

# Leitfaden zur digitalen Datensparsamkeit

## 1. Einleitung

Im Zuge der Technologisierung fast aller Bereiche der Geistes- und besonders der Naturwissenschaften können in immer kürzeren Zeitabschnitten zunehmend größere Mengen an Forschungsdaten generiert werden. Hierbei sind unter Forschungsdaten alle Daten zu verstehen, die im Zuge eines wissenschaftlichen Vorhabens z. B. durch Quellenforschungen, Experimente, Messungen, Erhebungen oder Simulation entstehen. Gleichzeitig steigen die Erwartungen an eine Archivierung und Publikation von Daten durch Forschungsförderer, wie sie beispielsweise die Deutsche Forschungsgemeinschaft in ihrem Leitfaden zur guten wissenschaftlichen Praxis darlegt.<sup>1</sup> Die Open-Science-Bewegung strebt an, alle Strategien, Verfahren und Bestandteile des wissenschaftlichen Prozesses offen zugänglich und nachnutzbar zu machen und dadurch der Wissenschaft, Gesellschaft und Wirtschaft neue Möglichkeiten im Umgang mit wissenschaftlichen Erkenntnissen zu eröffnen.<sup>2</sup> Dies umfasst auch die Bereitstellung und Nachnutzung von Daten.

Es lassen sich momentan vielfältige Entwicklungen hinsichtlich der wachsenden Gewinnung, Verfügbarkeit, Transparenz und Nachnutzbarkeit von Forschungsdaten beobachten, was in vielerlei Hinsicht zu begrüßen ist. Forschungsdatenmanagement nach den FAIR-Prinzipien wird zum Standard, die Publikation von Daten immer mehr zur Selbstverständlichkeit und datengetriebene und datenunterstützte Forschung findet sich in allen wissenschaftlichen Disziplinen. Allerdings bringen diese Entwicklungen technische und organisatorische Herausforderungen mit sich. Hierzu zählt ein Mehr an erforderlichen Ressourcen in allen Phasen des Forschungsdatenlebenszyklus, sowohl in technischer wie in personeller Hinsicht. Darüber hinaus können juristische und ethische Herausforderungen entstehen.<sup>3</sup>

Dieser Leitfaden soll daher praktische Ansätze liefern, wie durch digitale Datensparsamkeit<sup>4</sup> notwendige technische und organisatorische Ressourcen reduziert werden können. Diese praktischen Ansätze werden ergänzt durch eine Checkliste zum Überprüfen eigener Daten auf Sparpotentiale. Wesentlich ist, dass es sich hierbei um Empfehlungen handelt. Die spezifische Umsetzung muss fallbezogen in den einzelnen Fachdisziplinen, durch Festlegungen in Forschungsprojekten und ihren Forschenden erfolgen. Weiterhin können frühzeitig angestellte Überlegungen zum Thema Datensparsamkeit während des gesamten Forschungsprojektes Vorteile bringen, die im Folgekapitel beschrieben werden. Wichtig ist, dass die Eignung jeder Empfehlung immer von der individuellen

---

<sup>1</sup> Deutsche Forschungsgemeinschaft: Guidelines for Safeguarding Good Research Practice. Code of Conduct, 20.04.2022, <https://doi.org/10.5281/ZENODO.6472827>.

<sup>2</sup> AG Open Science: Mission Statement, [ag-openscience.de](https://ag-openscience.de/mission-statement/), Oktober 2014, <https://ag-openscience.de/mission-statement/>, Stand: 15.03.2024.

<sup>3</sup> Hillegeist, Tobias: Rechtliche Probleme der elektronischen Langzeitarchivierung wissenschaftlicher Primärdaten, Göttingen 2012 (Göttinger Schriften zur Internetforschung 8). Online: <https://doi.org/10.17875/gup2012-142>; Rösch, Hermann: 1.5 Forschungsethik und Forschungsdaten, in: Putnings, Markus; Neuroth, Heike; Neumann, Janna (Hg.): Praxishandbuch Forschungsdatenmanagement, Berlin; Boston 2021, S. 115–140. Online: <https://doi.org/10.1515/9783110657807-006>.

<sup>4</sup> Der Begriff stammt ursprünglich aus dem Datenschutz und beschreibt, dass für die Übernahme von Daten nur die verwendet werden sollen, die für den jeweiligen Zweck notwendig sind, statt einfach komplette Datensätze zu kopieren.

Maximilian Heber, Moritz Jakob, Matthias Landwehr, Jan Leendertse, Maximilian Müller, Gabriel Schneider, Dirk von Suchodoletz, Robert Ulrich

Forschungs- und Datensituation sowie dem Zeitpunkt innerhalb des Datenzyklus abhängt und dahingehend geprüft werden muss. Ein erster guter Ansatzpunkt dafür sind Datenmanagementpläne.<sup>5</sup>

## 2. Motivation digitaler Datensparsamkeit

Die Gründe für Sparsamkeit im Umgang mit Daten sind vielfältig, speisen sich insgesamt vor allem aus einer starken Zunahme der absoluten Datenmenge. So erzeugen beispielsweise bildgebende Verfahren und Videos mit höherer Auflösung größere Datenmengen als solche mit geringer Auflösung. Auch die stetig steigend verfügbare Rechenleistung für Simulationen und wissenschaftliches Rechnen erzeugt immer detailliertere Ergebnisse und damit mehr Daten.

Ein weiterer Grund liegt in gestiegenen Anforderungen an die Aufbereitung von Forschungsdaten zur Veröffentlichung und Archivierung. Hierfür haben sich im Forschungsdatenmanagement die FAIR-Prinzipien als Standard etabliert.<sup>6</sup> Diese sorgen durch konkrete Anforderungen dafür, dass Daten gut auffindbar (Findable), zugänglich (Accessible), interoperabel (Interoperable) und nachnutzbar (Reusable) sind. Um diese Prinzipien zu erfüllen, müssen Daten detailliert mit Metadaten versehen und zur Publikation in geeigneten Repositorien aufbereitet werden. Beides vergrößert die zu verwaltende Datenmenge.

Aufgrund einer oft fehlenden kohärenten und ausgewogenen Backup-Strategie werden Daten im Laufe ihres Lebenszyklus häufig unnötig mehrfach gespeichert. Dies kann auch daran liegen, dass Forschende den jeweiligen Speichersystemen nur bedingt vertrauen oder die von zentralen Services bereits angebotenen Redundanzen und Absicherungen nicht bekannt sind.

Die Gründe für eine Begrenzung der Datenmenge lassen sich in vier Kategorien unterteilen.

### 2.1. Organisatorische Gründe

Die Verwaltung großer Datenmengen ist aufwändig: Im Forschungsdatenmanagement müssen Datensätze versioniert, benannt, beschrieben, sortiert und gespeichert werden. Steigt die Menge, so leidet darunter die Übersicht über die eigenen Daten. Wenn Daten darüber hinaus extern geteilt werden und dadurch zusätzliche Versionen in die eigene Datenhaltung hinzukommen, kann das die Unübersichtlichkeit verstärken. Ein Fokus auf die eigentliche Forschung, deren Replikation und die zur Dokumentation notwendigen Daten, ist somit erschwert.

---

<sup>5</sup> Hausen, Daniela; Favella, Gianpiero; Fingerhuth, Matthias u. a.: Datenmanagementpläne in der Forschung – von Grundlagen zu Grundfragen, in: Bausteine Forschungsdatenmanagement 2022 (1), S. 103–120. Online: <https://doi.org/10.17192/BFDM.2022.1.8366>; Leendertse, Jan; Mocken, Susanne; Von Suchodoletz, Dirk: Datenmanagementpläne zur Strukturierung von Forschungsvorhaben, in: Bausteine Forschungsdatenmanagement 2019 (2), S. 4–9. Online: <https://doi.org/10.17192/BFDM.2019.2.8003>; Strötgen, Robert: Bedarfsplanung eines institutionellen Repositoriums für Forschungsdaten, in: Bausteine Forschungsdatenmanagement 2020 (2), S. 112–120. Online: <https://doi.org/10.17192/BFDM.2019.2.8106>.

<sup>6</sup> Wilkinson, Mark D.; Dumontier, Michel; Aalbersberg, IJsbrand Jan u. a.: The FAIR Guiding Principles for scientific data management and stewardship, in: Scientific Data 3 (1), 15.03.2016, S. 160018. Online: <https://doi.org/10.1038/sdata.2016.18>.

## 2.2. Technische Gründe

In manchen – überwiegend naturwissenschaftlichen – Arbeitsgruppen werden Daten im Petabyte-Bereich erzeugt und verwaltet. Für diese Daten werden ausreichend dimensionierte Speichersysteme benötigt,<sup>7</sup> idealerweise abgesichert durch eine Spiegelung oder ein Backup gleicher Größe. Im Unterschied zum wissenschaftlichen Rechnen wird die Ressource nach Projektabschluss nicht automatisch wieder frei. Neben der reinen Speicherung muss auch der Datentransfer berücksichtigt werden: Technische Systeme müssen dazu in der Lage sein, z.B. durch HD-Kameras erzeugte Bilddaten in Echtzeit zu verarbeiten; im späteren Verlauf des Lebenszyklus der Daten müssen entsprechende Datenmengen zu Computing-Systemen oder in Repositorien verschoben werden.

Der Erwerb, Betrieb und die regelmäßige Erneuerung von Netzwerk- und Speicherinfrastrukturen erfordern dabei technische, personelle und finanzielle Ressourcen.<sup>8</sup> Auch die Migration zwischen Speichersystemen muss berücksichtigt werden. Allein bei einer Vorhaltezeit von zehn Jahren nach Projektende bzw. Publikation von Daten bedeutet dies fast zwei Lebenszyklen<sup>9</sup> eines Speichersystems. Häufig ist nicht geklärt, wie diese Ressourcen aufgebracht werden können, die oft weit nach dem Ende von Projekten oder allgemeinen Forschungsvorhaben benötigt werden. Die Grundausstattung von Universitäten und anderen wissenschaftlichen Einrichtungen kann dies nicht im vollen Umfang leisten.

## 2.3. Datenschutz und -ethik

Der Umgang mit datenschutzrechtlichen Regelungen sowie Fragen der Datenethik ist Usus im Management von Forschungsdaten und deshalb in Datenmanagementplänen<sup>10</sup> zu berücksichtigen. Insbesondere beim Umgang mit personenbezogenen oder anderweitig sensiblen Daten sind geeignete Schutzmaßnahmen Routine. Ethische Fragen berücksichtigen, ob man alles speichern und verarbeiten sollte, auch wenn es keine rechtlichen Einschränkungen gibt. Kommt die Abwägung im Einzelfall zu einem entsprechenden Ergebnis, werden ebenfalls Schutzmaßnahmen ergriffen. Doch je stärker die Datenmenge und damit die Komplexität der Speicherverwaltung wächst, desto größer werden auch der Aufwand und die Prüfungen, ob die Daten noch von den Schutzmaßnahmen erfasst sind. Speicherort und Zugriffsrechte sollten entsprechend datenschutzkonform gewählt werden. Der mit der Datenmenge wachsenden Gefahr, eine Version oder Kopie von schützenswerten Daten zu übersehen, kann mit einer Reduzierung der Datenmenge begegnet werden: Daten, die nicht erhoben und nicht gespeichert werden, können nicht in falsche Hände geraten.

## 2.4. Nachhaltigkeit

Kernaspekt der Nachhaltigkeit ist ein möglichst ressourcenschonender Umgang auf allen Ebenen: Die Bereitstellung und der Betrieb von Netzwerk-, Rechen- und Speicherinfrastrukturen verbraucht eine enorme Menge an Rohstoffen und Energie mit entsprechend negativer CO<sub>2</sub>-Bilanz. Das Ziel des verantwortungsvollen Ressourcenumgangs ist durch das Schlagwort „Green IT“ längst in der

---

<sup>7</sup> Strötgen: Bedarfsplanung, 2020.

<sup>8</sup> Leendertse, Jan; Von Suchodoletz, Dirk: Kosten und Aufwände von Forschungsdatenmanagement, in: Bausteine Forschungsdatenmanagement 2020 (1), S. 1–7. Online: <https://doi.org/10.17192/BFDM.2020.1.8246>.

<sup>9</sup> Speichersysteme werden üblicherweise über einen Zeitraum von fünf bis sieben Jahren abgeschrieben.

<sup>10</sup> Leendertse et al: Datenmanagementpläne, 2019.

Maximilian Heber, Moritz Jakob, Matthias Landwehr, Jan Leendertse, Maximilian Müller, Gabriel Schneider, Dirk von Suchodoletz, Robert Ulrich

Datenverarbeitung angekommen. Mit einer Beschränkung der Datenmenge kann der ökologische Fußabdruck von Forschungsvorhaben verbessert werden.

### 3. Umsetzung digitaler Datensparsamkeit

Auch wenn sich die grundsätzliche Sinnhaftigkeit der Datensparsamkeit schnell erschließt: Die praktische Umsetzung stellt Forschende vor einige Herausforderungen. In der Folge werden typische Probleme und denkbare Lösungen skizziert. Darüber hinaus wird ein Weg aufgezeigt, wie Datensparsamkeit im Forschungsdatenlebenszyklus praxisnah mitgedacht werden kann. Zur Vervollständigung findet sich eine Checkliste zum Überprüfen von Daten auf Sparpotentiale.

#### 3.1. Herausforderungen bei der Umsetzung

**Entscheidungsbefugnis:** Es muss klar sein, wer die Befugnis hat, über die Verwendung der Daten zu entscheiden. Formal liegt dieses Weisungsrecht in der Regel bei der Leitung eines Forschungsprojektes, doch die wissenschaftliche Realität ist weniger eindeutig. Die eigentliche Arbeit und damit auch das Wissen über die Bedeutung der Daten liegt häufig bei anderen Beteiligten. Weiterhin muss man Urheber-, Nachnutzungs- und ggf. Eigentumsrechte an Daten berücksichtigen. Deshalb ist eine frühzeitige, gemeinsame und schriftlich dokumentierte Entscheidung mit klaren Rollenverteilungen zu empfehlen.<sup>11</sup>

**Endgültigkeit:** Gelöschte Daten sind meist unwiederbringlich verloren. Auch wenn es in speziellen Fällen technische Möglichkeiten gibt, gelöschte Daten wiederherzustellen, so sind diese aufwändig, ohne garantierten Erfolg und vor allem bei größeren Datenmengen kaum praktikabel. Diese Irreversibilität erfordert eine durchdachte und bewusste Entscheidung zur Löschung von Daten. Solche Entscheidungen sollten keineswegs leichtfertig oder gar unter Zeitdruck getroffen werden.

**Datenformate:** Diese unterscheiden sich in ihrer Offenheit, der langfristigen Zugreifbarkeit sowie hinsichtlich des Platzbedarfs. Daten zu komprimieren, trägt zu einem geringeren Speicherbedarf bei, allerdings muss immer gewährleistet sein, dass Daten auch in Zukunft wieder dekomprimiert werden können. Zudem sind Verfahren mit einer hohen Kompressionsrate bei Bildern oder Videos fast immer mit einem Qualitätsverlust verbunden. Es gilt folglich, Formate zu nutzen, die verlustfrei oder mit vertretbarem Informationsverlust arbeiten.<sup>12</sup>

**Technische Hürden:** Speichersysteme sind für Nutzende selten vollständig transparent. So ist bei Cloud-Systemen in der Regel nicht festzustellen, ob eine oder mehrere interne Kopien der Daten auf verteilten Servern abgelegt sind und ob und wie lange z.B. gelöschte Dateien noch in Backup-Systemen vorgehalten werden. Dies ist insbesondere dann problematisch, wenn man bei sensiblen Daten deren Löschung zusagt. Hier sind auch die Betreibenden von Diensten gefordert, eine entsprechende Transparenz herzustellen bzw. mit Angeboten wie einer „Garbage Collection“ technische Unterstützung anzubieten. Das Wissen über die Speicherredundanz erlaubt es Forschenden, qualifiziert zu entscheiden, ob sie Kopien ihrer Daten nochmal an anderer Stelle vorsehen (oder nicht).

---

<sup>11</sup> Entsprechende Regelungen können über institutionelle Policys (Open Science Policy oder Forschungsdatenmanagementpolicy) oder individuelle Vereinbarungen getroffen werden.

<sup>12</sup> Redaktion forschungsdaten.info: Formate erhalten, forschungsdaten.info, <https://forschungsdaten.info/themen/veroeffentlichen-und-archivieren/formate-erhalten/>, Stand: 15.03.2023.

Machbarkeit: Forschung ist sehr individuell und damit auch das Management der Forschungsdaten. Das bezieht sich auch auf die Möglichkeit, Daten einzusparen oder im Forschungsverlauf frühzeitig zu löschen. Daten, die während verschiedener Schritte des Forschungsprozesses anfallen, unterscheiden sich stark im Aufwand ihrer Reproduzierbarkeit. Daraus folgen unterschiedliche Möglichkeiten, Daten einzusparen oder sie im Forschungsverlauf frühzeitig zu löschen. Beispielsweise können aggregierte Datensätze oder textanalytische Daten schnell und einfach reproduziert und daher frühzeitig gelöscht oder gar nicht erst gesondert abgespeichert werden. Dem gegenüber stehen z.B. Rohdaten und Ergebnisdaten zeitintensiver Simulationen, die aufgrund der aufwändigen und zeitintensiven Reproduzierbarkeit kaum Einsparpotenzial bieten. Deshalb kann es nur rudimentäre pauschale Regelungen zur Datensparsamkeit geben, in allen Fällen sind Einzelfallbetrachtungen mit individuellen Lösungen anzuraten. Sinnvollerweise greift man hierzu auf die Empfehlungen z.B. der Nationalen Forschungsdateninfrastruktur (NFDI) für das Fachgebiet zurück und berät sich mit entsprechend qualifizierten Data Stewards.

Bewusstsein: Digitale Daten sind oft nicht sichtbar. Im Gegensatz zu Aktenordnern, die im Büroregal stehen, sind sie eine visuell nicht fassbare Menge, die sich im Computer bzw. auf entfernten IT-Infrastrukturen oder in einer Cloud befindet. Es ist sinnvoll, soweit möglich automatisierte Erinnerungsroutinen anzulegen, die beispielsweise an den Ablauf von Vorhaltefristen erinnern und Datensätze zur erneuten Entscheidung über den Verbleib vorlegen.

### 3.2. Umsetzung im Vorfeld eines Forschungsvorhabens

Der Prozess zur Datensparsamkeit ist von vielen Faktoren abhängig und muss ganzheitlich und individuell auf das eigene Forschungsvorhaben angewendet werden. Ebenso ist es sinnvoll, schon in der Planungsphase zu bestimmen, welche Daten man bis zum Projektende aufbewahren und auf welche man frühzeitig verzichten will. Zudem sollte schon zu Beginn überlegt werden, wie viele Kopien und Versionen von Daten für eine nachvollziehbare Forschung benötigt werden.

Diese Überlegungen lassen sich sehr gut in die allgemeine Datenplanungsphase aufnehmen. Bereits jetzt verlangen viele Forschungsförderorganisationen das Erstellen von Datenmanagementplänen (DMP).<sup>13</sup> Doch auch unabhängig davon empfehlen sich DMPs für alle Forschungsprojekte: Der DMP – als lebendes Dokument konzipiert – beinhaltet eine detaillierte Übersicht, welche Daten im Projektverlauf anfallen, wem welche Daten zur Verfügung stehen müssen, welche technische Infrastruktur genutzt werden soll und welche Daten am Ende wie publiziert und/oder archiviert werden sollen. Dabei sollte immer erwogen werden, ob Daten dauerhaft benötigt oder wieder gelöscht werden können.

Ein im DMP schriftlich niedergelegtes Lösch- und Rollenkonzept hat überdies den Vorteil, dass es für alle Beteiligten klar und jederzeit einsehbar ist. Damit wird auch über das Projektende hinaus definiert, wer über welche Daten wie entscheiden darf.

### 3.3. Umsetzung im Verlauf eines Forschungsvorhabens

Sollte das Forschungsprojekt bereits begonnen haben und ggf. kein DMP existieren, können die notwendigen Arbeiten in Verbindung mit einem Konzept zur Datensparsamkeit nachträglich erfolgen.

---

<sup>13</sup> Leendertse et al.: Datenmanagementpläne, 2019.

Maximilian Heber, Moritz Jakob, Matthias Landwehr, Jan Leendertse, Maximilian Müller, Gabriel Schneider, Dirk von Suchodoletz, Robert Ulrich

Beides erfordert, wie zu Beginn eines Projekts, eine ganzheitliche Betrachtung der Datensituation, aus der sich geeignete Maßnahmen ableiten lassen.

Die folgende Checkliste zur Überprüfung der Daten auf Sparpotentiale kann bereits vor bzw. beim Erheben der Daten benutzt werden. Mit geringem Aufwand können Vorkehrungen getroffen werden, die im weiteren Verlauf des Forschungsdatenlebenszyklus die Einsparung von Speicherplatz erleichtern.

Bei der Erhebung der Daten:

- Erfolgt die Speicherung der Daten in einer geordneten und nachvollziehbaren Verzeichnisstruktur?<sup>14</sup>
- Ist eine Beschreibung der Daten mit Metadaten vorhanden?<sup>15</sup>
- Sollen nur zu Test- oder Übungszwecken erzeugte Daten in einem separaten Ordner gespeichert werden?
- Werden Daten beispielsweise aus Methodeneinstellungen oder der Kalibrierung von Instrumenten, sofern sie für die spätere Beurteilung der Datenqualität oder für die weitere Arbeit nicht mehr benötigt werden, gelöscht?

In kurzen zeitlichen Abständen (z.B. monatlich):

- Wurden ältere Daten durch neue Daten oder Versionen redundant oder obsolet?
- Gibt es Zwischenergebnisse von beispielsweise längeren Prozess-Pipelines, die sich leicht aus anderen Daten reproduzieren lassen?
- Sind die Daten für eine Publikation geeignet (Datenqualität, Reproduzierbarkeit, Nachvollziehbarkeit)?
- Können die Daten für andere Forschende/Studierende von Nutzen sein?
- Sind die Daten nach Abschluss des Projekts noch in irgendeiner Weise von Nutzen oder können es zukünftig sein?

In längeren zeitlichen Abständen (z.B. halbjährlich):

- Wie viele Kopien bestehen von erhobenen Daten (auf derselben oder anderen Ressourcen)?
- Befinden sich erhobene Daten in akutem Gebrauch (Möglichkeit zur Archivierung)?
- Haben sich die Rechte an bestimmten Daten zwischenzeitlich geändert?
- Wurden die Daten zwischenzeitlich in einem Repositorium veröffentlicht und darüber zugänglich gemacht?

Bei diesen Überlegungen und Vorgängen können Data Stewards die entsprechenden Rahmen schaffen und beim laufenden Vorhaben qualifiziert unterstützen. Auch sind aus der NFDI heraus Standardisierungen, Regelwerke und Best Practices zu erwarten.

## 4. Empfehlungen

Dieser Beitrag bietet allgemeine Empfehlungen, wie eine gelebte Datensparsamkeit aussehen kann. Alle Empfehlungen sollten auf die individuelle bzw. projektspezifische Situation hin überprüft und ggf. angepasst werden.

---

<sup>14</sup> Redaktion forschungsdaten.info: Datenorganisation, [forschungsdaten.info](https://forschungsdaten.info/themen/organisieren-und-aufbereiten/datenorganisation/), <https://forschungsdaten.info/themen/organisieren-und-aufbereiten/datenorganisation/>, Stand: 15.03.2024.

<sup>15</sup> Ebd.

#### 4.1. Grundsätze

- Wie viele Kopien der Datensätze werden benötigt? Über welche Sicherungs- und Backup-Funktionen verfügen die genutzten Speichersysteme? Ein automatisches Backup eines Cloud-Dienstes kann z.B. eine eigene Kopie auf einem anderen Speichersystem unnötig machen.
- Wenn Daten von anderen Systemen für die eigene Forschung aus externen Quellen bezogen werden, reicht für die lokale Speicherung einfache Redundanz aus.<sup>16</sup>
- Ist es sinnvoll, Daten auf verschiedenen Systemen zu publizieren, beispielsweise einmal auf einem international anerkannten Publikationsserver und zum anderen auf dem lokalen Repositorium der eigenen Universität? Dies kann als Absicherung notwendig sein, wenn die Nachhaltigkeit des externen Publikationssystems nicht sichergestellt ist oder man befürchtet, dass die Daten perspektivisch hinter einer Paywall verschwinden.
- In welcher Auflösung werden die Resultate datenerzeugender Verfahren benötigt, kann die Anwendung von Kompressionsalgorithmen sinnvoll sein? Eine geringere Auflösung reduziert den Speicherbedarf signifikant.
- Wie soll der Umgang mit älteren Daten sein, welche durch neue Daten oder Versionen redundant bzw. obsolet wurden.
- Ist es zur Nachvollziehbarkeit notwendig, in jedem Zwischenschritt der Datenverarbeitung den vollständigen Datensatz zu speichern? Wenn die getätigten Schritte dokumentiert und auf den Originaldatensatz erneut angewendet werden können, können die Daten im Zwischenzustand eingespart werden.
- Bei der Arbeit mit Teilmengen von Daten – z.B. bei Rechnungen oder Simulationen – muss abgewogen werden, ob jede Teilmenge einzeln gespeichert werden muss oder ob die Teilmenge mit geringem Aufwand erneut aus den Daten generiert werden kann.
- Beim Teilen von Daten mit Projektbeteiligten ist festzulegen, ob diese immer eine vollständige Kopie bekommen oder ob man gemeinsam Daten an einer zentralen Stelle nutzt.
- Wie soll mit unbrauchbaren Daten umgegangen werden? Werden bei Simulationen oder Datenverarbeitung Resultate erzielt, die für die weitere Forschung nicht verwendet werden können oder sind Messungen invalide, muss entschieden werden, ob diese Daten direkt entfernt oder zu Dokumentationszwecken aufbewahrt werden.

Datensparsamkeit ist in diesem Sinne keine rein speicherkapazitätsbezogene Frage, sondern sollte auch vor dem Hintergrund ethischer Fragestellungen und Datenschutz sowie der eigenen vereinfachten Arbeit mit den Daten zu einem Kernelement aller Forschungsprojekte werden.

#### 4.2. Praxisbeispiele

##### Beispiel 1: Naturwissenschaften

In der fiktiven Arbeitsgruppe Mayer, für Umweltanalytik werden täglich Massenspektrometriedaten aus Umweltproben erhoben, welche sich häufig im zweistelligen Gigabyte-Bereich befinden. Diese umfassen Daten aus Methodenentwicklung, Testmessungen und den eigentlichen Messungen, welche letztlich für die Publikation wissenschaftlicher Abhandlungen gedacht sind. Damit die limitierten Speicher-Ressourcen nicht überbeansprucht werden, werden zunächst alle Daten in einem nach Projekten geordnetem Verzeichnis abgespeichert. Testmessungen und weitere nicht zur Archivierung vorgesehene Daten (temporäre Dateien) werden in einem gesonderten Ordner abgespeichert, welcher

---

<sup>16</sup> Wenn das lokale Speichersystem die Möglichkeit bietet, festzulegen, mit welcher Redundanz gespeichert werden soll.

Maximilian Heber, Moritz Jakob, Matthias Landwehr, Jan Leendertse, Maximilian Müller, Gabriel Schneider, Dirk von Suchodoletz, Robert Ulrich

regelmäßig gelöscht wird. Allen Messdaten wird eine Beschreibung der verwendeten Messmethode (e.g., Methodendatei der Instrument-Software, .txt-Datei) beigefügt, sodass diese auch reproduzierbar und replizierbar sind. Ebenfalls wird den Messdaten eine .txt-Datei beigefügt, welche sowohl den Kontext der Messung wiedergibt als auch weitere Informationen (e.g., Probenahmen, Probenaufbereitung, messende Personen, etc.) enthält. Zusätzlich sind alle Mitarbeitenden von Frau Prof. Mayer dazu angehalten, Messdaten, welche durch neue Messungen oder Erkenntnisse redundant oder obsolet wurden, unverzüglich zu löschen. In monatlichen Abständen löscht Frau Mayer den Ordner für temporäre Dateien und bittet alle Mitarbeiter\*innen gewissenhaft, ihre Daten hinsichtlich Datenqualität, Reproduzierbarkeit, Nachvollziehbarkeit und ob sie für andere Forschende zukünftig von Nutzen sein könnten, zu überprüfen und gegebenenfalls zu löschen. So behält Frau Mayer stets den Überblick über die Forschungsdaten ihrer Arbeitsgruppe und garantiert für andere Personen eine einfache Auffindbarkeit und Nachnutzbarkeit dieser Daten.

#### Beispiel 2: Geisteswissenschaften

In einem sozialwissenschaftlichen Forschungsprojekt werden dreißig leitfadengestützte Interviews mit Expertinnen und Experten von Bürgerstiftungen geführt. Ziel der Untersuchung ist eine textbasierte Auswertung der Befragungsergebnisse. Die Interviews werden online per Videokonferenz geführt und als Videodateien aufgezeichnet. Den Interviewpartner\*innen wird dabei aus Datenschutzgründen zugesichert, dass die Aufzeichnung nach erfolgter Transkription gelöscht wird. Nach Abschluss aller Interviews werden diese transkribiert und in Textdateien gespeichert. Diese dienen als Grundlage der weiteren Analyse. Die Videomitschnitte der Interviews können nach Abschluss der Transkription gelöscht werden.

#### Beispiel 3: Biodiversität

In einem Projekt zur Erforschung der Biodiversität von Vögeln erfolgt die Erhebung von Audiodaten durch Aufnahmen im Wald über einen längeren Zeitraum mit dem Ziel die Menge an vorhandenen Vögeln zu überwachen. Diese Daten sind unwiederbringlich, d.h. wenn Teile verloren gehen, lassen diese sich nicht wiederherstellen. Die Daten werden dann zur Beantwortung unterschiedlicher wissenschaftlicher Fragestellungen von mehreren Promovierenden analysiert und weiterverarbeitet. Zunächst liegen die Daten in einem unkomprimierten Audio-Format (WAV-File) vor. Nach der Verarbeitung müssen diese Rohdaten noch als Referenz zur Verfügung stehen, wobei es sich insgesamt um mehr als 100 TByte handelt, allerdings muss nicht mehr direkt auf diese Daten zugegriffen werden.

Anfänglich wurden die Daten auf ein großes iSCSI-Speichersystem gespeichert, welches über eine Virtuelle Linux-Maschine als Netzwerklaufwerk bereitgestellt wurde. Dieses iSCSI-System hat eine eingebaute RAID-Redundanz sonst aber keine weiteren Sicherungsmechanismen. Deshalb wurden die Daten nochmal komplett auf ein weiteres Speichersystem gespiegelt. Dieses hatte bereits eingebaute Sicherheitsmechanismen, welche bewirkten, dass die absolute Datenmenge durch diese Maßnahme auf ein Vielfaches anstieg. Dabei wurde weder Versionierung etc. in Betracht gezogen, sondern primär das Werkzeug *rsync* zum Datentransfer benutzt.

Nach einer Beratung zum Datenmanagement entschied sich die Projektleitung zu einem Übergang auf eine neue Speicherlösung. Alle Daten welche sich in akutem Gebrauch befinden werden fortan auf einem gesicherten zentralen Speichersystem gespeichert. Alle Daten welche sich nicht mehr in akutem Gebrauch befinden und lediglich als Referenz aufbewahrt werden sollen, werden in ein Object Storage bei höherer Redundanz und gleichzeitig hoher Speichereffizienz verschoben. Dieser Schritt führte letztlich zu einer erheblichen Abnahme der Bruttodatenmenge. In weiteren Schritten könnte eine verlustfreie Audiokompression in Betracht gezogen werden.



Maximilian Heber, Moritz Jakob, Matthias Landwehr, Jan Leendertse, Maximilian Müller, Gabriel Schneider, Dirk von Suchodoletz, Robert Ulrich

Das Papier wird vom Arbeitskreis der Leiterinnen und Leiter der wissenschaftlichen Rechenzentren in Baden-Württemberg (ALWR) sowie von der AG der Direktorinnen und Direktoren der Universitäts- und Landesbibliotheken Baden-Württembergs (AGBibDir) unterstützt.

**Autor\*innen:**

*Maximilian Heber, Universität Konstanz, <https://orcid.org/0000-0003-3399-7532>*

*Moritz Jakob, Universität Konstanz, <https://orcid.org/0009-0007-0772-3462>*

*Matthias Landwehr, Universität Konstanz, <https://orcid.org/0000-0001-9274-2578>*

*Jan Leendertse, Universität Freiburg, <https://orcid.org/0000-0001-5676-493X>*

*Maximilian Müller, Universität Konstanz, <https://orcid.org/0000-0003-2237-1147>*

*Gabriel Schneider, ZB MED. Informationszentrum Lebenswissenschaften, <https://orcid.org/0000-0001-6573-3115>*

*Dirk von Suchodoletz, Universität Freiburg, <https://orcid.org/0000-0002-4382-5104>*

*Robert Ulrich, Karlsruher Institut für Technologie, <https://orcid.org/0000-0001-9063-2703>*

**Zitierfähiger Link (DOI):**

<https://doi.org/10.5281/zenodo.11445843>

Dieses Werk steht unter der Lizenz [Creative Commons Namensnennung 4.0 International](https://creativecommons.org/licenses/by/4.0/).