

Leitfaden für die Publikation und Archivierung von Forschungsdaten an der TU Braunschweig

Die TU Braunschweig bietet ihren Forschern die Möglichkeit, ihre Forschungsdaten auf dem [Publikationsserver](#) zu archivieren und, falls gewünscht, zu publizieren. Im Folgenden wird auf die wichtigsten Punkte eingegangen.

1) Forschungsdaten vorbereiten

Kurz und knapp

- Mehrere Dateien sollten in hierarchisch strukturierten Ordnern mit sinnvoller Namensgebung gespeichert werden.
- Ordnerbenennungen sollten auch für Dritte nachvollziehbar, d.h. systematisch und inhaltsbezogen sein.
- Die abgelegten Dateien sollen selbständige Nachvollziehbarkeit und Nachnutzbarkeit ermöglichen.
- Im Vorfeld der Archivierung/ Publikation ist es ratsam eine gewissenhafte Datenselektion durchzuführen und nur Daten auszuwählen, die wissenschaftlich relevant und archivierungswürdig sind.
- Die Dateien sollten in standardisierten, verbreiteten Dateiformaten für kostenfreie Darstellungsprogramme vorliegen (z.B. XML, CSV, TXT, TIFF).
- Liegen Daten in einem proprietären Format vor (z.B. MatLab), sollten diese nach Möglichkeit auch in einem offenen Format abgelegt werden.
- Die Daten dürfen nicht schreibgeschützt sein.

Ordnerstrukturierung

Um eine möglichst nachvollziehbare, klare Strukturierung der Forschungsdaten zu erlangen, eignet sich eine hierarchische Struktur. Wie diese genau aussieht, hängt vom Datensatz ab und muss jedes Mal individuell festgelegt werden. Bei heterogener Struktur der Daten (viele verschiedene Dateierarten) kann man Ordnerarten anlegen, z.B.:

- Materialien → Daten, Dokumentationen, Publikationen
- Daten → Texte, Fotos, Modelle, Tabellen

Sind die Daten homogener, können sie auch nach Zeit, Ort oder Methode unterteilt werden, z.B.:

- Projekt_XY → Untersuchungsgebiet01 → Messungen

Wichtig ist dabei, dass die gewünschten Dateien schnell und intuitiv gefunden werden können.

Dateibenennung

Bei der Benennung von Dateien sollte von Anfang an eine Namenskonvention festgelegt werden, an der man durchs gesamte Projekt konsistent festhält, gerade wenn mehrere Parteien an einem Projekt beteiligt sind. Die Benennung sollte Aufschluss über den Inhalt der Daten geben und helfen, sie eindeutig zu identifizieren. Dafür bieten sich z.B. besonders das Erstellungsdatum, eine kurze Beschreibung, der Ort, Projektname, Probennummer, Analyseverfahren oder Versionsnummer an.

Zudem sollte man folgende Regeln beachten:

1. Dateinamen sollten kurz aber erklärend sein.
2. Es sollten keine Sonderzeichen, Leerzeichen, Satzzeichen, oder Umlaute verwendet werden.
3. Es sollten Großbuchstaben und Unterstriche stattdessen verwendet werden.
4. Großbuchstaben beeinflussen die Sortierung, weswegen auf eine konsistente Benennung geachtet werden sollte.
5. Als Datumsformat eignet sich am besten: YYYYMMDD.
6. Wenn Nummern angegeben werden, sollten diese immer zweistellig oder gar dreistellig sein, sodass die 0 als Platzhalter dient (z.B. Probe01 und nicht Probe1).
7. Die einzelnen Elemente des Dateinamens sollten so angeordnet sein, dass die Datei schnellstmöglich gefunden wird, Namensanfänge wie „Test“ oder „Entwurf“ sollten vermieden werden.
8. Wenn verschiedene Versionen einer Datei vorhanden sind, sollte ein V mit Nummerierung verwendet werden, um Ordnung zwischen den Versionen zu halten (V01, V02, etc.).
9. Wiederholungen von Informationen aus Ordnernamen in Dateinamen sollten vermieden werden.

Sinnvolle Elemente für die Dateibenennung

Sinnvolle Elemente	Hinweise und Beispiel
Name des Verfassers oder Initialen	LG
Inhaltliche/ Fachliche Beschreibung	Schlagwörter
Datum der Erstellung v.a. für Versionskontrolle	YYYYMMDD
Bearbeitungsstatus	Original, Entwurf, Teil
Version	V02

Beispiele für Dateibenennungen

Gut	Schlecht
20200813Agenda	13Aug2020Agenda
2020UBWorkshopFDM	Der_2020_Workshop_zum_Thema_Forschungsdatenmanagement_der_UB!
20200804_ProjektA_Protokoll_LG	Protokoll Laura

Vertiefende Informationen: http://www.data.cam.ac.uk/files/gdl_tilsdocnaming_v1_20090612.pdf

Die Namenskonvention der Dateien und Ordnerstruktur sollten in der beiliegenden README-Datei des Datensatzes erläutert werden, falls Kürzel verwendet werden.

Datenselektion

Um zu entscheiden, welche Daten geeignet für eine Archivierung/ Publikation sind, können sie in folgende Kategorien eingeteilt werden.

Kategorie 1: Sehr wertvolle Forschungsdaten

Forschungsdaten, die

- einzigartig oder sehr aufwändig wiederherzustellen sind und
- für eine große Community und/oder für ein breites Spektrum an Nachnutzungsmöglichkeiten interessant sind.

→ Unbedingt archivieren / publizieren!

Kategorie 2: Begrenzt wertvolle Forschungsdaten

Forschungsdaten, die entweder

- mit vertretbarem Aufwand wiederherzustellen sind oder
- nur für kurze Zeit, eine kleine Zielgruppe oder nur zum Nachweis von Forschungsergebnissen interessant sind.

→ Archivieren/ publizieren, wenn Aufwand und Kosten nicht dagegen sprechen.

Kategorie 3: Wenig wertvolle Forschungsdaten

Forschungsdaten, die

- nur zu Testzwecken entstanden,
- nicht lesbar oder nicht ausreichend beschrieben sind.

→ Daten sind nicht geeignet für Archivierung/ Publikation.

Empfohlene Dateiformate

Nicht jedes Dateiformat ist für eine Langzeitarchivierung geeignet. Damit Forschungsdaten auch nach einigen Jahren lesbar und zu verwenden sind, empfehlen sich offene, nicht proprietäre Dateiformate.

	Empfohlene Formate	Ungeeignet für Archivierung	Konvertierungsmöglichkeiten
Textformate	PDF/A, TXT (UTF8 kodiert), XML	DOC, PPT	- Word & PowerPoint zu PDF/A-1b - LaTeX & TeX zu PDF/A
Tabellen/ Spreadsheets	CSV/ TVS	XLS	Tabellen zu ASCII Text-Dateien (CVS)
Statistische Umgebungen	SDD, R-Dateien, Matlab-Dateien in HDF-Format		- Matlab-Dateien als v7.3 MAT abspeichern (HDF5 basierter Standard) - Wichtige Tabellen im Workspace separat als CSV speichern
Rastergrafiken	TIFF, PNG, JPEG2000	EPS	
Vektorgrafiken	SVG		
Multimedia	WAV, AVI, MOV, MPEG-4	WMV	

Geodaten	GML, KML, ESRI Shapefile, Georeferenced TIFF		
Numerische Daten	NetCDF		

Vertiefende Informationen:

<https://www.ianus-fdz.de/it-empfehlungen/dateiformate>

<https://uni-tuebingen.de/forschung/forschungsinfrastruktur/escience-center/forschungsdaten-archivfaehigkeit/#c1168492>

2) Dokumentation vorbereiten

Kurz und knapp

- Die Dokumentation soll dazu dienen, die publizierten Daten umfassend zu beschreiben, sodass eine Reproduzierbarkeit und Nachnutzung der Daten ermöglicht wird.
- Der Datensatz sollte von einer README-Datei (TXT oder PDF/A) mit beschreibenden Metadaten begleitet werden, damit interessierte Wissenschaftler den Datensatz besser verstehen und nachnutzen können.
- Das README muss zumindest folgende Angaben beinhalten:
 - Titel der Datenpublikation
 - AutorInnen, UrheberInnen, RechteinhaberInnen
 - Institution und Projekt
 - Entstehungsjahr oder -zeitraum
 - Abstract/ Beschreibung des Datensatzes
 - Methode/ Datenerhebung
 - Verweis auf andere Publikationen, die den Datensatz beschreiben/ auswerten
 - Lizenz
- Auch für die Einreichung in ein Repositorium werden Metadateninformationen verlangt. Je präziser diese sind, desto besser kann die Publikation später gefunden und verstanden werden.
- Zusatzmaterial wie Laborbucheinträge, Software-Code und andere Begleitmaterialien sollten auch bereitgestellt werden, falls man sie zum Verständnis benötigt.

Strukturierung der README-Dateien

Bei sehr heterogenen Datensätzen (z.B. bestehend aus Geodaten, Matlab-Skripten und Tabellen mit Messdaten) können (1) alle Informationen in einer README-Datei beschrieben werden, wobei Bezug auf einzelne Datenspezifika genommen wird, oder es kann (2) ein "Haupt"-README erstellt werden, das den Zusammenhang des Gesamtdatensatzes beschreibt und ergänzend spezifische READMEs für Unterdatensätze oder auch einzelne Dateien, welche im jeweiligen Unterverzeichnis abgelegt werden. Bei mehreren README-Dateien sollte darauf geachtet werden, dass die Benennung der READMEs Aufschluss darauf gibt, welchen Datensatz/ Datei sie beschreiben. Es sollte auf eine einheitliche Formatierung der README-Dateien geachtet werden.

README-Vorlage

Wir haben eine eigene Dokumentvorlage für README-Dateien erstellt. Falls Sie Interesse haben, diese zu nutzen, sprechen Sie uns an!

3) Exkurs Forschungssoftware

Kurz und knapp

- Software sollte versioniert werden (z.B. mit der Hilfe von GitVersion).
- Software (hier Programm im ausführbaren Zustand) sollte immer mit zugehörigem Quell- oder Sourcecode versehen werden.
- Der Zweck der Software sollte kurz aber aussagekräftig formuliert werden, z.B. in in einer beiliegenden README-Datei.
- Der Code sollte modularisiert werden, um eine bessere Übersicht zu erhalten.
- Dabei sollte bei jeder Funktion erkennbar sein, was ihre Intention ist (z.B. durch selbsterklärende Namen oder Kommentare).
- Der Software sollte ein Testdatenset beigelegt werden, mit dem schnell untersucht werden kann, ob das Programm vernünftigt läuft. Noch besser wäre die Software direkt testgetrieben zu entwickeln.

Veröffentlichung von Forschungssoftware

Software nimmt einen immer größeren Stellenwert in der Forschung ein. Sie hilft nicht nur bei der Analyse von Daten, sondern ist oftmals auch für deren Herstellung entscheidend. Oftmals braucht man die zugrundeliegende Software, um Forschungsdaten richtig verstehen und vor allem auch reproduzieren zu können. Zudem kann selbst entwickelte Forschungssoftware ein eigenständiges Forschungsergebnis darstellen. Deshalb sollte Forschungssoftware wie Forschungsdaten zur Nachnutzbarkeit publiziert werden.

Metadaten für Forschungssoftware

Forschungssoftware benötigt eigene Metadaten, um sie gut zu beschreiben und nachvollziehbar zu machen. Ein Beispiel für einen Metadatenstandard ist [CodeMeta](#).

Hilfreiche Tools, um Forschungssoftware ausführbar zur Verfügung zu stellen

ExecShare: <http://www.execandshare.org/CompanionSite/>

jupyter-Notebooks: <http://jupyter.org/>

Docker-Container: <https://www.docker.com/>

Archivierung von Forschungssoftware

Die Langzeitarchivierung von Software gestaltet sich schwierig. Zum einen muss sie verständlich beschrieben sein und zum anderen muss dafür gesorgt werden, dass sie ausführbar gehalten wird. Darin liegt ein Problem, denn oft ist Software stark abhängig von der Rechenumgebung in der oder für die sie entwickelt wurde, sprich von Betriebssystem, Hardware oder Software-Bibliotheken. Diese ändern sich mit fortschreitender Technik sehr schnell, sodass entwickelte Software nicht mehr gelesen oder ausgeführt werden kann. Um dieses Problem zu lösen, gibt es noch keine endgültige Lösung. Ein

Ansatz ist die Nutzung von Emulation, bei der zusätzlich zur eigentlichen Software auch noch der ausführende Rechner samt Betriebssystem nachgebildet wird.

Weiterführende Literatur

Matthias Katerbow, Georg Feulner et al. (2018): Handreichung zum Umgang mit Forschungssoftware, <http://doi.org/10.5281/zenodo.1172970>

4) Rechtliche Aspekte abklären und Lizenz wählen

Kurz und knapp

- Alle rechtlichen (Urheber-/ Verwertungs-/ Nutzungsrechte, Datenschutz, Dienstvereinbarungen) und ethischen Bestimmungen müssen eingehalten werden.
- Auch die Wahl einer geeigneten Lizenz (z.B. Creative Commons) ist für die gewünschte Art der Nachnutzung essentiell.
- Es gilt: So offen wie möglich, so geschlossen wie nötig.
- Beachten Sie die [Vertragsbedingungen](#) des Publikationsservers der TU Braunschweig.

Urheberrecht

Mit dem Urheberrecht werden geistige Schöpfungen und Leistungen geschützt. Liegt ein Urheberrecht vor, ist die Nutzung, Weitergabe und Verarbeitung des Werks nur mit Einwilligung des Urhebers möglich. Forschungsdaten unterliegen nur selten dem Urheberrecht, es kann jedoch Ausnahmen geben. Oftmals werden publizierte Forschungsdaten allerdings mit Lizenzen versehen, die eine mögliche Nachnutzung klären. Datenbanken bilden im Übrigen eine Ausnahme und sie genießen laut deutschem Recht einen spezifischen Schutz, wodurch die Ersteller das alleinige Recht zur Verbreitung und Vervielfältigung haben (§ 87b UrhG). Allgemein endet der urheberrechtliche Schutz 70 Jahre nach dem Tod des Urhebers. Danach gelten die Werke als gemeinfrei.

Beschreibende Metadaten von Forschungsdaten sind i.d.R. als nicht urheberrechtlich geschützt zu betrachten, da es sich meist um relativ kurze, rein beschreibende Darstellungen handelt.

Neben einem Urheberrecht können Forschungsdaten in bestimmten Fällen auch einem Patentrecht unterliegen, wenn die Forschungsdaten technische Lehre beschreiben. In solchen Fällen darf nur der Patentinhaber die patentierten Daten nutzen.

Arbeits- & dienstvertragliche Vorgaben

Open-Access-Leitlinie der TU Braunschweig: https://www.tu-braunschweig.de/fileadmin/Redaktionsgruppen/Einrichtungen/UB/PDF/OA/Leitlinie_zu_Open_Access_an_der_TU_Braunschweig.pdf

Leitlinie zum Umgang mit Forschungsdaten der TU Braunschweig: <https://www.tu-braunschweig.de/fileadmin/Redaktionsgruppen/Einrichtungen/UB/PDF/forschungsdatenleitlinie.pdf>

Es ist im Allgemeinen darauf zu achten, was im Arbeitsvertrag bezüglich der Nutzungsrechte (gehören i.d.R. dem Arbeitgeber) an erzeugten Forschungsdaten festgehalten ist und ob z.B. ein Betriebsgeheimnis vorliegt.

Auch Förderer von Forschungsprojekten können Ansprüche haben, z.B. wenn es um die Rechte an erstellter Software im Zuge des Forschungsprojektes geht. Im Zweifelsfall sollte im Vorfeld einer Veröffentlichung nachgefragt werden, um ggf. eine Zustimmung für die Verbreitung einzuholen.

Datenschutz

Beim Umgang mit personenbezogenen Daten müssen verschiedene datenschutzrechtliche Aspekte beachtet werden. Durch eine Anonymisierung können datenschutzrechtliche Restriktionen vermieden werden. „Anonymisieren ist das Verändern personenbezogener Daten derart, dass die Einzelangaben über persönliche oder sachliche Verhältnisse nicht mehr oder nur mit einem unverhältnismäßig großen Aufwand an Zeit, Kosten und Arbeitskraft einer bestimmten oder bestimmbaren natürlichen Person zugeordnet werden können“ (§ 3 Abs. 6, BDSG). Des Weiteren gibt es die Möglichkeit zur Pseudonymisierung. Diese beschreibt „die Verarbeitung personenbezogener Daten in einer Weise, in der die Daten ohne Hinzuziehung zusätzlicher Informationen nicht mehr einer spezifischen betroffenen Person zugeordnet werden können, sofern diese zusätzliche Informationen gesondert aufbewahrt werden und technischen und organisatorischen Maßnahmen unterliegen, die gewährleisten, dass die Daten keiner betroffenen Person zugewiesen werden können“ (§ 46 Abs. 5, BDSG neu).

Liegt ein Personenbezug vor, dürfen Forschungsdaten nur mit der Unterzeichnung einer Einwilligungserklärung der Probanden erhoben und weiterverarbeitet werden.

Muster für Einwilligungserklärungen lassen sich hier finden:

<https://www.econstor.eu/dspace/bitstream/10419/97181/1/785263330.pdf>

Der Bereich Datenschutz ist ein weites Feld. Die TU Braunschweig hat eigens einen Datenschutzbeauftragten, der für jegliche Fragen rund ums Thema Datenschutz zur Verfügung steht (<https://www.tu-braunschweig.de/datenschutz>).

Publikationsvereinbarungen mit dem Repositorium

Vertragsbedingungen des Publikationsservers der TU Braunschweig:

<https://publikationsserver.tu-braunschweig.de/content/publish/contract.xml>

Repositorien stellen Nutzungsverträge /-vereinbarungen für die Publikation/ Archivierung von Forschungsdaten auf. Diese gilt es im Vorfeld gut durchzulesen und es sollte sichergestellt sein, dass alle angegebenen Vorgaben erfüllt sind.

Datennachnutzungsvereinbarungen

Zugriffsberechtigung

Es ist empfehlenswert, Forschungsdaten so offen wie möglich zur Verfügung zu stellen. Jedoch können die Daten auch mit einem Embargo versehen (Zugriffverzögerung für eine bestimmte Zeit) oder eingeschränkt werden, wer Zugriff auf die Daten erhält. Sensible Forschungsdaten, die z.B. personenbezogene Informationen enthalten, die sich nicht vollständig anonymisieren lassen, wären ein Fall für einen solchen eingeschränkten Zugriff.

Folgende Varianten bietet der Publikationsserver der TU Braunschweig für Forschungsdaten an:

1. Publikation (uneingeschränkter Zugriff)

Bei dieser Variante gibt es keine Beschränkungen für den Zugriff auf die Daten. Jeder kann die Daten einsehen und herunterladen. Durch die Wahl einer passenden Nutzungslizenz kann bestimmt werden, wie die Daten nachgenutzt werden dürfen.

2. Embargo (verzögerter Zugriff)

Ein Embargo kann für die Veröffentlichung des Datensatzes festgelegt werden. Bis zum Ende des Embargos ist der Datensatz gesperrt und für andere nicht zugänglich. Erst nach Beendigung der selbst gewählten Embargofrist wird der Datensatz veröffentlicht.

Die Verwendung eines Embargos kann sinnvoll sein, wenn die Veröffentlichung von Forschungsdaten verzögert werden soll, z. B. bis zur Veröffentlichung eines zugehörigen Zeitschriftenartikels oder aus kommerziellen Gründen bei der Anmeldung eines Patents.

3. Publikation (eingeschränkter Zugriff)

Verschiedene Arten von Zugriffsbedingungen können gewählt werden. Der Publikationsserver ermöglicht es z.B., dass Forschungsdaten nur von Mitgliedern der TU Braunschweig eingesehen werden können. Wenn das nicht genügt, kann auch ein eingeschränkter Zugriff ausgewählt werden, bei dem nur die Metadaten des Datensatzes einsehbar sind, der Datensatz selbst aber nicht. Bei Interesse eines Forschers kann sich dieser beim Autoren melden und dieser entscheidet individuell, ob der Forscher Zugriff zu dem Datensatz erhalten soll. Dies kann hilfreich sein, wenn rechtliche oder ethische Gründe es nicht erlauben, den Datensatz frei zur Verfügung zu stellen. Zudem ist es so auch möglich mit dem interessierten Forscher abzuklären, wofür er die Daten nutzen möchte.

4. Archivierung

Natürlich ist es auch möglich, dass die Forschungsdaten gemäß der guten wissenschaftlichen Praxis für mindestens 10 Jahre archiviert werden, ohne dass sie öffentlich gemacht werden. In diesem Fall können nur Mitarbeiter der UB den Datensatz einsehen, um ihn zu verwalten.

Lizenzen für Daten

Die Creative Commons (CC) Lizenzen sind weltweit anerkannt und sind ab Version 4.0 auch für Forschungsdaten im internationalem Kontext geeignet (<https://creativecommons.org/>). Durch sie kann auf simple Weise geklärt werden, wie ein Forschungsdatensatz genutzt werden darf, sodass andere Wissenschaftler über eine mögliche Art der Nachnutzung informiert sind.

Lizenz	Abkürzung	Beschreibung	Open-Access-konform
	CC0	Keine Rechte vorbehalten, gemeinfrei	✓
	CC BY	Namensnennung: uneingeschränkte Weiterverarbeitung und kommerzielle Nutzung erlaubt	✓

	CC BY-SA	Namensnennung + gleiche Lizenzvergabe (SA = share alike): uneingeschränkte Weiterverarbeitung und kommerzielle Nutzung unter gleicher Lizenzvergabe	✓
	CC BY-NC	Namensnennung + nicht kommerziell: uneingeschränkte Weiternutzung aber nicht zu kommerziellen Zwecken ist erlaubt	✗
	CC BY-NC-SA	Namensnennung + nicht kommerziell + gleiche Lizenzvergabe: uneingeschränkte Weitervergabe für nichtkommerzielle Zwecke ist unter der gleichen Lizenzvergabe gestattet	✗
	CC BY-ND	Namensnennung + keine Bearbeitung: eine kommerzielle Nutzung ist gestattet, aber keine Bearbeitung	✗
	CC BY-NC-ND	Namensnennung + nicht kommerziell + keine Bearbeitung: Weiterverarbeitung und kommerzielle Nutzung sind nicht gestattet	✗

Um den Missbrauch von Forschungsdaten zu verhindern, ist es zudem möglich, den Datensatz zunächst nicht unmittelbar abrufbar zu machen, sondern erst eine Kontaktierung des Urhebers zu fordern, sodass dieser bei berechtigtem wissenschaftlichem Interesse eine Freigabe erlauben kann. In diesem Falle ist zunächst nur die beschreibende Metadatenfile frei zugänglich.

ACHTUNG: Eine einmal vergebene Lizenz kann nicht wirksam zurückgezogen werden!

Lizenzen für Software

Für erstellte Software gibt es eigene Open-Source-Software-Lizenzen. Gebräuchliche Lizenzen sind:

GPL: GNU General Public License

LGPL: GNU Lesser General Public License

BSD: Berkeley Software Distribution

MIT: MIT-Lizenz

Eine ausführliche Auflistung von Open-Source-Lizenzen findet sich hier:

<https://ifross.github.io/ifrOSS/Lizenzcenter>

Hilfe bei der Wahl der passenden Lizenz

<https://creativecommons.org/choose/?lang=de>

<https://choosealicense.com/>

Weiterführende Links und Literatur

Hilfreiches Flussdiagramm „Forschungsdaten veröffentlichen“ von forschungsdaten.info:

<https://www.forschungsdaten.info/themen/veroeffentlichen-und-archivieren/daten-publizieren/entscheidungshilfe-daten-veroeffentlichen/>

Allgemeine Infos zum Thema:

<https://www.forschungsdaten.info/themen/rechte-und-pflichten/>

https://tu-dresden.de/gsw/jura/igetem/jfbimd13/ressourcen/dateien/dateien/DataJus/DataJus_Zusammenfassung_Gutachten_12-07-18.pdf?lang=de

5) Publizieren

Sind alle vorangegangenen Punkte bedacht und umgesetzt? Dann steht einer Archivierung bzw. Publikation der Forschungsdaten nichts mehr im Weg. Eingereicht werden die Daten entweder selbstständig per Shibboleth und der TU Braunschweig Kennung auf dem [Publikationsserver](#) oder ein Mitarbeiter der UB (forschungsdaten@tu-braunschweig.de) gibt Unterstützung dabei (z.B. bei größeren Datenmengen).

Nach der Einreichung der Daten werden diese durch die Mitarbeiter der UB geprüft und wenn alles in Ordnung ist, werden die Forschungsdaten archiviert oder mit einer DOI versehen und veröffentlicht.