nach Vorhaben unterschiedlich gewichtet sein können.

*Anforderungs- und Umfeldanalyse*

Um möglichst sicher zu stellen, dass die eingesetzte technische Lösung den wesentlichen Zielgruppen – darunter natürlich insbesondere den Forschenden selbst – auch tatsächlich nützt, ist zunächst und typischerweise eine Anforderungsanalyse durchzuführen. Die pragmatische und eher agile Variante wäre, einen aus der bisherigen Praxis heraus entwickelten konkreten Infrastrukturvorschlag zu machen (in Form einer prototypischen Anwendung oder eines Dienstes), der an die konkreten Nutzungsbedürfnisse dann noch weiter anzupassen wäre. In jedem Fall sollte die Anforderungs- und Umfeldanalyse auf Einsichten in die (jeweilige) Praxis beim Umgang mit Forschungsdaten beruhen und dabei sowohl bestehende Praxen aufgreifen, als auch neue Zugangswege eröffnen. Hinsichtlich der Durchführung und Form einer Anforderungsanalyse bieten sich ebenfalls verschiedene Varianten an: von Interviews über Befragungen von Experten und Expertinnen bis hin zu funktionalen Spezifikationen, die bereits relativ detaillierte Vorgaben hinsichtlich der softwaretechnischen Umsetzung machen. Neben diesen fachlich getriebenen Spezifikationen sind – auch und gerade in einem betrieblich möglichst abgesicherten Forschungsdatenmanagement – nicht-funktionale Anforderungen wichtig und zu berücksichtigen, wie z.B. die allgemeine Performanz und Skalierbarkeit einer Lösung.

*Stakeholderanalyse*

Für wen errichtet man eine technische Infrastruktur zum Management von Forschungsdaten? Diese Frage und ihre Beantwortung mögen zunächst trivial erscheinen – dabei sollte man sie sich auch im weiteren Fortgang immer wieder bewusst stellen und die zwischenzeitlichen technischen Entwicklungen entsprechend abgleichen. Folgende Personengruppen kommen typischerweise bei der technischen Entwicklung einer Infrastruktur für Forschungsdaten in Betracht:

- Die Datengebenden und -nutzenden bzw. Institute als derjenige Nutzerkreis, dessen Beiträge oder Inhalte den wesentlichen Bezugspunkt für ein Forschungsdatenmanagement bilden. Dabei ist es eine durchaus offene Frage, ob diese Interessengruppe mit der technischen Infrastruktur selbst unmittelbar in Berührung kommt – denkbar ist beispielsweise auch eine einfache Ablage, während die Verwaltung und Prozessierung der Daten im Hintergrund und ggf. ausgelagert abläuft. Für den Betrieb solcher Cloud-Dienste kommen größere Universitätsrechenzentren in Frage wie das der Universität Jena[6], aber ggf. auch kommerzielle Anbieter wie Amazon Web Services[7] oder Microsoft Azure[8]. Speziell das Management größerer Volumina an Forschungsdaten erfordert finanzielle, materielle und personelle Ressourcen, die i.d.R. dann nur noch größere IT-Unternehmen oder institutionelle, im Wesentlichen staatlich finanzierte Einrichtungen wie Rechenzentren, Forschungsdatenzentren, statistische Ämter oder auch akademische Bibliotheken mit enger Anbindung an Rechenzentren aufbringen können. Auch die Forschungsförderung spielt gerade bei der Initiierung von Infrastrukturen zum Forschungsdatenmanagement eine wichtige Rolle, während ihr nachhaltiger Betrieb und ihre Weiterentwicklung wiederum eigene Finanzierungsmodelle auf Seiten der betreibenden Organisationen erfordern.
- Datenkuratoren und -kuratorinnen. Sie pflegen, überarbeiten und veröffentlichen die von Datengebenden bereitgestellten (Meta-)Daten mit Hilfe von fachspezifischen Werkzeugen und Systemkomponenten.
- Forschungsfördernde Organisationen, die ein ausgeprägtes Interesse an einer möglichst nachhaltigen und effizienten Verwaltung von finanzierten Forschungsprojekten haben, in denen regelmäßig Daten anfallen.
- Kooperationspartner und -partnerinnen, die gerade bei Drittmittelprojekten ihre eigene Interessenlage einbringen können.
- Technische Infrastruktureinrichtungen wie Rechenzentren, oder auch Bibliotheken mit erhöhtem Infrastrukturanteil.
- Administrative Einrichtungen im universitären Umfeld, wie z.B. Referate zur Forschungsevaluierung.

[6] FSU-Cloud, https://www.uni-jena.de/FsuCloud.html.

[7] AWS Research Cloud Program, https://aws.amazon.com/de/government-education/research-and-technical-computing/reserach-cloud-program.

[8] Microsoft Azure for Research, https://www.microsoft.com/en-us/research/academic-program/microsoft-azure-for-research.

*Sichtung und Analyse der am Markt befindlichen Lösungen*

Vor dem Hintergrund einer systematischen Anforderungsanalyse sind technische und auch IT-betriebliche Lösungen zu sichten, wobei sich hier prinzipiell eine Bandbreite bietet: Von kompletten in-house-Eigenentwicklungen über die lokale Installation und Nachnutzung von bestehenden Komponenten bis hin zu gehosteten Lösungen sowohl für die Programme („Software-as-a-Service“), als auch für die Daten selbst.