Anforderungen der Ingenieurwissenschaften an das Forschungsdatenmanagement der Universität Stuttgart – Ergebnisse der Bedarfsanalyse des Projektes DIPL-ING

Dorothea Iglezakis, Universitätsbibliothek Stuttgart
Björn Schembera, Höchstleistungsrechenzentrum Stuttgart

Zusammenfassung:
Forschungsdaten sind die Grundlage aller wissenschaftlichen Praxis und der Ausgangspunkt für alle daraus gewonnenen Erkenntnisse. Dieser Wert spiegelt sich allerdings oft nicht im Management von Forschungsdaten wider. Insbesondere in den Ingenieurwissenschaften gibt es Nachholbedarf, was das zweckgerichtete Forschungsdatenmanagement angeht, um die Daten nachnutzbar, nachvollziehbar und nachprüfbar zu machen. Die vorliegende Veröffentlichung fasst die Ergebnisse der Bedarfsanalyse des Projektes DIPL-ING zusammen, welches das Ziel hat, gemeinsam mit Ingenieurwissenschaftlerinnen und Ingenieurwissenschaftler Konzepte für das Forschungsdatenmanagement in den Ingenieurwissenschaften bereitzustellen. Anhand von konkreten Anwendungsfällen aus der technischen Thermodynamik und der Aerodynamik wurden Problembereiche und Anforderungen der Ingenieurwissenschaften an das Forschungsdatenmanagement ermittelt. Spezifische Anforderungen ergeben sich dadurch, dass die Forschung zu einem nicht unerheblichen Teil auf Software und Code beruht, der zum Teil sehr große Mengen an Roh- und auch verarbeiteten Daten generiert und weiterverarbeitet. Ziel ist es, eine sinnvolle interne wie externe Nachnutzung von Forschungsdaten zu ermöglichen. Dafür werden fachspezifische Metadatenstandards benötigt, die den Entstehungsprozess der Daten und Codes dokumentieren können und so sowohl die Suchbarkeit als auch die Verständlichkeit der Daten fördern. Zudem fehlen klare fachspezifische Richtlinien, welche Daten für welchen Zeitraum ökonomisch sinnvoll abgelegt werden sollen. Für die Veröffentlichung der Daten werden Infrastrukturen benötigt, die sowohl mit großen Datenmengen wie auch mit Software und Code umgehen können und eine Qualitäts- wie auch Zugriffskontrolle ermöglichen.

Summary:
The importance of research data as the basis of all scientific reasoning is not always reflected in their management. Particularly in the engineering disciplines there is a backlog demand in research data management to make data reusable, reproducible and verifiable. The following publication summarizes the outcomes of the requirement analysis conducted for the project DIPL-ING. This project aims to deliver concepts for data management in infrastructures, processes and life cycles in engineering. By means of concrete use cases from the fields of technical thermodynamics and aerodynamics, problems and requirements of engineering disciplines were identified and addressed. Specific requirements result from the fact that the research in these fields is mainly based on software and code which produce huge amounts of raw and analyzed data. To facilitate reasonable reuse, either internally or externally, discipline-specific metadata standards are needed to document the provenance of data and codes and ensure the retrieval as well as the comprehensibility of the data. In addition, there is a lack of discipline-specific guidelines recommending which part of the data should be archived and for what period of time. The publication of data requires infrastructures that can handle large volumes of data as well as software code and provide quality and access control.

Zitierfähiger Link (DOI): https://doi.org/10.5282/o-bib/2018H3S46-60
Autorenidentifikation: Dorothea Iglezakis, ORCID: http://orcid.org/0000-0002-8524-0569; Björn Schembera, ORCID: http://orcid.org/0000-0003-2860-6621
Schlagwörter: Forschungsdatenmanagement, Ingenieurwissenschaften, Repositorien, Anforderungen

1. Einleitung

Forschungsdaten in den Ingenieurwissenschaften unterscheiden sich in einigen Punkten von denen aus anderen Wissenschaftszweigen, gleich ist ihnen jedoch deren Relevanz im wissenschaftlichen Erkenntnisprozess: Mittels Datenerzeugung und darauf folgender Datenanalyse lassen sich Erkenntnisse gewinnen, ob sich z.B. Turbulenzen hinter einem Flügel entwickeln oder wie sich Trajektorien von Molekülen verhalten.

Generell hängt die Einstellung und Intention der Personen, die Daten verwenden, vor allem von der empfundenen Nützlichkeit der Daten, aber auch davon ab, ob Datenrepositorien zur Verfügung stehen und in wieweit eine Datennachnutzung in der Fachgemeinschaft üblich und erwünscht ist.[1]

Während es in den ähnlich arbeitenden Klimawissenschaften das anerkannte Metadatenmodell CERA-2,[2] Repositorien[3] oder Metadaten-Container wie NetCDF[4] gibt, fehlen im Ingenieurbereich einheitlichen Konzepte zur Datenverwaltung, standardisierte Metadatenschemata[5] und fachspezifische Forschungsdatenrepositorien[6] oder existieren nur in sehr spezieller Form als Prototypen oder Insellösungen (siehe z.B. zur Veröffentlichung von Turbulenz-Daten[7, 8]; MoSGrid[9] und DCMS[10] als Ausführungsumgebungen für molekulare Simulationen).

1 Joo, Yeon Kyoung; Kim, Youngseek: Engineering researchers' data reuse behaviours. A structural equation modelling approach, in: The Electronic Library 35 (6), 2017, S. 1141–1161. Online: <http://dx.doi.org/10.1108/EL-08-2016-0163>.
2 Lautenschlager, Michael; Toussaint, Frank; Thiemann, Hannes u.a: The CERA-2 data model, 1998, <https://www.pik-potsdam.de/cera/Descriptions/Publications/Papers/9807_DKRZ_TechRep15/cera2.pdf>, Stand: 07.02.2018.
3 Cerasearch, DKRZ, <https://cera-www.dkrz.de/WDCC/ui/cerasearch/>, Stand: 05.10.2018.
4 Malik, Tanu: Geobase: Indexing NetCDF Files for large-scale Data Analysis, in: Big Data Management, Technologies, and Applications, 2014 (IGI Global), S. 295–313. Online: <http://dx.doi.org/10.4018/978-1-4666-4699-5.ch012>.
5 Im Metadatenschemata-Verzeichnis der Research Data Alliance finden sich für die Ingenieurwissenschaften nur allgemeine Standards oder Standards aus Nachbarwissenschaften, die lediglich für einen sehr kleinen Teil der Ingenieurwissenschaften anwendbar sind, <http://rd-alliance.github.io/metadata-directory/standards/>, Stand: 07.02.2018.
6 Siehe im Verzeichnis re3data , <https://www.re3data.org/>, Stand: 07.02.2018.
7 Meneveau, Charles, Marusic, Ivan: Turbulence in the Era of Big Data. Recent Experiences with Sharing Large Datasets, in: Pollard, Andrew; Castillo, Luciano; Danaila, Luminita u.a. (Hg.): Whither turbulence and big data in the 21st century? Cham 2017, S. 497–507. Online: <http://dx.doi.org/10.1007/978-3-319-41217-7>.
8 Sillero, Juan A.; Jiminéz, Javier: Public Dissemination of Raw Turbulence Data, in: in: Pollard, Andrew; Castillo, Luciano; Danaila, Luminita u.a. (Hg.): Whither turbulence and big data in the 21st century? Cham 2017, S. 509–515. Online: <http://dx.doi.org/10.1007/978-3-319-41217-7_28>.
9 Krüger, Jens; Grunzke, Richard; Gesing, Sandra u.a.: The MoSGrid Science Gateway. A Complete Solution for Molecular Simulations, in: Journal of Chemical Theory and Computation 10 (6), 2014, S. 2232–2245. Online: <http://dx.doi.org/10.1021/ct500159h>.
10 Kumar, Anand; Grupcev, Vladimir; Berrada, Meryem u.a: DCMS. A data analytics and management system for molecular simulation, in: Journal of Big Data 2 (1), 2014, S. 9. Online: <https://doi.org/10.1186/s40537-014-0009-5>.

Im aus diesen Problemstellungen hervorgegangenen und vom Bundesministerium für Bildung und Forschung (BMBF) geförderten Projekt DIPL-ING[11] erarbeiten die Infrastruktureinrichtungen (Universitätsbibliothek, Technische Informations- und Kommunikationsdienste und Höchstleistungsrechenzentrum) gemeinsam mit den Instituten aus der Thermodynamik und Aerodynamik Konzepte für das Forschungsdatenmanagement (FDM) in den Ingenieurwissenschaften und Lösungen für jene zwei ingenieurwissenschaftliche Institute der Universität Stuttgart. Um die Problemlage in den Ingenieurwissenschaften besser zu verstehen, wurden in mehreren qualitativen Interviews konkrete Anwendungsfälle entwickelt, die beispielhaft den Forschungsprozess im Detail beschreiben. Schwerpunkt waren Forschungsprozesse, die auf Computer-Simulationen beruhen. Festgehalten wurden je Anwendungsfall Akteur(e), Ziel und Beschreibung des Anwendungsfalls, beteiligte Hardware- Systeme, verwendete Software und Skripten, Eingabe- und Ausgabedaten sowie genannte Problembereiche beim Forschungsdatenmanagement.

Im Folgenden werden die Ergebnisse der Bedarfsanalyse vorgestellt. Dabei wird in Kapitel 2 der Forschungsprozess in den Ingenieurwissenschaften beschrieben. In Kapitel 3 werden die von den Ingenieurinnen und Ingenieure genannten Problembereiche dargelegt, die einem geregelten Forschungsdatenmanagement entgegenstehen. Da die erhobenen Anwendungsfälle zwar einen sehr detaillierten, aber auch sehr spezifischen Blick auf zwei Fachbereiche der Universität Stuttgart bieten, wurden die ermittelten Erkenntnisse in Kapitel 4 mit den ingenieurspezifischen Daten der landesweiten Befragung bwFDMCommunities abgeglichen.[12] Aus den Problemstellungen heraus werden in Kapitel 5 Anforderungen an ein Forschungsdatenmanagement für die Ingenieurwissenschaften abgeleitet.

2. Der ingenieurwissenschaftliche Forschungsprozess – Code, Systeme und Daten

2.1. Forschungsprozess

Der Forschungsprozess läuft an den betrachteten Instituten nach einem ähnlichen Muster ab und ist in Abbildung 1 dargestellt. Die Forscherinnen und Forscher arbeiten in der Regel eigenständig an einem Projekt, das Eigenschaften oder Verhalten eines Systems betrachtet. Das System kann beispielsweise eine Box aus einzelnen Molekülen (Thermodynamik) oder auch der Luftraum um einen Helikopter (Aerodynamik) sein. Es gibt in den Ingenieurswissenschaften drei Ansätze zur Datengewinnung: experimentelle Messung, Simulation und theoretische Analyse.

11 FDM-Projekt DIPL-ING: Datenmanagement in Infrastrukturen, Prozessen und Lebenszyklen in den INGenieurwissenschaften, FKZ 16FDM008, 2018, <http://www.ub.uni-stuttgart.de/dipling>, Stand: 16.02.2018,

12 Tristram, Frank; Streit, Achim: Öffentlicher Abschlussbericht von bwFDM-Communities. Technischer Bericht, Karlsruher Institut für Technologie, 2015, <https://bwfdm.scc.kit.edu/downloads/Abschlussbericht.pdf>, Stand: 16.02.2018.

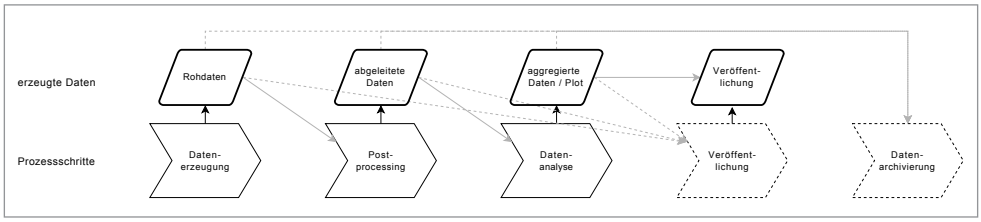

Abb. 1: Forschungsprozess in den datenintensiven Ingenieurwissenschaften

Dazu wird, ausgehend von einer Startkonfiguration, ein bestimmtes Verhalten des Systems hergestellt und beobachtet. Die Beobachtung erfolgt durch Messung von Sensoren (Experimente) oder das Herausschreiben von Systemeigenschaften in bestimmten Zeitschritten und an bestimmten Messpunkten. Dabei entstehen insbesondere bei Simulationen, teils aber auch bei experimentellen Ansätzen, sehr große Datenmengen im Gigabyte- bis Terabyte-Bereich, wohingegen die Datenmengen bei den analytischen Lösungen zu vernachlässigen sind.

Die resultierenden Daten werden anschließend weiter verarbeitet und nach interessierenden Eigenschaften gefiltert, zusammengefasst und in Grafiken (Plots) visualisiert, die in Publikationen veröffentlicht werden. Dabei werden die Daten zum Teil auch mit Daten aus anderen Quellen (Simulation, Experiment) verglichen, kombiniert oder verifiziert. Nicht alle Daten durchlaufen den gesamten Prozess. Bis zu 90 Prozent der erzeugten Roh- und Analysedaten werden nicht weiter verwendet, da sie nicht das erwartete Verhalten abbilden, fehlerhafte Annahmen oder Parameter beinhalten oder aus anderen Gründen nicht zielführend für die untersuchte Forschungsfrage sind.

Ein Projekt besteht meist aus vielen solcher Beobachtungen und mehreren resultierenden Publikationen. Nach Abschluss des Projektes werden ein Teil der Daten typischerweise auf einen institutseigenen Datenablageserver gespeichert, die umfangreichen Rohdaten entweder gelöscht oder in einem Archivsystem des jeweiligen Rechenzentrums abgelegt.

Die Verwaltung und Beschreibung der Forschungsdaten geschieht individuell ohne zentrale Vorgaben oder Richtlinien. Die Forscherinnen und Forscher legen die Daten in der Regel in einer individuellen Ordnerstruktur ab. Eine Dokumentation der Daten geschieht nur in Form von Datei- und Ordner-Bezeichnungen, seltener mittels readme-Dateien. Sie ist im Allgemeinen nicht standardisiert, sondern geschieht individuell abhängig von der Institutskultur.

Meist werden die Daten institutsintern verarbeitet, teilweise aber auch auf persönliche Anfrage mit Forschungs- oder Industriepartnern geteilt. Es gibt für die betrachteten Bereiche kein anerkanntes fachspezifisches Repositorium, in dem Forschungsdaten veröffentlicht werden können.[13] Einzelne Forscherinnen und Forscher veröffentlichen die direkt einem Plot zugrundeliegenden Daten als Appendix zusammen mit den wissenschaftlichen Publikationen. Im Bereich der technischen Thermodynamik

13 Siehe dazu z.B. das Verzeichnis re3data, <https://www.re3data.org/>, Stand: 07.02.2018.

gibt es eine größere Fachgemeinschaft, die von einer Veröffentlichung der zugrundeliegenden Daten und Skripten direkt profitieren würde. Im Bereich der Aerodynamik sind die Fragestellungen allerdings oft so speziell, dass nur eine sehr kleine Anzahl von Forscherinnen und Forschern etwas mit den Daten anfangen könnte. In beiden Bereichen sind Daten, die mehrere Jahre alt sind, nur noch in Ausnahmefällen für die aktuelle Forschung relevant. Dies gilt vor allem für Daten, die aus Berechnungen (Simulationen, theoretische Analyse) hervorgehen. Experimentell erhobene Daten können dagegen für einen längeren Zeitraum nachnutzbar bleiben.

2.2. Rechensysteme

Bei ingenieurwissenschaftlichen Forschungen, die auf Simulationen beruhen, nehmen die Rechneranlagen die Rolle des Werkzeugs ein, mit dem Daten produziert werden. Während einfache Berechnungen auf lokalen Arbeitsrechnern oder kleineren, institutseigenen Rechenclustern ausgeführt werden, benötigen umfangreiche Berechnungen Hoch- oder Höchstleistungsrechner.[14]

Höchstleistungsrechner sind auf Leistungsfähigkeit für Berechnungen hin optimiert, wohingegen die Datenverwaltung nachgeordnet ist und durch die Grenzen von POSIX-Dateisystemen nur durch die Benennung von Dateien und Verzeichnissen möglich ist.[15] Wissenschaftlerinnen und Wissenschaftler aus den Simulationswissenschaften bringen typischerweise ihre Programme zur Simulation oder Analyse in einem privaten Arbeitsbereich zur Ausführung. Dabei kann sich ein komplettes Rechenprojekt über Monate erstrecken. Nach Ablauf der Projektlaufzeit werden die Daten im Arbeitsbereich gelöscht und der Account geschlossen.

2.3. Daten und Code

Da bei Simulation auch die Erzeugung, in jedem Fall aber die Analyse und Visualisierung der Forschungsdaten auf Software und Skripten beruht, spielt diese eine besonders wichtige Rolle. Zum Einsatz kommen sowohl öffentlich frei verfügbare OpenSource-Pakete, die von einer Gemeinschaft gepflegt werden und auch versioniert sind (z.B. Gromacs[16]), als auch institutseigener Code, der teils versioniert, meist aber nicht veröffentlicht ist.

Die im Forschungsprozess in den Ingenieurwissenschaften anfallenden Daten sowie der Zusammenhang zwischen Eingabedaten und -code sowie Ausgabedaten sind in Abbildung 2 dargestellt: Die Ein- und Ausgabedaten im Forschungsprozess liegen in einem Format vor, das spezifisch für die jeweilige Software ist. Die Rohdaten lassen sich in der Regel auch nur mit dieser Software verarbeiten. Während die Größe der Eingabedaten meist im Bereich von Kilo- bis Megabyte, höchstens aber im Gigabytebereich liegt, können die Ausgabedateien auch mehrere hundert Gigabyte bis hin zu einigen Terabyte umfassen, was von der iterativen Natur der numerischen Lösungsansätze herrührt. Die Eingabekonfiguration legt das zu untersuchende System fest (Komponenten, zeitliche und räumliche

14 Ein Höchstleistungsrechner ist ein sehr leistungsfähiges Rechensystem, das massiv parallel arbeitet und aus vielen, über ein schnelles Kommunikationsnetzwerk verbundenen, einzelnen Rechnern besteht.

15 Schembera, Björn; Bönisch, Thomas Bönisch: Challenges of Research Data Management for High Performance Computing, International Conference on Theory and Practice of Digital Libraries, 2017, S. 140–151. Online: <https://doi.org/10.1007/978-3-319-67008-9_12>.

16 Weitere Informationen zu Gromacs: <http://www.gromacs.org/>, Stand: 06.02.2018.

Auflösung). Restart-Dateien enthalten alle Informationen, um eine Simulation oder Berechnung ab diesem Zeitpunkt weiterlaufen zu lassen. Skripte zur Steuerung der Erzeugung, Analyse, Auswertung und Visualisierung werden meist mit allgemeinen Tools oder Sprachen wie Bash, gnuplot, Python oder Matlab von den einzelnen Forscherinnen und Forschern erstellt.

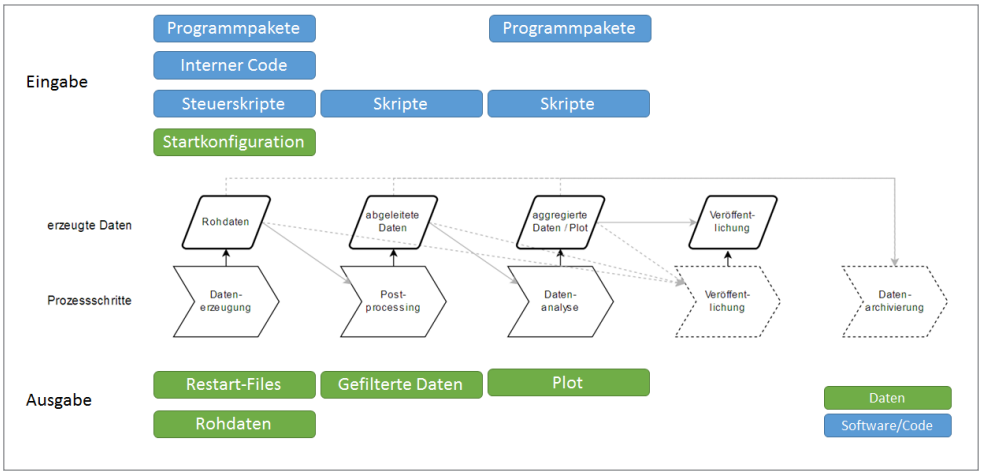

Abb. 2: Ein- und Ausgabedaten und verwendete Arten von Software und Codes je nach Schritt im Forschungsprozess.

3. Problembereiche

Im Rahmen der durchgeführten halbstrukturierten Interviews[17] wurden von den befragten Wissenschaftlerinnen und Wissenschaftlern viele verschiedene Probleme genannt, die aktuell beim Umgang mit Forschungsdaten auftreten und Leidensdruck erzeugen. Zusammengefasst lassen sich die nachfolgenden Problembereiche identifizieren:

Fehlende Nachnutzbarkeit von Daten

Während das Nachvollziehen der Ergebnisse anderer durchaus eine große Rolle im Forschungsprozess spielt, geschieht eine direkte Nachnutzung von Daten in den betrachteten Bereichen sehr selten und das nicht nur zwischen, sondern auch innerhalb von Forschungsgruppen. Dies liegt zu einem großen Teil daran, dass eine einheitliche Dokumentation der Daten fehlt und damit Daten und Code unverständlich für andere sind. Hemmend wirkt aber auch das schnelle Veralten von Daten und Codes durch Voranschreiten der Rechenleistung und eine große Abhängigkeit von der Systemumgebung. Da standardisierte Dateiformate fehlen, ist für das Lesen und Verarbeiten von Daten in der Regel die Software in der Version und Rechenumgebung notwendig, mit der sie ursprünglich erzeugt wurden, mindestens aber Informationen über das Datenformat.

17 Siehe Einleitung für eine Beschreibung des Vorgehens.

Schwieriges Handling von Daten

Zudem kann das Handling von im Bereich der Ingenieurwissenschaften erzeugten Daten sehr viel aufwändiger sein als in anderen Forschungsbereichen. Die erzeugten und zu analysierenden Daten können von einigen hundert Gigabyte bis zu mehreren Terabyte groß werden. Diese Datenmengen zu bewegen und zu verarbeiten ist sowohl zeitaufwändig wie auch ressourcenintensiv. Die Dateisysteme im Höchstleistungsrechner-Umfeld sind auf Übertragungsraten und hohe Kapazitäten spezialisiert, bieten aber wenig Unterstützung bei der strukturierten Ablage und Suche von Daten. Die Möglichkeiten zur Dateiorganisation beschränken sich auf entsprechende Datei- und Verzeichnisnamen.

Fehlende Motivation zur Veröffentlichung von Forschungsdaten

Nicht zuletzt aus diesen Gründen gehört die Veröffentlichung von Forschungsdaten in den betrachteten Forschungsbereichen nicht zur Fachkultur. Die Anforderungen der guten wissenschaftlichen Praxis der Deutschen Forschungsgemeinschaft (DFG)[18] werden zwar durch die Ablage der Daten zum Teil erfüllt, wegen der mangelnden Nachnutzbarkeit aber gleichzeitig in Frage gestellt. Die Motivation und Bereitschaft der einzelnen Wissenschaftlerinnen und Wissenschaftler, Forschungsdaten zur Verfügung zu stellen, wird durch verschiedene Faktoren reduziert: Eine Veröffentlichung der Daten bietet zu wenig Anreize, bedeutet aber viel Aufwand für Dokumentation und Aufbereitung der Daten. Bei Daten, die keinen erkennbaren langfristigen Nutzen aufweisen, wird der Sinn einer langfristigen Archivierung generell in Frage gestellt. Diese Einschätzung aber von der Art der Daten ab. Für experimentell erhobene Daten, die nicht einfach wiederhergestellt werden können, ist die Bereitschaft zur Archivierung deutlich größer als bei Daten, die auf numerischen Simulationen beruhen. Ein Aspekt der Motivation ist auch die Größe der Fachcommunity. In manchen Bereichen ist die Forschung so spezialisiert, dass aus Sicht der Forscher nur eine sehr kleine Menge von Personen Nutzen aus den Daten ziehen kann, eine Veröffentlichung dieser aber gleichzeitig einen hart erkämpften Forschungsvorsprung gefährden könnte.

Rechtliche Fragestellungen

Nicht nur, aber vor allem in Industrieprojekten sorgen rechtliche Unklarheiten für Hindernisse bei der Verwaltung und Veröffentlichung von Forschungsdaten: Bei Forschung mit externen Daten von Industriepartnern besteht oft Unklarheit, für was diese Daten verwendet, für welchen Zeitraum sie gespeichert und ob sie oder auf ihnen beruhende Daten veröffentlicht werden dürfen.

3.1. Bewertung der Probleme

Ein Großteil der Probleme von Forschungsdatenmanagement in den Ingenieurwissenschaften ist technisch bedingt: Die Größe der Daten, die unterschiedlichen Dateiformate sowie die Abhängigkeit von der Rechnerumgebung benötigt auf diese Disziplinen konkret zugeschnittene technische Lösungsansätze. Um eine sinnvolle Nachnutzung der Forschungsdaten zu ermöglichen, muss aber im Bereich der Motivation angesetzt werden. Der Aufwand, der für die dafür notwendige Dokumentation und

18 Deutsche Forschungsgemeinschaft: Leitlinien zum Umgang mit Forschungsdaten. Technischer Bericht, Bonn, 2015. Online: <http://www.dfg.de/download/pdf/foerderung/antragstellung/forschungsdaten/richtlinien_forschungsdaten.pdf>, Stand: 16.02.2018.

strukturierte Verwaltung der Daten entsteht, muss aufgewogen werden durch unmittelbare Vorteile im eigenen Forschungsprozess.

Unsicherheit in rechtlichen Fragestellungen ergeben sich in den Ingenieurwissenschaften weniger im Persönlichkeitsrecht und Datenschutz als im Bereich des Urheberrechts, insbesondere bei Beteiligung von Industriepartnern. Hier spielen auch weitere Fragestellungen eine Rolle, wie z.B. Produkthaftung, patentrechtliche Fragen oder Verschwiegenheitsverpflichtungen. Klare rechtliche Regelungen können hier Abhilfe schaffen.

4. Reanalyse bwFDMCommunities für Ingenieurwissenschaften

Die in der Bedarfsanalyse von DIPL-ING erhobenen Anwendungsfälle bieten zwar einen qualitativen Blick auf den Forschungsprozess in den betrachteten Forschungsbereichen, können dabei aber keinen Anspruch auf Allgemeingültigkeit für die Ingenieurwissenschaften erheben. Innerhalb vom Projekt bwFDMCommunities wurden 627 teilstrukturierte Interviews in den Jahren 2014 und 2015 an den baden-württembergischen Universitäten zum Thema Forschungsdatenmanagement durchgeführt.[19] Alle Wünsche, die innerhalb der Interviews geäußert wurden, liegen als User Stories in der Form „Als (Rolle) wünschen wir uns (Wunsch) um (Ziel) zu erreichen“ vor.[20] Aus den Ingenieurwissenschaften liegen 589 User Stories vor, die anhand zweier Fragestellungen reanalysiert wurden: Welche Abschnitte des Datenlebenszyklus werden in den Wünschen der Ingenieurwissenschaftler am meisten genannt? Welche Art der Unterstützung wird in den User Stories konkret gewünscht?

Dabei zeigt sich, dass die befragten Forscherinnen und Forscher Unterstützungsbedarf vor allem bei der Beschreibung von Daten (54 Nennungen), der Datenverwaltung (42 Nennungen), beim Nachnutzen von externen Daten (41) und bei der Archivierung von Daten oder Software (40) haben. Teilen, Veröffentlichen und Nachnutzen eigener Daten spielt dagegen eine wesentlich geringere Rolle. Als Art der Unterstützung werden am häufigsten technische Lösungen (Systeme, Implementierungen) (152 Nennungen) oder Personal (105 Nennungen) gewünscht, neben technischer Ausstattung (hier vor allem schnelle Netzwerke und Infrastrukturen für große Datenmengen), Informationen, neuen Regelungen oder Richtlinien.

Insgesamt decken sich die Daten aus bwFDMCommunities mit den Erkenntnissen aus der Bedarfsanalyse. Die Bedarfe der Forscherinnen und Forscher liegen vordringlich im lokalen Datenmanagement, die Veröffentlichung der Daten ist nachgelagert. Die Wissenschaftlerinnen und Wissenschaftler möchten sich gerne auf ihr Forschungsgebiet konzentrieren und wünschen sich daher für die Aufgaben rund um das Datenmanagement technische Lösungen oder zusätzliches Personal.

19 Tristram: Öffentlicher Abschlussbericht von bwFDM-Communities, 2015.
20 Tristram, Frank; Streit, Achim: Daten zu bwFDM-Communities, 2015, <http://bwfdm.scc.kit.edu/cgi-bin/daten/>, Stand: 16.02.2018.

5. Anforderungen

Im Folgenden werden die Anforderungen gruppiert und zusammengefasst, welche sich aus der Bedarfsanalyse direkt ergeben.

5.1. Dokumentation und Metadaten

Da die in den Ingenieurwissenschaften verwendeten Daten (siehe Abschnitt 2.3) in der Regel nicht selbsterklärend sind, ergeben die Aufbewahrung, das Veröffentlichen oder Teilen von Code und Daten nur dann Sinn, wenn diese ausreichend dokumentiert sind. Dafür werden Metadaten benötigt, die einerseits wichtige Suchkriterien abdecken und andererseits die Daten soweit beschreiben, dass ein Verstehen und Nachvollziehen möglich wird.

5.1.1. Funktionale Anforderungen an die Metadaten

Laut den FAIR-Prinzipien (Wilkinson, et al. 2016) sollen Forschungsdaten auffindbar (F = findable), zugreifbar (A = accessible), zwischen verschiedenen Systemen austauschbar (I = interoperable) und nachnutzbar (R = reusable) sein.[21] Grundlegend dafür sind persistente Identifier, über die die Daten auffindbar und erreichbar sind, standardisierte Protokolle und Vokabulare und Metadaten, die sowohl maschinen- als auch menschenlesbar sind.

Für die Ingenieurwissenschaften stellen sich auf inhaltlicher Ebene darüber hinaus folgende Anforderungen:

Auffindbarkeit: Die Beschreibung sollte alle Informationen enthalten, nach der Forscherinnen und Forscher aus dem betreffenden Fachgebiet suchen möchten. Potentielle Suchkriterien sind neben dem Autor, dem Jahr und Titel/Beschreibung der Daten auch die erhobenen und kontrollierten Variablen, angewendete Methoden und Spezifika des untersuchten Systems. Die Daten können auch über eine zugehörige Publikation oder über einen persistenten Identifier (z.B. eine DOI) gesucht werden.

Teilbarkeit: Die Beschreibung sollte alle relevanten Informationen enthalten, die zum Verständnis des Datensatzes notwendig sind. Um die Daten verstehen und nachvollziehen zu können, müssen darüber hinaus auch die Software und die Rechenumgebung bekannt sein, mit der die Daten erzeugt, verarbeitet und/oder analysiert wurde. Eine etwaige Publikation kann auch als zusätzliche Dokumentation der Daten dienen. Für den internen Gebrauch ist hilfreich zu wissen, ob der Ansatz erfolgreich war und warum dies der Fall war, um auch Negativergebnisse dokumentieren zu können.

Zitierbarkeit: Die Beschreibung eines Datensatzes muss alle Pflichtinformationen eines standardisierten Metadatenformats erfüllen. Dabei ist Kompatibilität mit dem DataCite-Schema[22] von Bedeutung, da dies grundlegend für die DOI-Vergabe ist.

21 Wilkinson, Mark D.; Dumontier, Michel; Aalbersberg, IJsbrand Jan u.a.: The FAIR Guiding Principles for scientific data management and stewardship, in: Scientific Data 3, 2016. Online: <http://dx.doi.org/10.1038/sdata.2016.18>.

22 Data Cite Metadata Working Group: DataCite Metadata Schema for the Publication and Citation of Research Data. Version 4.1, 2017, <http://dx.doi.org/10.5438/0015>.

Archivierbarkeit: Die Beschreibung sollte technische Daten beinhalten, die für eine Langzeitarchivierung von Bedeutung sind, wie Datei-Formate und Prüfsummen zur Beurteilung der Integrität der Daten.

Anpassbarkeit: Terme der Beschreibung sollten fachspezifisch übersetzbar und um fachspezifische Metadaten erweiterbar sein.-

Interoperabilität: Die Beschreibung sollte in die Metadatenstandards von beteiligten Organisationen (PID-Vergabe, Repositorien,[23] Metasuchmaschinen) überführbar sein und standardisierte Vokabulare verwenden.

Um diese Anforderungen zu erfüllen, muss also nicht nur der Datensatz selbst inhaltlich allgemein beschrieben werden (deskriptive Metadaten), sondern auch das betrachtete System aus Sicht des Fachbereichs (fachspezifische Metadaten) und der Weg der Entstehung zusammen mit der genutzten Rechenumgebung (Software, Hardware) und verwendeten Instrumenten dokumentiert werden (Prozessmetadaten). Beteiligte Personen und Institutionen sollten genauso festgehalten werden wie mit dem Datensatz verbundene Publikationen und Daten, ebenso die damit verbundenen Rechte und Lizenzen, Prüfsummen und Dateiattribute (technische Metadaten).

5.2. Entscheidungskriterien zur Auswahl von Daten und Dauer der Archivierung

Bedingt durch die Größe und Anzahl der erzeugten Daten ist es ökonomisch nicht sinnvoll oder überhaupt technisch möglich, alle Forschungsdaten und -codes zu speichern, die im Forschungsprozess anfallen. Die Frage stellt sich vor allem bei Daten, die auf Simulationen beruhen, da sie – im Gegensatz zu experimentellen Daten – in der Regel wieder hergestellt werden können.

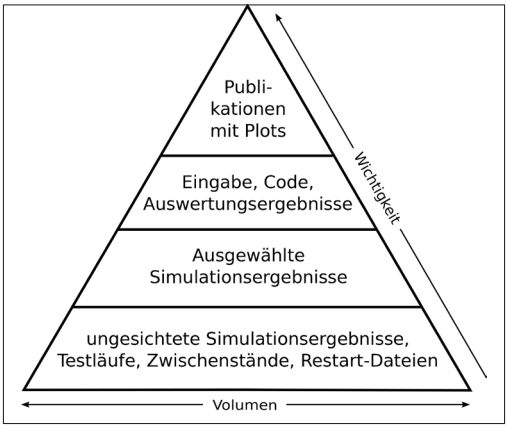

Abb. 3: Datenpyramide für die Simulationswissenschaften, adaptiert nach (Reilly, et al. 2011)[24]

23 Vgl. zum Beispiel fachspezifische Repositorien oder allgemeine Repositorien wie RADAR, figshare oder Zenodo.

24 Reilly, Susan; Schallier, Wouter; Schrimpf, Sabine u.a.: Report on integration of data and publications, 17.10.2011, <http://epic.awi.de/31397/1/ODE-ReportOnIntegrationOfDataAndPublications-1_1.pdf>, Stand: 16.02.2018.

Grundlegend für die Entscheidung, welche Simulationsdaten für welchen Zeitraum gespeichert werden, kann die in Abbildung 3 gezeigte Datenpyramide sein, die die Relation zwischen Volumen der Daten und deren Wichtigkeit darstellt. Die ungesichteten Simulationsergebnisse, Testläufe und Zwischenstände bilden die Basis für alle weitere Forschungsarbeit (Stufe 1). Hieraus ausgewählte Simulationsergebnisse (Stufe 2) bilden die Grundlage für die dann folgenden Analyseschritte. Die Ergebnisse der Analyse (Stufe 3) liefern die Forschungserkenntnisse, die zusammengefasst als Plot in die Publikation eingehen (Stufe 4). Die zugrundeliegenden Inputfiles, Codes und Skripten bieten bei relativ kleiner Größe viele Informationen, mit denen die Daten verstanden oder wieder erzeugt werden können (Stufe 3). Auswahlkriterien können sich nun aus der Relation Wichtigkeit/Volumen ergeben: Publikationen und deren direkt zugrundeliegenden Analyseergebnisse, Skripten und Inputfiles müssen längerfristig gespeichert werden, wohingegen die Testläufe und Zwischenstände retrospektiv keinen Nutzen bringen und gelöscht werden können. Inwieweit Simulationsergebnisse als Rohdaten der Stufe 2 archiviert werden sollen, hängt von dem vorhandenen Speicherplatz und den Möglichkeiten der Wiederherstellung ab.

Für die gespeicherten Daten ist es dann notwendig, sinnvolle Aufbewahrungszeiträume zu definieren, nach deren Ablauf Daten auf ihre weitere Verwendbarkeit geprüft und gegebenenfalls auch gelöscht werden können.

Nicht immer ist es während oder nach Abschluss eines Projektes möglich, die potentielle Nutzbarkeit von Daten oder Code einzuschätzen. Es werden also Daten und Kennzahlen benötigt, die während der Archivierung oder nach der Veröffentlichung Aussagen über die Nutzung von Daten und Code liefern. Auf Grund dieser Kennzahlen könnten Fristen neu angepasst werden.

5.3. Technische Anforderungen

5.3.1. Automatisierung der Datendokumentation im Workflow

Sollen die im Absatz 5.1.1 definierten Anforderungen erfüllt werden, ist eine Vielzahl von Metadaten für jedes Forschungsdatum zu definieren. Dies wird im Forschungsalltag realistisch nur dann zu bewerkstelligen sein, wenn die Datendokumentation während des Forschungsworkflows so automatisiert wie möglich geschieht.

Viele der zu dokumentierenden Informationen ändern sich selten (wie die Kontaktinformationen von Personen) oder wiederholen sich häufig (wie Projekteigenschaften, z.B. Projektmitarbeiter, Keywords oder Finanzierungsinformationen). Solche Informationen können in Profilen gespeichert und eingebunden werden.

Einige Metadaten können automatisiert erfasst oder erzeugt werden. Dies sind insbesondere die POSIX-Dateieigenschaften, wie Dateigröße, -name und zu erzeugende Merkmale wie Prüfsummen, aber auch die fachspezifischen Informationen, die in den Log-Dateien der Simulationscodes zu finden sind.

Zwar benötigen solche Automatisierungen einmalig Anpassungs- oder Entwicklungsaufwand, wirken aber insofern unterstützend, indem sie auf lange Sicht die Belastung der Forscherinnen und Forscher reduzieren und damit machen eine umfangreiche Dokumentation der Daten erst möglich.

5.3.2. Einbindung des lokalen Datenmanagements

Das lokale Datenmanagement ist die Grundlage für alle weiteren Schritte im Datenlebenszyklus und bietet die besten Anknüpfungspunkte, um schnell Verbesserungen im Arbeitsalltag für die Forscherinnen und Forscher spürbar zu machen. Jegliches technische System, welches Forscherinnen und Forscher unterstützen soll, muss also am lokalen Datenmanagement ansetzen und zunächst lokale Lösungen für die Metadatenannotation und die Verwaltung und Suchbarkeit der Daten finden.

5.3.3. Anforderungen an Datenrepositorien zur Veröffentlichung

Ein ideales Datenrepositorium für die Ingenieurwissenschaften wie in Abbildung 4 erfüllt die folgenden Anforderungen:

Umgang mit großen Datenmengen: Die umfangreichen Datenmengen, die bei Simulationen oder auch Experimenten entstehen, verlangen neue Lösungen für Speicherung, Import und Ausgabe der veröffentlichten Daten. Große Datenmengen sollten an beliebigen physikalischen Orten liegen können (Ortstransparenz der Daten), um günstigere Speichermedien zu ermöglichen und aufwändigen Transport der Daten zu vermeiden. Die Bandbreiten müssen groß genug und Übertragungstechnologien[25] schnell sein, damit auch Hunderte von Terabytes ihren physikalischen Speicherplatz in praktikabler Zeit wechseln können.

Umgang mit Code und Daten: In den simulierenden Ingenieurwissenschaften hängen die Daten von Software und Code ab. Dadurch ist es von hoher Priorität, auch die Software genau zu referenzieren oder (gemeinsam mit den Daten) zu archivieren. Die Herausforderung besteht dabei darin, Nachnutzern die Ausführung von Software und Code und damit das Nachvollziehen der Ergebnisse ohne großes technisches Wissen zu ermöglichen.

Fachspezifische Sicht auf Metadaten: Gerade die fachspezifischen Metadaten sind für die Forschenden entscheidende Suchkriterien. So wie für Forschende aus den Geowissenschaften die räumliche Lage entscheidend ist, müssen Ingenieurwissenschaftlerinnen und Ingenieurwissenschaftler nach physikalischen Parametern, Komponenten oder betrachteten Ensembles suchen können. Ein fachspezifisches Repositorium muss also die in Kapitel 5.1.1 genannten Metadaten nicht nur anzeigenkönnen, sondern auch eine (facettierte) Suche auf ihnen ermöglichen.

Qualitätsmanagement: Durch die großen Datenmengen und die damit verbundenen Speicherkosten ist es wichtig, die Qualität zu speichernder Datensätze einschätzen zu können, um den Aufwand auf die Datensätze zu konzentrieren, die am meisten Potential für die Nachnutzung bieten. Ein Repositorium

25 Eine solche Übertragungstechnologie ist z.B. GridFTP. Vgl. Allcock, William; Bresnahan, John; Kettimuthu, Rajkumar u.a.: The Globus Striped GridFTP Framework and Server, Proceedings of the 2005 ACM/IEEE Conference on Supercomputing, Seattle, 2005, S. 54. Online: <http://dx.doi.org/10.1109/SC.2005.72>.

sollte daher sowohl ein Peer-Review-Verfahren für Datenunterstützen, als auch Downloadstatistiken und die Möglichkeit einer Rückmeldung durch Nutzerinnen und Nutzer nach der Veröffentlichung anbieten. Diese Metriken können eine fundierte Entscheidung über die Archivierung, bzw. die Archivierungsdauer unterstützen. Hierbei können die in Kapitel 5.2 erarbeiteten Entscheidungskriterien eine Anleitung geben, um die Wichtigkeit der Daten mit ihrer Größe abzuwägen.

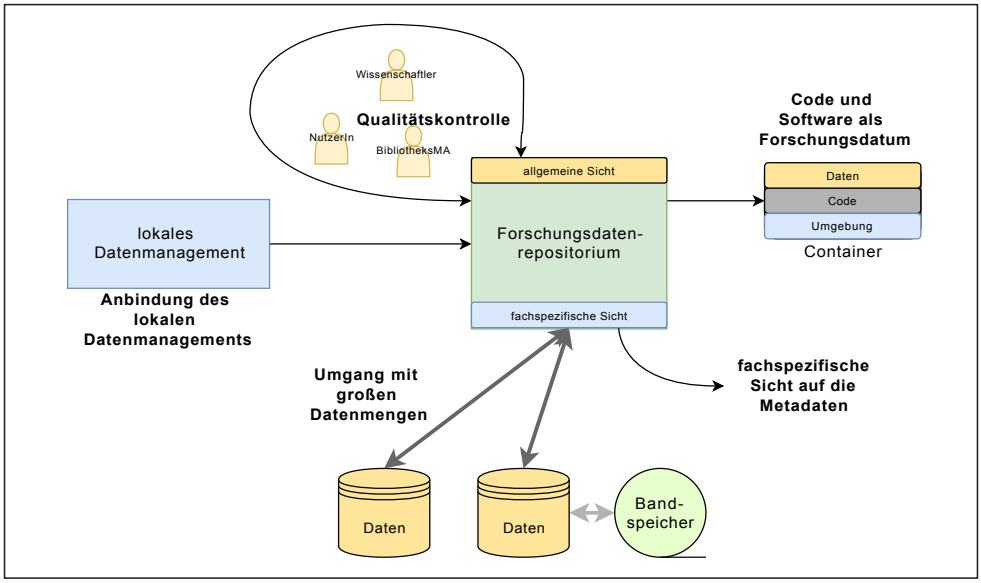

Abb. 4: Ein ideales Datenrepositorium

6. Zusammenfassung und Ausblick

Um Forschungsdatenmanagement in den Ingenieurwissenschaften zu etablieren, gilt es zunächst die spezifischen Probleme des Fachbereichs zu verstehen (siehe Kapitel 3): große Datenmengen, die schnell veralten, Code- und Systemabhängigkeit und fehlende Verankerung des Forschungsdatenmanagements in den Fachdisziplinen. Die aktuell übliche individuelle Dokumentation der Daten durch Ordner und Dateibezeichnungen sorgt dafür, dass Daten unverständlich und kaum nachnutzbar in Datengräbern landen.

Um den herausgearbeiteten Problemen beizukommen und den ingenieurwissenschaftlichen Forschungsprozess durch institutionalisiertes Forschungsdatenmanagement zu unterstützen, sind einige Anforderungen zu erfüllen: Notwendig ist zunächst ein strukturiertes Metadatenschema zur Beschreibung der Daten, um die Auffindbarkeit, die Verständlichkeit und die Nachnutzbarkeit zu ermöglichen. Bibliotheken sind hier in der Pflicht, über vorhandene fachspezifische Metadatenschemata zu informieren oder Forscher dabei zu unterstützen, ein geeignetes lokales Metadatenschema

auf der Basis vorhandener Schemata zu erstellen. Diese Beschreibung der Daten muss mit möglichst wenig manuellem Aufwand bereits bei der Erzeugung ansetzen und so viele Metadaten wie möglich automatisiert erfassen. Hier sind Rechenzentren gefordert, Schnittstellen im Forschungsworkflow zu schaffen. Damit wird ein lokales Datenmanagement möglich, mit dem Daten langfristig einfach gefunden und verstanden werden können und auch Fehlschläge und Negativergebnisse dokumentiert werden können. Die Veröffentlichung der so dokumentierten und strukturierten Daten ist dann kein großer Schritt mehr. Klare Entscheidungskriterien und Qualitätsmaße für die Auswahl und Dauer der Speicherung konzentrieren den Aufwand auf die Datensätze, die das meiste Potential für eine interne wie externe Nachnutzung bieten. Ein weitgehend automatisierter Prozess macht das Datenmanagement auch für schnell veraltende Daten und Codes möglich. Wenn ein institutionelles oder fachspezifisches Repositorium es im letzten Schritt etwaigen Nachnutzern einfach macht, die veröffentlichten Daten und Codes zu verstehen und zu verwenden, ist ein großer Schritt in Richtung Open Science getan.

Literaturverzeichnis

- Allcock, William; Bresnahan, John; Kettimuthu, Rajkumar u.a.: The Globus Striped GridFTP Framework and Server, Proceedings of the 2005 ACM/IEEE Conference on Supercomputing, Seattle, 2005, S. 54. Online: <http://dx.doi.org/10.1109/SC.2005.72>.

- Data Cite Metadata Working Group: DataCite Metadata Schema for the Publication and Citation of Research Data. Version 4.1, 2017, <http://dx.doi.org/10.5438/0015>.

- Deutsche Forschungsgemeinschaft: Leitlinien zum Umgang mit Forschungsdaten. Technischer Bericht, Bonn, 2015. Online: <http://www.dfg.de/download/pdf/foerderung/antragstellung/forschungsdaten/richtlinien_forschungsdaten.pdf>, Stand: 16.02.2018.

- Joo, Yeon Kyoung; Kim, Youngseek: Engineering researchers' data reuse behaviours. A structural equation modelling approach, in: The Electronic Library 35 (6), 2017, S. 1141–1161. Online: <http://dx.doi.org/10.1108/EL-08-2016-0163>.

- Krüger, Jens; Grunzke, Richard; Gesing, Sandra u.a.: The MoSGrid Science Gateway. A Complete Solution for Molecular Simulations, in: Journal of Chemical Theory and Computation 10 (6), 2014, S. 2232–2245. Online: <http://dx.doi.org/10.1021/ct500159h>.

- Kumar, Anand; Grupcev, Vladimir; Berrada, Meryem u.a: DCMS. A data analytics and management system for molecular simulation, in: Journal of Big Data 2 (1), 2014, S. 9. Online: <https://doi.org/10.1186/s40537-014-0009-5>.

- Lautenschlager, Michael; Toussaint, Frank; Thiemann, Hannes u.a: The CERA-2 data model, 1998, <https://www.pik-potsdam.de/cera/Descriptions/Publications/Papers/9807_DKRZ_TechRep15/cera2.pdf>, Stand: 07.02.2018.

- Malik, Tanu: Geobase: Indexing NetCDF Files for large-scale Data Analysis, in: Big Data Management, Technologies, and Applications, 2014 (IGI Global), S. 295–313. Online: <http://dx.doi.org/10.4018/978-1-4666-4699-5.ch012>.

- Meneveau, Charles, Marusic, Ivan: Turbulence in the Era of Big Data. Recent Experiences with Sharing Large Datasets, in: Pollard, Andrew; Castillo, Luciano; Danaila, Luminita u.a. (Hg.): Whither turbulence and big data in the 21st century? Cham 2017, S. 497–507. Online: <http://dx.doi.org/10.1007/978-3-319-41217-7>.

- Reilly, Susan; Schallier, Wouter; Schrimpf, Sabine u.a.: Report on integration of data and publications, 17.10.2011, <http://epic.awi.de/31397/1/ODE-ReportOnIntegrationOfDataAndPublications-1_1.pdf>, Stand: 16.02.2018.

- Schembera, Björn; Bönisch, Thomas Bönisch: Challenges of Research Data Management for High Performance Computing, International Conference on Theory and Practice of Digital Libraries, 2017, S. 140–151. Online: <https://doi.org/10.1007/978-3-319-67008-9_12>.

- Sillero, Juan A.; Jiminéz, Javier: Public Dissemination of Raw Turbulence Data, in: in: Pollard, Andrew; Castillo, Luciano; Danaila, Luminita u.a. (Hg.): Whither turbulence and big data in the 21st century? Cham 2017, S. 509–515. Online: <http://dx.doi.org/10.1007/978-3-319-41217-7_28>.

- Tristram, Frank; Streit, Achim: Daten zu bwFDM-Communities, 2015, <http://bwfdm.scc.kit.edu/cgi-bin/daten/>, Stand: 16.02.2018.

- Tristram, Frank; Streit, Achim: Öffentlicher Abschlussbericht von bwFDM-Communities. Technischer Bericht, Karlsruher Institut für Technologie, 2015, <https://bwfdm.scc.kit.edu/downloads/Abschlussbericht.pdf>, Stand: 16.02.2018.

- Wilkinson, Mark D.; Dumontier, Michel; Aalbersberg, IJsbrand Jan u.a.: The FAIR Guiding Principles for scientific data management and stewardship, in: Scientific Data 3, 2016. Online: <http://dx.doi.org/10.1038/sdata.2016.18>.

