Einstieg ins FORSCHUNGSDATENMANAGEMENT in den Geowissenschaften

DFG PROJEKT **EWIG**

Freie Universität Berlin

GFZ Helmholtz-Zentrum Potsdam

ZIB

INHALT

1. FORSCHUNGSDATEN IN DEN GEOWISSENSCHAFTEN 2

Was sind Forschungsdaten? 2
Was ist Forschungsdatenmanagement? 3
... und was habe ich davon? 3

2. WIE ORGANISIERE ICH DATEN? 4
VON DER ERHEBUNG BIS ZUR NACHNUTZUNG

Bestehendes Datenmanagement – Institutionelle Infrastruktur 4
Datensicherung – Datenversionen 4
Richtlinien 5
Datenerhebung – vorab zu klärende Fragen 5
Metadaten – Dokumentation 6
Analyse 7
Archivierung und Publikation 7

3. WELCHE WERKZEUGE HELFEN MIR? 8
HILFSMITTEL UND STRATEGIEN IM DATENWORKFLOW

Verteilte Datenerfassung und Kollaboration 8
Sicherung der Datenqualität 8
Bedeutung der Formate 9
Reproduzierbarkeit der Bearbeitung 9
Nachnutzung durch Publikation der Daten 10

4. WARUM MUSS ICH MICH UM DIE KOSTEN KÜMMERN? 11
WERTSCHÖPFUNG UND KOSTENSCHÄTZUNG

Kriterien zur Bestimmung des Wertes der Daten 11
Kalkulation der Kosten für das Datenmanagement 12
Wer übernimmt die Kosten für das Datenmanagement? 16

5. WO KANN ICH DIE ANREGUNGEN VERTIEFEN? 17
ANHANG, TABELLEN, GLOSSAR

Weiterführende Literatur 17
Ausgewählte Internetquellen zu Datenportalen und Metadaten 19
Ausgewählte Internetquellen zu Werkzeugen 20
Kurzes Glossar 21

Nachwort 23
Impressum 24

1. FORSCHUNGSDATEN IN DEN GEOWISSENSCHAFTEN

Der Prozess des wissenschaftlichen Arbeitens ist in den empirisch basierten Geowissenschaften immer verbunden mit der Erzeugung von Forschungsdaten – als Ergebnis von Messungen oder Probenbeschreibungen, aber immer häufiger auch durch Neubearbeitung bereits vorhandener Daten. Dabei werden Daten zunehmend rein digital erzeugt, zumindest aber digital ausgewertet. Dass Forschungsdaten nicht nur die eigene Arbeit nachvollziehbar dokumentieren und damit die eigene Forschungsleistung über Textpublikationen hinaus präsent machen, sondern auch anderen Wissenschaftlerinnen und Wissenschaftlern als Anregung dienen können, ist mittlerweile anerkannt. Die DFG verlangt zur Sicherung guter wissenschaftlicher Praxis nicht zuletzt eine sichere Aufbewahrung der „Primärdaten" der Forschung – und zwar mindestens für zehn Jahre. Seit 2010 werden Antragsteller einer DFG-Förderung gefragt, welche „Maßnahmen ergriffen werden, um die Daten nachhaltig zu sichern und ggf. für eine erneute Nutzung bereit zu stellen". Oft fällt erst am Ende eines Projekts auf, dass Maßnahmen zur Sicherung und Nachnutzung ergriffen werden müssen. Zu einem so späten Zeitpunkt erscheint das als gewaltige Aufgabe und bedeutet einen wesentlichen Mehraufwand.

Diese Handreichung soll aufzeigen, wie man von Beginn an im Umgang mit Daten sicherstellen kann, dass dieser „Schreckmoment" am Ende des Projekts nicht eintritt und man mit vertretbarem Mehraufwand archivierbare und nachnutzbare Daten erhält. Er soll als Anregung für Masterstudierende und (Post-) Doktoranden ebenso nützlich sein wie für Lehrende, die einen Kurs zum **Forschungsdatenmanagement** planen. Koordinierte Doktorandenausbildung in Graduiertenschulen und -kollegs sowie Summer Schools dienen nicht nur der Bildung von konkreten Fachkompetenzen, sondern unterstützen beim Projekt- und Zeitmanagement sowie im Umgang mit Forschungsdaten und Publikationen. Bereits im Vorfeld sollte im Rahmen von Veranstaltungen im B.Sc. und M.Sc. geeignet sensibilisiert werden, um ein Bewusstsein für einen kritischen und verantwortungsvollen Umgang mit Forschungsdaten und -ergebnissen zu schulen. Solche Ausbildungskomponenten sind hervorragende Plattformen, um die Forschungs- und Publikationskultur kennenzulernen und Erfahrungen im Austausch mit anderen zu sammeln.

Inhaltlich orientiert sich die Handreichung an den einzelnen Schritten im **Daten-Lebenszyklus,** der eine für alle Wissenschaftsdisziplinen grundlegend ähnliche Arbeitsweise als Kreislauf darstellt: eine Abfolge von Datenerhebung, Dokumentation, Analyse, Publikation, Archivierung und Nachnutzung von Forschungsdaten.

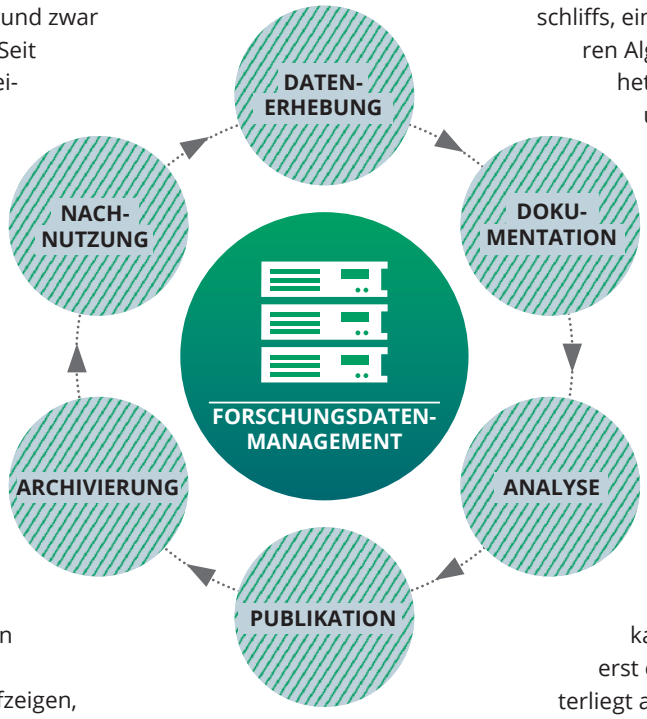


Abb. 1: Lebenszyklus von Forschungsdaten

Was sind Forschungsdaten?

Eine Tabelle mit Messwerten oder ein Datenbankeintrag, der Scan eines Bohrkerns, das Foto eines Dünnschliffs, eine Softwaresimulation und deren Algorithmus, ein umfangreiches heterogenes GIS-Projekt usw., die unterschiedlichsten Formate werden in den Geowissenschaften als Forschungsdaten erzeugt. Forschungsdaten bilden die eigentliche **Grundlage des Forschungsprozesses.** Es handelt sich um Produkte aus Experimenten, Messungen, Beschreibungen, Erhebungen, Quellenforschungen oder auch Befragungen, die der Überprüfung von Hypothesen dienen (s. auch Glossar im Anhang). Die digitale Publikation von Forschungsdaten hat erst eine kurze Geschichte und unterliegt auch aufgrund der Vielfältigkeit des Materials noch einem starken Wandel, während wissenschaftliche Artikel bereits in über 250-jähriger Tradition ihre heutige Form gefunden haben. Doch schon beim Probenahmedesign oder der Auswahl eines speziellen Sensors treffen Forschende eine Entscheidung über die weitere Nutzbarkeit von Daten. Diesen Prozess auszugestalten und zu dokumentieren und Forschungsdaten somit für die Zukunft nachnutzbar und interpretierbar zu halten, macht Forschungsdatenmanagement aus. Die eigentlichen Messwerte zusammen mit Metadaten bilden erst einen vollwertigen Forschungsdatensatz. Für Metadaten gilt: The more, the merrier. Metadaten enthalten in der einfachsten Form beispielsweise die Probennummer, eine Bezeichnung oder einen Titel, den (Fund-)Ort, das Datum der Erfassung, den Namen des Autors oder Bearbeiters. Damit Experimente und Messungen langfristig nachvollziehbar bleiben, sind darüber hinaus möglichst umfassende Informationen sinnvoll. In jedem Wissenschaftsgebiet innerhalb der Geowis-

senschaften haben sich spezielle Community-Standards herausgebildet und Empfehlungen, welche Metadaten für das Verständnis eines Forschungsdatensatzes notwendig sind. In der Klimaforschung hat sich beispielsweise daraus der CERA Standard (siehe Anhang) für die Beschreibung von Modelldaten etabliert.

Was ist Forschungsdaten-management?

Forschungsdatenmanagement umfasst kurz gesagt alle Methoden und Verfahren, die zur Sicherung der langfristigen Nutzbarkeit von Forschungsdaten angewendet werden [1]. Nachhaltiges Forschungsdatenmanagement zu betreiben ist nicht nur Aufgabe unterstützender Infrastruktureinrichtungen wie Universitätsbibliotheken oder Forschungsdatenzentren oder auch Verlagen, die Forschungsdaten in Form von „supplementary material" bereithalten. Die **Abgrenzung von Verantwortungsbereichen** lässt sich nach „Domänen" aufteilen [2]. Die Verantwortung für den korrekten Umgang mit produzierten Forschungsdaten beginnt in der privaten Domäne der Wissenschaftlerin oder des Wissenschaftlers und wächst mit der Sichtbarkeit der Daten hin zur Zugangsdomäne.

Abhängig von den Domänen haben die Phasen des Datenworkflows unterschiedliche Bedeutung. Datenerhebung, Dokumentation und wissenschaftliche Analyse werden individuell durch den oder die Wissenschaftlerinnnen oder Wissenschaftler verantwortet (private Domäne), dies geschieht oft über ein kollaboratives Arbeiten an und mit den Daten in einer begrenzten Arbeitsgruppe (Gruppendomäne). Publikation, Archivierung und öffentliche Nachnutzung dagegen sind durch geteilte Verantwortung zwischen den Forschenden und einer externen, institutionellen Gruppe gekennzeichnet, die an dieser Stelle Dienstleistungen für den Wissenschaftsbetrieb erbringt (Dauerhafte Domäne bzw. Zugangsdomäne). Die Verbindung der Domänen und ihre fließenden Übergänge ergeben das Gesamtbild einer Forschungsdateninfrastruktur. Soweit die Theorie: In der Praxis erfolgen die Übergänge zwischen den Domänen vielfach noch nicht so reibungslos, wie es im Sinn einer offenen und transparenten Wissenschaft wünschenswert ist. Deshalb steht am Beginn von Forschungsdatenmanagement die bewusste Ausgestaltung des Datenworkflows innerhalb des Lebenszyklus.

… und was habe ich davon?

Mit der Zugänglichkeit und im idealen Fall einer Publikation von Daten entstehen nicht nur für die Gesellschaft sondern auch für die Wissenschaft selbst konkrete Mehrwerte. An erster Stelle steht dabei die Steigerung der **Sichtbarkeit der eigenen Forschungstätigkeit.** Es hat sich gezeigt, dass Publikationen, deren zugehörige Daten offen verfügbar sind, signifikant häufiger zitiert werden [3]. Sichtbarkeit erhöht zudem das Potenzial für Kollaboration und Vernetzung. Zur Wahrung der Regeln der guten wissenschaftlichen Praxis können Wissenschaftlerinnen und Wissenschaftler Daten an Archive übergeben, ohne dass selbst weitere Maßnahmen zu ergreifen sind. Dadurch bleiben Daten und Metadaten langfristig verfügbar. Die so entste-

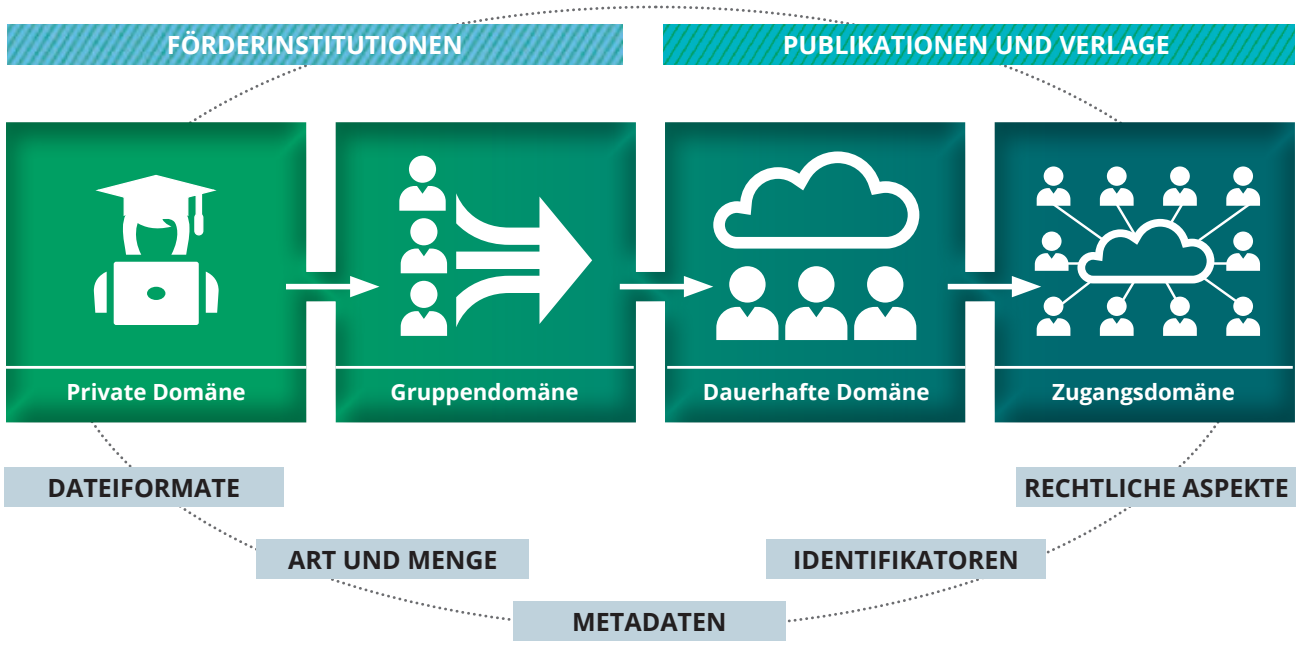


Abb.2. Domänen-Modell zum Forschungsdatenmanagement nach [2] (auch data curation continuum)

hende Datensammlung kann für Folgeprojekte interessant sein. Aus Sicht der Forschungsförderer ergibt sich durch ein solides Nachnutzungskonzept ein weiteres Kriterium zur Mittelvergabe. Synthesestudien werden erleichtert und es kann weitere wissenschaftliche oder kommerzielle Tätigkeit stimuliert werden. Zudem spart die Allgemeinheit Kosten, da dieselben Daten nicht doppelt erhoben werden müssen (falls das überhaupt möglich ist).
Wesentlich für die **Nachnutzbarkeit** ist dabei eine möglichst offene Verfügbarkeit von Forschungsdaten nach dem Prinzip der „Intelligent Openness" [4]. Hiermit verbindet sich die Forderung, dass Forschungsdaten zugänglich und auffindbar (accessible) sowie für Dritte verständlich (intelligible) sein sollen. Daten müssen zudem so aufbereitet und präsentiert werden, dass Aussagen durch Dritte überprüft werden können und ein Urteil über die Validität von Methoden und Ergebnissen möglich wird bis hin zur weiteren Verwendung der Daten. Um die Forderungen nach „Intelligent Openness" zu erfüllen, müssen die zu publizierenden Daten vorausschauend ausgewählt werden. Der potentielle Kreis der Nutzerinnen und Nutzer der Daten muss abgesteckt werden, denn schließlich stellen Datennutzer bestimmte Anforderungen an die Qualität und den Grad der Datenaufbereitung, sonst ist eine weitere Arbeit mit den Daten vielfach gar nicht oder nur mit erheblichem Aufwand möglich. Mit der Festlegung von Nutzungsrechten – in Form von Lizenzen – lassen sich dann zukünftige Anwendungsszenarien für Daten ausgestalten. Förderorganisationen weltweit haben ein starkes Interesse an der Bereitstellung von Daten über einen einfachen Zugang und versuchen Forschende bei der Veröffentlichung zu unterstützen. So hat die Europäische Union in Horizon 2020 dem Zugang zu im Projekt erstellten Daten eine hohe Wertigkeit gegeben [5]. Die National Science Foundation in den USA ändert ihr Vokabular bei Forschungsanträgen und möchte nun wichtige Produkte (Texte, Forschungsdaten, Software) anstatt nur Publikationen des Antragsstellers aufgeführt wissen. Auch die DFG und nationale Förderer in Großbritannien und Australien agieren ähnlich. Für Wissenschaftlerinnen und Wissenschaftler bedeutet dies zum einen sanft steigenden Druck von Seiten der Fördereinrichtungen, zum anderen steigen auch die Anreize für eine möglichst offene, nachhaltige Arbeitsweise.

2. WIE ORGANISIERE ICH DATEN?
VON DER ERHEBUNG BIS ZUR NACHNUTZUNG

Der Verlust von aufwendig erhobenen Daten wiegt umso schwerer, wenn dies kurz vor Fertigstellung einer Arbeit oder dem Abschluss eines Forschungsprojektes passiert. Manchem wird erst bei einer solchen Gelegenheit der Wert von Forschungsdaten bewusst. Selbst wenn die Messung wiederholbar sein sollte, sind gerade in der abschließenden Phase eines Projekts viele Informationen, die begleitend bei der Datenerhebung und Analyse angefallen sind, nicht mehr oder nur schwer zu reproduzieren. Ohne solche zusätzlichen Metadaten wird nicht nur die Bewertung der eigenen Forschungsergebnisse erschwert, sondern auch eine potentielle Nachnutzung weitgehend behindert bis unmöglich gemacht. Bei der Erhebung nur einmalig erfassbarer Daten, wie es bei vielen Zeitreihenmessungen der Fall ist, kommt unter Umständen ein Bruch in der Kontinuität der Messreihen hinzu. Eine durchdachte Datenablage kann die Suche nach Daten deutlich beschleunigen.

Bestehendes Datenmanagement – Institutionelle Infrastruktur

An erster Stelle steht die Orientierung über das an der Institution **bestehende Datenmanagement.** Der Einblick in das institutionelle Datenmanagement, im Allgemeinen der Gruppendomäne zugeordnet, erleichtert die Strukturierung des persönlichen Arbeitsbereiches. Dazu muss man die entsprechenden Ansprechpartner finden, auf diese zugehen und auch existierende Schulungen annehmen. Lehrende sind gefordert, einzelne Schritte im Workflow von der Datenerhebung bis zu deren Nachnutzung durch Beispiele aus der Praxis zu illustrieren. Die gängige Praxis sollte nicht gedankenlos übernommen werden, die Zweckmäßigkeit der einzelnen Arbeitsschritte, beginnend bei der Art und Weise der Datenhaltung bis hin zum gesamten Forschungsdatenmanagement, darf gerne kritisch vor einer Neuerhebung oder Ergänzung von Daten beurteilt werden. Trotzdem muss das Rad nicht neu erfunden werden. Die Nutzung der **institutionellen Infrastruktur** zur Verwaltung der Daten (z.B. Filesystemstruktur, Datenbanksystem) ist ebenso wichtig wie die Einbindung der Daten in institutionelle Datensicherungsmaßnahmen bereits während deren Erfassung.

Weiterführende Literatur [6], [9], [14], [15]

Datensicherung – Datenversionen

Unabhängig von den institutionellen Datensicherungsmöglichkeiten sollte man sich eine eigene Strategie zur **Datensicherung** (Backup) überlegen. So schnell wie heute mit aktuellen Rechnersystemen und Software große Mengen an Daten produziert werden, so schnell können diese auch wieder unbeabsichtigt verloren gehen, sofern keine Werkzeuge zur Datensicherung eingesetzt werden.

Das Risiko möglicher Datenverluste infolge von Hard- oder Softwareausfällen oder gar eigener Unachtsamkeit ist den meisten bewusst. Dennoch treten immer wieder Probleme mit verlorenen Daten auf. Mangelnde Lösungen zur Datensicherung zerstören die Ergebnisse der Arbeit oft vieler Monate – teilweise unwiederbringlich. Die kurz- bis mittelfristige Sicherung von Daten im Projekt ist nicht als Option anzusehen, sondern gehört zu einem verantwortungsvollen Umgang mit Forschungsdaten ebenso dazu, wie die geeignete Dokumentation der Ergebnisse. Es zahlt sich daher aus, bereits in der Anfangsphase seiner Arbeiten Überlegungen zur Sicherung der Forschungsdaten anzustellen.
Es empfiehlt sich, zu unterscheiden zwischen Daten, die ohne erhebliche manuelle Interaktion automatisiert reproduziert werden können, wie beispielsweise systematisch verarbeitete Daten aus Rohdaten, und Daten, deren Erstellung einen erheblichen zeitlichen und/oder intellektuellen Anteil individueller Arbeit erfordert haben (Texte, Grafiken, Karten, annotierte Abbildungen, manuelle Datenanpassungen). Hierzu gehören abgeleitete höherwertige Daten, die häufig am Ende einer langen Verarbeitungskette stehen, aber auch die eigenen Arbeiten für die Dokumentation und Publikation (Artikel, Präsentationen, Poster usw.). Ferner sollten mindestens drei Bearbeitungszustände der Daten unterschieden werden:

- **Originaldaten** sind Daten der Erhebung und damit nicht bereinigt. Als primäre Datenquelle sind diese von bearbeiteten Daten zu trennen und langfristig zu archivieren.
- **Bereinigte Daten** stellen die Daten dar, die nach Kontrolle und Korrektur aus den Originaldaten abgeleitet worden sind. Je nach Umfang der manuell durchgeführten Arbeiten werden diese bereinigten Daten geeignet gesichert und werden zur eigentlichen Analyse verwendet.
- **Analysedaten** sind die Daten, mit denen aktiv gearbeitet wird und von denen im Projektverlauf meist eine Reihe unterschiedlicher Kopien oder Modifikationen angelegt werden. Die veröffentlichten Ergebnisdaten sind ein Teil der Analysedaten.

Weiterführende Literatur [4], [5], [7]

Richtlinien

Für den Umgang mit Daten können von verschiedenen Seiten Richtlinien – in Form von Vorschriften, Regeln und Policies – wirksam werden, die eingehalten werden sollten. Hilfreich ist bereits im Vorfeld die Prüfung von entsprechenden Auflagen der Institution, des Geldgebers, aus bestehenden Kooperationsverträgen oder Absprachen in Forschungsverbünden. Solche Richtlinien regeln den Lebenszyklus der Forschungsdaten, also die Bereitstellung, den Austausch, die Verfügbarkeit, die Pflege, die Archivierung und die Nutzung der Daten. Die Sortierung der Daten nach ihrer Quelle (eigene, neu erhobene, von Dritten erworbene, ...), eine Aufteilung in Daten und Metadaten und die Definition von verschiedenen Datenebenen (Levels) vereinfachen die Zusammenarbeit mit den Partnern und sorgen für eine zeitnahe Nutzung der Daten. In diesen Richtlinien, Regeln oder Policies wird von der Verwendung der Daten innerhalb des Projektes bis hin zur Nachnutzung durch Dritte all das festgeschrieben, was der wissenschaftlichen Gemeinschaft einen einfachen Zugang zu Daten und Forschungsergebnissen ermöglichen soll. Hinzu kommen Vorgaben bei der Nutzung von Infrastrukturen (Geräte, Instrumente, usw.) als auch zu beachtende Auflagen von Zeitschriften, wenn die Daten als Supplement publiziert werden sollen.

Weiterführende Literatur [8], [10], [11], [12], [13]

Das Management von Daten, die in Forschungsprojekten ebenso wie bei einer Daueraufgabe genutzt, erhoben und/oder erzeugt werden, kostet Zeit und damit auch Geld. Beides muss von vornherein eingeplant werden. Die eingeplante Zeit muss dann auch genommen, manchmal sogar eingefordert werden. Nur so wird Datenmanagement ein selbstverständlicher Bestandteil des Forschungsalltags.

Datenerhebung – vorab zu klärende Fragen

Als **Orientierung bei der Datenerhebung** kann die Beantwortung folgender Fragen dienen:

Welche Typen von Daten werden generiert?
Die Einmaligkeit von Daten sollte als erstes bewertet werden. Geowissenschaftliche Beobachtungs- und Messdaten sind in der Regel einmalig, sofern nicht gleichzeitig mit gleichen Instrumenten die gleiche physikalische Größe am gleichen Ort gemessen wurde. Bei Labordaten spielen die Kosten der Reproduzierbarkeit eine wichtige Rolle. Bei Simulationen wiederum können das Modell und die Randdaten wertvoller sein als das berechnete Ergebnis. Abgeleitete und prozessierte Daten können häufig reproduziert werden, wenn alle Informationen zur Bearbeitung bekannt sind.

Welche Formate liegen vor bzw. werden erzeugt, gibt es standardisierte Formate, die bevorzugt zu verwenden sind?
Daten werden auf sehr unterschiedlichen Ebenen zu digitalen Objekten zusammengefasst. Dabei unterscheidet man zunächst die Art der Kodierung auf binärer Ebene. Ein oder mehrere Bytes beschreiben entweder einen Maschinenbefehl, einen Buchstaben oder eine Zahl. Der Inhalt eines digitalen Objektes erschließt sich erst, wenn man die Bedeutung dieser atomaren Elemente kennt; genau das wird durch das Format definiert. Das verwendete Format

sollte sich nach einem verbreiteten Standard richten und nach den Programmen, die man zur Bearbeitung einsetzen möchte.

Sind es genuin digitale Daten oder werden über analoge Objekte zusätzlich Eigenschaften digital erfasst?
Die Digitalisierung erschließt oft wenig zugängliche Datenquellen. Umso wichtiger sind alle Informationen, die mit der Erzeugung der digitalen Objekte zusammenhängen. Die nicht-digitalen Quellen sollten nach ihrer Digitalisierung möglichst verlustfrei gelagert werden.

Wie viele Daten sind zu erwarten?
Der Umfang der Daten kann über die Zeit anwachsen. Dies kann durch die Vervollständigung von Messreihen oder durch verschiedene Bearbeitungsstufen verursacht werden. Fallen verschiedene Bearbeitungsstufen an, sollte man diese mit einem Versionierungssystem nachvollziehbar machen.

Kann ich vorhandene Daten nachnutzen?
Die Nachnutzung von Daten verlangt neben deren akzeptabler Qualität und einer Abwägung des Aufwands der Beschaffung im Vergleich zur Neuerhebung auch die Recherche der Zugriffsmöglichkeiten und Urheberrechte.

Welche Qualitätsstandards sind bei der Auswahl und Aufbereitung von Daten zu beachten?
Zur Bewertung von Daten sollte man sich an Kriterien orientieren, die eine hohe Qualität erwarten lassen (dokumentiert, geprüft, korrigiert, reproduzierbar). Großen Stellenwert haben dabei sowohl das verwendete Format als auch die Standardisierung der Datenbearbeitung.

Für wen und zu welchem Zweck werden die Daten erhoben?
Oft gibt es potentielle Nachnutzer für neue erhobene Daten, die man bereits von Beginn an einbeziehen sollte, z.B. bei der Planung von Messkampagnen. Der Mehraufwand dafür sollte in einem guten Verhältnis zum dabei entstehenden Wertzuwachs der Daten stehen.

Wem gehören die Daten?
Bei der Nachnutzung von Daten sollten vorab Urheberrechte und Zugriffsrechte geklärt werden. Bei der Neuerhebung sind diese Rechte zu dokumentieren.

Wie lange sollen die Daten aufbewahrt werden?
Projektgeförderte Wissenschaft hängt immer an einem festen Zeitrahmen. Aber auch Qualifizierungsarbeiten oder Publikationen werden abgeschlossen. Damit endet in der Regel die formelle Verantwortung für die Daten des Vorhabens. Überlegungen zur Nachnutzung von Daten sollten deshalb bereits zu Beginn ihres Lebenszyklus einbezogen werden, damit man den Blick auf notwendige Zusatzinformationen nicht verliert.

Welche Werkzeuge sollen eingesetzt werden?
Die Datenerhebung, das Format und die Bearbeitung hängen sehr stark von der eingesetzten Software ab. Das zwingt teilweise zu einer Konvertierung zwischen verschiedenen Datenformaten. Dabei müssen die Metainformationen semantisch stabil bleiben, d.h. es sollte kein Bedeutungsverlust eintreten.

Weiterführende Literatur [1], [3], [4]
Tabelle „Datenportale“

Metadaten – Dokumentation

Wie beschreibe ich die Daten?

Daten ohne ergänzende Informationen sind wertlos! Erhobene Daten sind auch bei der persönlichen Nutzung und bei der Nutzung in der Gruppendomäne umso wertvoller, je mehr Informationen ihnen mitgegeben werden. Für eine langfristige Erhaltung und eine Nachnutzung sind solche beschreibenden **Metadaten** unverzichtbar. Eine Erfassung von Metadaten gleich bei der Erzeugung der Daten sichert darüber hinaus im Lebenszyklus die Konsistenz der Beschreibung. Forschungsdaten ohne ausreichende Metadaten sind nicht vollständig und entsprechen nicht der guten wissenschaftlichen Praxis. Solche beschreibenden Metadaten umfassen zuallererst Informationen zum Kontext, in dem Daten erhoben werden. Beispielsweise wird durch einen einprägsamen Titel (Name des Projekts, der Datenerhebung und dergleichen) die eindeutige Zuordnung und Einordnung in einen innerhalb der Domäne verfügbaren Katalog erleichtert. Weitere wichtige Informationen sind Name und Adresse der Institution oder der Personen, die an der Erhebung der Daten beteiligt sind, eine formlose Beschreibung der thematischen Inhalte, Angaben zum Zeitrahmen (Projektlaufzeit, Aufbewahrungsfristen), den Rechten der Datennutzung. Querverweise wie die Projektnummer, unter der die Datenerhebung durchgeführt wurde, erleichtern die Einordnung der Daten für Außenstehende. Zu geowissenschaftlichen Daten selbst gehören Angaben zur geographischen Verortung und zur zeitlichen Zuordnung, bei Messungen noch der Zeitpunkt oder die Zeitspanne der Datenerhebung. Die Anreicherung mit weiteren Metadaten sollte bei nachgenutzten Daten über die Herkunft, das Abrufdatum und die Zugangsquelle Auskunft geben. Die dynamische Anreicherung und Ergänzung während des Lebenszyklus ist hilfreich oder sogar erforderlich zur Beurteilung von Daten. Informationen zur Methode, dem Gerät, der verwendeten Software, den Verarbeitungsschritten, Protokolle der Experimente, Hinweise auf den Methodenteil relevanter Artikel, Versionsangaben, verwendete Datenformate, eine Liste mit allen Dateien des Datensatzes und Checksummen können die Metadaten sehr nützlich ergänzen. Je nach Fachrichtung, spezifischen Anforderungen oder internen/externen Vorgaben können weitere Metadaten erforderlich sein. Die Dokumentation

der Daten mit Metadaten kann in einem eigenen Versionsdokument erfolgen oder man nutzt einen entsprechenden bereits ausgearbeiteten und erfolgreich getesteten Metadatenstandard. Für die spätere Verwendung in Katalog- und Publikationssystemen ist es vorteilhaft, wenn sich die Metadaten gut in publikationsrelevante Metadatenschemata wie ISO19115, GCMD-DIF oder Dublin Core abbilden lassen. Als integrierendes Schema beginnt sich durch weltweite Standardisierungsprozesse und gesetzliche Rahmenbedingungen die ISO19115/ISO19139 durchzusetzen. Gute Anlaufstellen für verbreitete Daten- und Metadatenformate sind die GEOSS-Registry und die Formatregistry der Library of Congress, wobei letztere eher auf die Archivierung abzielt.

Weiterführende Literatur [7], [9], [17]
Tabelle „Metadatenschemata“

Je weiter man sich von der systeminternen Ebene entfernt, umso abstrakter kann sich der Zugriff auf die digitalen Objekte gestalten. Das entbindet einerseits von sehr technischen Entscheidungen, schafft aber eine umso stärkere Abhängigkeit von Informationsinfrastrukturen und somit zum Teil sehr spezifischer Software. Im einfachsten und häufigsten Fall wird man Daten auf der Dateisystemebene ordnen. Der Einsatz einer Datenbank oder die Nutzung einer komplexen IT-Infrastruktur verlangt einen Einsatz von Ressourcen, die nicht immer vorhanden sind. Beginnt man bei der **Datenablage** mit der Einrichtung von Ordnern und der Namensgebung von Dateien, so sollte man die verwendete Semantik dokumentieren. Kodiert man im Dateinamen bereits wichtige Metadaten, kann das beim Umsortieren der Daten zum Verlust der Zuordnung zwischen Daten und Metadaten führen und die Daten wertlos machen. Bei der Erfassung der Daten sollte der Arbeitsablauf so gestaltet werden, dass Metadaten automatisch oder sehr bequem mit eingepflegt werden können und eine stabile Bindung zwischen Daten und Metadaten erstellt wird. Der Weg zu selbst beschreibenden Datensammlungen sollte nach Möglichkeit durch einen Datenmanagementplan begleitet werden.

Weiterführende Literatur [14], [16]

Analyse

Eine technische Abgrenzung von Datengruppen, wie sie für die Datensicherungsstrategie sinnvoll ist, sollte auch für die (Zwischen-) Ergebnisse der wissenschaftlichen **Analyse** vorgenommen werden. Die Gruppierung nach Wertschöpfung erleichtert die spätere Bewertung der Daten. Die Nachvollziehbarkeit der Arbeitsschritte bildet eine Grundlage guter wissenschaftlicher Praxis und der Workflow sollte als Teil der Metadaten erfasst werden. Die Effizienz des wissenschaftlichen Arbeitsprozesses wird gegenwärtig nach der Anzahl der Veröffentlichungen bemessen. Die Publikation, als Abschluss einer Forschungsarbeit bzw. als ein Produkt der wissenschaftlichen Arbeit, umfasst neben der auf die Analyse von Daten bezogenen Veröffentlichung der Ergebnisse in Textform auch eine Publikation der zugrunde liegenden Daten. Datensupplemente zu Textpublikationen haben, für junge Forschende vielleicht überraschenderweise, eine lange Tradition. Aus Kostengründen verschwanden sie aber im Lauf der Jahre aus den damals noch ausschließlich papiergebundenen Zeitschriften. Viele Verlage fordern allerdings heute wieder zugehörige Supplementdaten in digitaler Form ein. In zunehmendem Maß wird dabei auf Datenrepositorien verwiesen.

Weiterführende Literatur [6]

Archivierung und Publikation

Die **Archivierung** von Forschungsdaten soll möglichst transparent sein. Das bedeutet dokumentiert, registriert und zugänglich, damit eine Nachnutzung vereinfacht wird. Dabei müssen Rechte und Zuständigkeiten geklärt sein und man sollte nach folgender Regel vorgehen: „So offen wie möglich, so geschlossen wie nötig“. Bei einer durchaus mehrstufigen Archivierung im Hinblick auf potentielle Nutzergruppen orientiert man sich am data curation continuum (siehe Abb. 2, Kapitel 1). Gerade in den Geowissenschaften gibt es eine lange Tradition, Forschungsdaten für die Community zugänglich zu machen. Beispiele dazu findet man in der Liste von Datenportalen im Anhang. Datenrepositorien erweitern gegenwärtig ihre Anwendungen, um alle oder einzelne Datenpakete in ihren Systemen publizierbar und zitierbar zu machen. Aber auch die eigenständige Veröffentlichung generierter Datensätze in sogenannten **Data Journals** gewinnt zunehmend an Bedeutung. In den letzten Jahren ist eine ganze Reihe solcher Datenzeitschriften entstanden. Dies sind Zeitschriften, deren Fokus auf der Beschreibung selbständiger Datensätze oder -zusammenstellungen liegt, die zuvor in einem zuverlässigen Datenrepositorium publiziert worden sind. Inzwischen beginnt sich darüber hinaus eine Kultur der Publikation von Forschungsdaten analog zur Textpublikation zu entwickeln. Mit der Möglichkeit der Zuordnung von persistenten Identifikatoren – wie DOIs – zu Forschungsdaten werden publizierte Forschungsdatenpakete zitierbar gemacht. Grundvoraussetzung ist dabei die Zuordnung elementarer Metadaten im Stil traditioneller Textpublikationen (Autoren, Titel, Erscheinungsjahr, Quelle). Die Vorbereitung solcher Datenpublikationen erfordert neben einer gezielten Auswahl der Forschungsdaten auch einen quasi redaktionellen Prozess, der mit einer Qualitätsprüfung der Daten einhergeht. Dieser Mehraufwand führt letztendlich zu einer Wertsteigerung der Daten und erhöht deren Sichtbarkeit.

Weiterführende Literatur [7], [9]
Tabelle „Portale zur Datenpublikation“

3. WELCHE WERKZEUGE HELFEN MIR?
HILFSMITTEL UND STRATEGIEN IM DATENWORKFLOW

Ehe man mit der **Erfassung von Daten** beginnt, sollte zunächst der Datenworkflow beschrieben werden. Diesen kann man u.U. einem Datenmanagementplan entnehmen oder ihn selbst entwerfen. Der Workflow erfasst mögliche Datenformate, die bei einer Messung oder Erhebung entstehen und stellt Programmkomponenten mit ihren Datenformaten dar, die später für die Bearbeitung genutzt werden können. Daten, die neu erfasst werden, sollten sofort mit Metadaten versehen werden. Natürlich sollte dies in der Arbeitsumgebung praktikabel sein, denn eventuell empfiehlt es sich im Gelände nicht immer, ein Tablet oder einen Laptop in diese „elektronikfeindliche" Umgebungen mitzunehmen. Falls die Daten nicht direkt digital bearbeitet und angereichert wurden, sollte dies unbedingt zeitnah nachgeholt werden. Gerade bei der Erfassung der Metadaten werden Aufwand und Kosten dafür gern vernachlässigt. Das macht die Daten später sehr oft nahezu wertlos. Im erstellten Datenworkflow sollte besonderes Augenmerk auf Einschränkungen der verwendeten Programme und Datenformate gelegt werden. So waren ältere Versionen von Tabellenkalkulationen nicht in der Lage, Arbeitsblätter mit mehr als 65535 Zeilen zu bearbeiten. Wollte man etwa sehr große tabellarische Daten auswerten, konnte das zum Problem werden, aber auch beispielsweise Grafikprogramme können bei Bildern ab einer bestimmten Größe ihren Dienst einstellen. Informationen über Programme und Datenformate für den Datenworkflow zu sammeln, sollte im Gespräch mit Kollegen beginnen. Mitunter sind die Grenzen beliebter Softwarelösungen bekannt aber nicht dokumentiert. Ebenso kann es sein, dass in dem eigenen Fachgebiet nur mit einem bestimmten, schon etablierten Workflow gearbeitet wird.

Verteilte Datenerfassung und Kollaboration

Angefangen bei der Ausgestaltung gemeinsamer Projekte und Projektanträge über die Durchführung von Forschungsarbeiten bis zur Publikation der Ergebnisse in Zeitschriften oder Projektberichten sind heutzutage kaum Arbeiten in vollständiger Isolation durchführbar. Im Gegenteil, nie schien der Erfolg der Forschung mehr von einem kollaborativen Gedanken getragen worden zu sein als heute, und die Finanzierung von größeren Forschungsvorhaben ist ohne eine hervorragend vernetzte Projektstruktur kaum denkbar. Je größer diese Netze werden, desto wichtiger ist es, effiziente Werkzeuge zur Verfügung zu haben, die den Austausch von Daten und Dokumenten ermöglichen und alle Projektteilnehmer und Mitautoren nicht nur mit einem stets aktuellen Stand der Arbeiten in nahezu Echtzeit versorgen, sondern dies auch plattformübergreifend und jederzeit zugreifbar zu leisten vermögen. Das Bereitstellen und Verteilen von Daten, auch File-Sharing, ist eine universale Methode, Daten schnell zugreifbar und effizient in Form einer Ordnerstruktur bereitzustellen. Das Deutsche Forschungsnetz, Firmen aus Deutschland und kommerzielle Anbieter aus dem Ausland bieten mit verschiedenen Geschäftsmodellen und rechtlichen Implikationen Online-Speicher an und erfreuen sich wachsender Beliebtheit. Meist sind auch ältere Versionen von Dateien wieder herstellbar – eine richtige Versionierung oder Datensicherung wird jedoch oft nicht garantiert. Jenseits der Verteilung von Daten werden auch eine Reihe dedizierter Funktionalitäten zur Projekt- und Wissensorganisation, dem kollaborativen Code-Management, dem gemeinsamen Schreiben an Dokumenten oder zum Management von Literatur- und anderer Datenquellen auf Online-Plattformen bereitgestellt. Nicht zuletzt sind inzwischen zahlreiche Communities im Bereich des wissenschaftlichen Social Networkings entstanden. Trotz hervorragender Angebote im Internet sollte die Speicherung von Daten in ‚der Cloud' stets mit hohem Verantwortungsbewusstsein gegenüber der Forschung, den Förderern und den Heimatinstituten geprüft werden und die Frage nach der Sicherung und Zugreifbarkeit der Daten und den Besitzrechten recherchiert werden, bevor Forschungsdaten oder -ergebnisse in nahezu unbekannte Hände gegeben werden. In den Allgemeinen Geschäftsbedingungen kommerzieller Anbieter etwa wird diesen gern das Recht eingeräumt, Benutzer und ihre Daten unter bestimmten Umständen zu löschen. Auch sollte man immer im Hinterkopf behalten, dass das Geschäftsmodell einiger Firmen das Sammeln und Auswerten von Daten ihrer Nutzer ist. Daher ist das Studium der Daten- und Datenschutzrichtlinien vor der Entscheidung obligatorisch.

Weiterführende Literatur [3], [4], [5], [6]

Sicherung der Datenqualität

Bei der Beurteilung der Qualität von Daten sollte man sich verdeutlichen, dass die Entscheidung über deren Hochwertigkeit oder Unbrauchbarkeit immer der potentielle Nutzer trifft. Diese Entscheidung kann von ihm nur dann sinnvoll getroffen werden, wenn der Werdegang der Daten dokumentiert ist. Die **Qualität der Metadaten** spielt dabei die entscheidende Rolle, d.h.

- **wer** hat
- **wann,**
- zu welchem **Zweck,**
- **was** und
- **womit**

gemessen oder modelliert, muss darin enthalten sein. Oft sind Metadaten implizit durch den Auftrag eines Messnetzes oder einen Projektkontext gegeben und in einer wissenschaftlichen Veröffentlichung dokumentiert. Trotz allem sollten diese Metadaten in einem geeigneten Format auch die Datensätze begleiten. Stehen Datensätze aber gar ohne schriftliche Publikation für sich allein, müssen diese Informationen unbedingt dokumentiert sein.

Gute Metadaten sind noch keine Garantie für die **Qualität von Daten** selbst. Es ist zwingend erforderlich, die Daten selbst einer Qualitätskontrolle zu unterziehen. So können beispielsweise Messwerte auf Plausibilität kontrolliert und bei Vorliegen von Messfehlern gelöscht oder interpoliert werden. Dieser Vorgang lässt sich in vielen Fällen automatisieren und der Einsatz von Software ist möglich. Trotzdem sollte man diesen Automatismus stichprobenartig prüfen. Wenn beispielsweise nur unbearbeitete Datenströme zur Nutzung bereitgestellt werden, ist die Definition der Qualität von Daten schwierig und unter Umständen sogar ein Streitpunkt. In einer Zeit wachsender Anforderung an die Wissenschaft, hochfrequent und gleichzeitig qualitativ hochwertig zu veröffentlichen, stellt die Qualitätssicherung eine zunehmend größer werdende Herausforderung dar. Die dahinter stehende Motivation umfasst dabei die Bewertung sowohl bereits existierender wissenschaftlicher Daten, abgeleiteter Ergebnisse und publizierter Literatur als auch die Bewertung der eigenen Daten und wissenschaftlichen Ergebnisse. Gerade in den Geowissenschaften stellt die Qualitätssicherung bei den üblicherweise hohen Datenaufkommen eine zusätzliche Herausforderung an Strategien und Werkzeuge dar.

Bedeutung der Formate

In jeder Disziplin gibt es Präferenzen für einen de-facto Standard im Umgang mit Daten und deren Format. Einige Disziplinen arbeiten diesbezüglich bereits mit wohldefinierten Regeln, in anderen Forschungszweigen befinden sich diese erst im Aufbau. Es empfiehlt sich, vor jeder Datenerhebung die verwendeten Datenformate kritisch zu betrachten. Lassen sich beispielsweise Daten bequem austauschen und können diese von anderen gelesen werden? Sind die Formate unabhängig von den Werkzeugen einsetzbar (bspw. GeoTiff, Shapefiles, NetCDF) oder werden die Formate im Zuge von Softwareaktualisierungen ebenfalls verändert und sind damit nicht mehr für ältere Versionen lesbar? Tritt gerade bei kommerzieller Software das Problem proprietärer Formate auf, so lassen sich vielleicht die verwendeten Formate in offen standardisierte Formate umwandeln. Bei der Umwandlung der Formate dürfen keine Metadaten verloren gehen. Das passiert oft, wenn keine äquivalente Beschreibung der Attribute möglich ist. Offene und weitverbreitete Formate sind proprietären stets vorzuziehen, wenn sie das Gleiche leisten oder mit wenig Aufwand entsprechend genutzt werden können. In wenigen Jahren ist ein Softwarepaket und mit ihm das Datenformat unter Umständen vom Markt verschwunden, während etablierte systemunabhängige Formate weiterhin Verwendung finden. Grundsätzlich sollten Abhängigkeiten für die erfolgreiche Weiterverwendung von Daten weitestgehend minimiert werden, d.h., die Daten sollten systemunabhängig und ohne zusätzlichen Installationsaufwand auf dem Empfängersystem lesbar sein.

Reproduzierbarkeit der Bearbeitung

Forschungsinstitutionen und Universitäten stellen heutzutage nahezu immer kostenfreie Möglichkeiten der Datensicherung über das lokale Rechenzentrum oder einen anderen Dienstanbieter bereit. Im Gespräch mit der lokalen Systemadministration sollte bereits frühzeitig die Art und der Umfang einer Sicherung (vollständig vs. inkrementell) diskutiert werden, um eine gemeinsame Strategie zu entwickeln. Möglicherweise sind sehr umfangreiche zentral durchgeführte **Backup-Lösungen** mit Kosten verbunden, sodass die Entwicklung einer geeigneten Strategie nicht nur kostenrelevant ist. Bei der persönlichen Datensicherung sollte sichergestellt werden, dass mindestens eine aktuelle Kopie vorhanden ist, auf die im Ernstfall problemlos zugegriffen werden kann. Ist diese Kopie auf einem externen Medium, wie beispielsweise einem USB-Gerät oder gar auf den Servern eines Cloud-Anbieters, sollte sichergestellt werden, dass die Daten verschlüsselt sind und – gerade bei Cloud-Diensten – keine Besitzrechte an Externe abgegeben werden (s. AGB/Terms of Service).

Bei umfangreichen und kollaborativen Arbeiten und auch bei wesentlicher Änderung einer Bearbeitungsstrategie empfiehlt sich die Einführung eines **Versionierungssystems** (sofern das in der Community praktikabel ist). So ein System lässt sich zur Speicherung von kleinen Arbeitsschritten (ein commit, oder jeweils nach Abschluss eines Tests, etc.) wie auch zur parallelen Entwicklung an einem Projekt einsetzen. Eine Versionskontrolle erlaubt auch nach einem längeren Bearbeitungszeitraum nicht nur die vielleicht letzte Version vom Vortag wiederherzustellen, sondern auch problemlos auf eine frühere Version eines früheren Entwicklungszweigs zuzugreifen. Die einzelnen Arbeitsschritte (commits) werden in der Regel automatisch nummeriert. Die ebenfalls wichtige Nummerierung von selbst definierten Bearbeitungsstufen und/oder Projektzielen basiert auf folgendem Schema: Nach DDI-Standard [2] empfiehlt sich eine dreiteilige Versionsnummerierung „<Major>.<Minor>.<Revision>“, wobei wesentliche Änderungen und erreichte Arbeitsziele durch den Zähler <Major>, beginnend bei „1“, und nachfolgende untergeordnete Schritte in den Zählern <Minor> und <Revision>, beginnend bei „0“ angezeigt werden (bspw. Version 1.0.1: erste Revision der Hauptversion 1).
Wird eine einheitliche Strategie erarbeitet und konsequent verfolgt, erfordert eine Datensicherung kaum persönlichen

Zusatzaufwand und stellt sicher, dass nicht nur Daten bei technischen Ausfällen jederzeit wiederhergestellt, sondern auch Bearbeitungsstände früherer Projektzeitpunkte beliebig aufgerufen werden können.

Weiterführende Literatur [1,2]

Nachnutzung durch Publikation der Daten

Eingehend sollte geklärt werden, wer der Urheber der Daten ist und wer, etwa im Rahmen von Forschungskooperationen Rechte geltend machen kann. Dabei ist allerdings zu beachten, dass in Deutschland Messwerte nicht schützbar im Sinne des Urheberrechts sind, da einem Wert, den ich der Natur entnehme, keine **Urheberschaft** zuzuordnen ist (der juristische Begriff ist hier „mangelnde Schöpfungshöhe"). Für eine Vertiefung des Themas in internationalem Rahmen siehe **[13]**. Um die Nachnutzung von Daten zu ermöglichen, müssen diese entsprechend der potentiellen Nutzergruppe aufbereitet, mit Metadaten auffindbar, an einer zugänglichen Stelle abgelegt und unter klaren Nutzungsbedingungen (Lizenzen) freigegeben sein. Mit der Nutzungslizenz stehen und fallen spätere Anwendungsmöglichkeiten. Weiter- und Nachnutzung der Daten sind wissenschaftsimmanent wünschenswert und wissenschaftspolitisch erwünscht. Entsprechend sollte die Lizenz so offen wie möglich gewählt werden. Unterstützend wirken hier Webangebote wie die der Creative Commons, die es ermöglichen, anhand einer Kriterienliste die passende Lizenz auszuwählen. Sinnvoll ist es, einer Sammlung von Messdaten eine im Sinn des Open Access möglichst freie, gleichzeitig der wissenschaftlichen Praxis entgegenkommende Lizenz zu vergeben, beispielsweise die Creativ-Commons-Lizenz CC-BY. Damit wird der Nutzer dazu verpflichtet, den Autor/Datenerzeuger zu zitieren.

Die Aufbereitung von Daten findet immer im und für den jeweiligen Fachbereich statt. Sofern publiziert wird, ist es wünschenswert, ein offenes und archivierungsfreundliches Format zu verwenden, um die Erhaltung der Daten zu vereinfachen und deren spätere **Auffindbarkeit** zu gewährleisten. Ein passendes Vokabular zur Einordnung eines Datensatzes in die Teilbereiche der Geowissenschaften ermöglicht später eine komfortablere Suche. Für eine grobe Kategorisierung sind dabei der multilinguale GEMET Thesaurus und die GCMD Science Keywords der NASA beliebt (siehe Anhang). Für Spezialgebiete gibt es darüber hinaus weitere Thesauri.

Werden Daten öffentlich zugänglich gemacht, sollten sie in einem **Repositorium** abgegeben werden, das die Daten für Interessierte bereitstellt (Zugangsdomäne). Unter Umständen existiert eine solche Möglichkeit in der eigenen Institution. Einige Fachbereiche haben für die Verbreitung von Daten aber schon etablierte Repositorien und Archive, die für die Veröffentlichung und den Erhalt von Daten sorgen. Das vereinfacht die Übergabe entsprechender Forschungsdaten spürbar, da keine „Selbstverständlichkeiten" der jeweiligen Disziplinen geklärt werden müssen. Sollte noch nach einem Archiv gesucht werden, hat das re3data-Projekt (siehe Anhang) sicherlich ein Passendes erfasst. Um veröffentlichte Daten später zitieren zu können, ist eine eindeutige **Identifizierung** und eine zuverlässige Erreichbarkeit der Datensätze zu gewährleisten. Beides leisten heute etablierte Systeme der persistenten Identifikatoren. Persistente Identifikatoren wie etwa der DOI (Digital Object Identifier, siehe Anhang) ermöglichen es, einem Datensatz eine Art Nummer oder Zeichenkette als eindeutiges Kennzeichen zu geben. Dieser Identifikator dient anstelle der realen Internetadresse als Referenzlink auf den publizierten Datensatz. Ändert sich nun etwas am Aufenthaltsort der Daten, ändert man nur die Internetadresse der Zuordnung, der Identifikator und somit die Zitierweise bleibt erhalten.

Weiterführende Literatur [9,10,11,12]

4. WARUM MUSS ICH MICH UM DIE KOSTEN KÜMMERN? WERTSCHÖPFUNG UND KOSTENSCHÄTZUNG

„Sind das Forschungsdaten oder kann das weg?" – Forschungsdatenmanagement ist kein Selbstzweck. Vielmehr zielt es darauf ab, solche Daten zu erhalten und in ihrer Qualität zu sichern, die einen gewissen wissenschaftlichen Wert haben – sowohl für das Projekt, in dessen Kontext sie entstanden sind, als auch für Forschende, die die Daten nach Projektende in neuen Forschungskontexten nachnutzen wollen. So besteht ein wichtiger Schritt des Datenmanagements darin zu entscheiden, welche Daten „wertvoll" genug sind, über die Projektlaufzeit hinaus aufbewahrt zu werden. Eine erste Auseinandersetzung mit dieser Frage sollte bereits in der Planungsphase erfolgen, d.h. bevor Daten überhaupt erhoben werden. Denn das Datenmanagement kann zwar einerseits dabei helfen, die Qualität – und damit den Wert – von Daten nicht nur zu sichern, sondern zu steigern (z.B. durch umfassende Dokumentation); andererseits erfordert das Datenmanagement den Einsatz von Ressourcen. Diese sollten nur aufgewendet werden, wenn ein Nutzen für das Forschungsprojekt oder durch eine mögliche Nachnutzung der Daten zu erwarten ist. Beispiel: Die Forschenden entschließen sich, Daten, die mit einem fehlerhaft kalibrierten Gerät erhoben wurden, nicht oder nur kurzfristig aufzubewahren. In jedem Fall aber sollte die Entscheidung über die Löschung und deren Gründe dokumentiert werden.

Kriterien zur Bestimmung des Wertes der Daten

Der wohl größte Einflussfaktor auf den Wert und wissenschaftlichen Nutzen von Daten – sowohl im Projektkontext, als auch für die Nachnutzung – ist die **Qualität der Daten.** Hierbei handelt es sich um einen komplexen Begriff, denn die Datenqualität selbst wird wiederum von einer Reihe Faktoren beeinflusst. An dieser Stelle soll nur eine Auswahl genannt werden:

- **Objektivität:** Sind die Daten genau, konsistent und verlässlich? Dies ist u.a. abhängig von der verwendeten Messapparatur/dem Messinstrument (Kalibrierung, Algorithmen, etc.) und davon, ob die gewählte Methode der Datenerhebung angemessen ist und korrekt angewendet wurde.

- **Integrität:** Wurden die Daten vor Korruption / unautorisierter Veränderung geschützt und wurden alle vorgenommenen Änderungen dokumentiert?

- **Verständlichkeit:** Ist transparent und nachvollziehbar, wie die Daten entstanden sind und was sie bedeuten? Hier spielt die Dokumentation der Daten eine wesentliche Rolle (s.u.).

Die Nutzbarkeit (Nützlichkeit) der Daten im Projekt und darüber hinaus hängt stark von deren **Zugänglichkeit** ab. Dies umfasst sowohl technische, als auch intellektuelle Aspekte wie etwa Verständlichkeit. So kann der Wert von Daten dadurch steigen, dass sie besonders aufwändig aufbereitet wurden – z.B. um sie besser durchsuchbar zu machen. Zum Beispiel: Historische Messdaten werden digitalisiert; als einfaches gescanntes Bild, PDF mit OCR, in ein Tabellenformat übertragen oder als Datenbank mit zusätzlichen Funktionen gespeichert; in den Sozialwissenschaften werden Daten teilweise bis hinunter auf Variablenebene erschlossen, anstatt nur den Datensatz als Ganzes zu beschreiben. Von besonderer Bedeutung für die Zugänglichkeit von Daten sind:

- **Formate:** Das Format, in dem Daten vorliegen, beeinflusst wesentlich, wie mit diesen Daten gearbeitet werden kann (z.B. Auswertung oder Weiterverarbeitung in bestimmten Programmen) und damit den Wert, den diese Daten für die Forschung haben. Liegen die Daten in einem offenen oder einem proprietären Format vor? Handelt es sich bei dem verwendeten Format um einen Standard, der innerhalb der Community weit verbreitet ist? Zum Beispiel: In der empirischen Sozialforschung ist ein wichtiges Standardformat .sav, das Datenformat des proprietären Statistikprogramms SPSS. Obwohl es sich um ein proprietäres Format handelt, hat es in der Community eine große Verbreitung; damit sind Daten, die in SPSS-Formaten zur Nachnutzung zur Verfügung gestellt werden tendenziell „wertvoller", als weniger verbreitete Formate. Hier ist aber beispielsweise auch denkbar, dass ein Projekt eine Entscheidung für ein Format trifft, das weder offen, noch in der Community verbreitet ist, weil es ganz bestimmte Auswertungs- oder Weiterverarbeitungsmöglichkeiten bietet, eine solche Entscheidung sollte aber gut abgewogen werden. Vielfach bietet sich reines ASCII-Format an, das universell weiterverarbeitbar ist.

- **Dokumentation:** Die Qualität der Dokumentation hat wesentlichen Einfluss auf den Wert von Daten. Nur wenn ausreichend Kontextinformationen zum Forschungsprozess verfügbar sind, ist eine Nutzung der Daten im Projekt und darüber hinaus möglich. Ist die vorhandene oder geplante Dokumentation dazu geeignet, Datenerhebung und Analyse – den gesamten Forschungsprozess – transparent und nachvollziehbar zu machen?

Sind die Daten einmalig, oder lassen sie sich reproduzieren? Die **Erhebungskosten** sind den **Kosten für die Reproduzierbarkeit** gegenüber zu stellen, um den Wert zu ermitteln. Zum Beispiel: Meteorologische Messungen einer

Exkursion des Forschungsschiffs Polarstern lassen sich nicht reproduzieren; Messungen nach der Radiokarbonmethode lassen sich wiederholen, solange noch Proben des Materials vorliegen. Je schwieriger eine Wiederbeschaffung der Daten ist, desto größer ist ihr potenzieller Wert. Hierbei gilt es auch in die Überlegungen mit einzubeziehen, welche Kosten (z.B. Personal, Material) bei der Erhebung von Daten oder deren Wiederbeschaffung anfallen. Sind diese im Vergleich zu einer Aufbewahrung gering und ist es möglich, die Daten mit wenig Aufwand wiederzubeschaffen, so spricht dies gegen eine langfristige Aufbewahrung.

Bei der Überlegung, Daten anderen Forschenden zur Nachnutzung zur Verfügung zu stellen, gilt es abzuschätzen, welche **Relevanz** die Daten über den Kontext des eigenen Projekts hinaus haben könnten. Können sie zur Beantwortung anderer Forschungsfragen herangezogen werden? Haben die Daten ein „Verfallsdatum" und sind somit nur für einen bestimmten Zeitraum relevant? Die Beantwortung dieser Frage ist nicht einfach – es ist oft überraschend, in welchen anderen Kontexten und Disziplinen Daten plötzlich genutzt werden können, auch wenn das im ursprünglichen Erhebungskontext gar nicht vorgesehen war. Beispiel: Jemand hat Postkarten einer Gaststätte in den Alpen gesammelt. Über viele Jahre gibt es immer verschiedene Aufnahmen. Nun fällt einer Forschergruppe auf, dass im Hintergrund immer derselbe Gletscher zu sehen ist. Mit einem Mal werden die Postkarten zu einem Klimaarchiv, mit dessen Hilfe der Gletscherrückgang approximativ dokumentiert werden kann.

Zu beachten ist, dass auch **„projektexterne" Faktoren** bei der Entscheidung eine Rolle spielen, ob und wie lange Daten aufbewahrt werden müssen. Hier gibt es möglicherweise konkrete Vorgaben des Forschungsförderers, der eigenen Universität oder beteiligter Forschungsinstitute. So fordert beispielsweise die Deutsche Forschungsgemeinschaft (DFG), dass Primärdaten, die die Grundlage von Publikationen bilden, zehn Jahre aufbewahrt werden müssen [8].

weiterführende Literatur [1], [2], [3], [4]

Kalkulation der Kosten für das Datenmanagement

Datenmanagement ist ein Kostenfaktor. Allerdings handelt es sich nur zu einem gewissen Teil um zusätzliche Kosten, denn viele Aktivitäten des Datenmanagements sind ohnehin Bestandteil des Forschungsprozesses, insbesondere wenn dieser den allgemeinen Regeln guter wissenschaftlicher Praxis folgt [8]. Zusätzliche Kosten können aber anfallen, wenn besondere Maßnahmen ergriffen werden – z.B. um eine bessere Sicherung oder Nachnutzbarkeit von Daten zu gewährleisten, etwa durch den Einsatz besonderer Werkzeuge zur Erstellung der Dokumentation, der Entwicklung eines Repositoriums oder einer virtuellen Forschungsumgebung. Diese Kosten können und sollten in Förderanträgen berücksichtigt werden. Zur Beurteilung der Kosten müssen verschiedene Kostentypen herangezogen werden. Einen Ansatz liefert folgende Aufstellung:

- Personalkosten (auch für das Datenmanagement)
- Materialkosten, unterteilt in Kosten für Ausstattung (z.B. Server, Geräte) oder Werkzeuge (Software) und Gerätezeit, bei gemeinsam genutzten Großgeräten
- Dienstleistungskosten, ergeben sich aus: Gebühren etwa für nachgenutzte Daten, für die Übernahme der Daten in ein Langzeitarchiv, Publikationsgebühren, Schulungen
- Overhead / Gemeinkosten: das sind indirekte Kosten, die nicht unmittelbar dem Forschungsprojekt zugerechnet werden können. Hierbei handelt es sich insbesondere um Kosten, die im Zusammenhang mit einer Bereitstellung von Infrastruktur entstehen (Miete, Heizkosten, Strom, Telefon, etc.).

Bei der Kalkulation der anfallenden Kosten gilt es auch zu überlegen, ob es sich um einmalige oder regelmäßige Kosten handelt und wie lange diese ggf. anfallen. Zu beachten ist auch, dass bestimmte Datenmanagementkosten möglicherweise bereits über die Gemeinkosten abgedeckt sind, die je nach Praxis des Forschungsförderers ausgezahlt werden. Die folgende Checkliste soll Forschungsprojekten eine frühzeitige Orientierung über die Kosten für das Datenmanagement und die Veröffentlichung von Daten nach Projektende ermöglichen. Bei vielen der genannten Punkte gilt: Werden die entsprechenden Maßnahmen frühzeitig geplant und umgesetzt, fallen die Kosten für das Datenmanagement tendenziell geringer aus, als wenn erforderliche Maßnahmen nachträglich oder mit großer Verzögerung umgesetzt werden. Zum Beispiel: Wenn von vornherein ein Schema / Standard für die Metadaten und Dokumentation vereinbart wird, können alle Projektbeteiligten sich direkt daran orientieren. Wenn zunächst jede/r nach Gutdünken Metadaten anlegt, müssen Angaben nachträglich vereinheitlicht und eventuell „rekonstruiert" werden, weil wichtige Informationen fehlen. Die folgende Tabelle basiert auf [9]:

weiterführende Literatur [7], [8], [9]

Wertschöpfung und Kosten

Schritt im Lebenszyklus	Kostenfaktor	Erläuterungen/Kommentare
1. Datenerhebung	**Datenorganisation**	Wurde im Vorfeld der Datenerhebung festgelegt, wie Dateien benannt, geordnet und versioniert werden, so entstehen hierfür kaum zusätzliche Kosten. Kosten entstehen, wenn nachträglich Dateibenennungen vereinheitlicht oder Verzeichnisstrukturen zusammengeführt und/oder neu organisiert werden müssen.
	Datenbereinigung und -aufbereitung	Dies umfasst z.B. Maßnahmen zur Verifizierung/Validierung der erhobenen Daten, Qualitätskontrollen, etc. Erfolgen diese Maßnahmen direkt mit der Erhebung nach festgelegten Regeln, entstehen nur geringe zusätzliche Kosten. Kosten können auch anfallen, wenn Daten genutzt werden, die vom Projekt nicht selbst erhoben wurden. Diese Daten müssen u.U. ebenfalls aufbereitet werden (z.B. Harmonisierung von Daten aus unterschiedlichen Quellen).
	Datenzugang und -übermittlung	Wird spezielle Software oder Hardware benötigt, um Daten in der Erhebungsphase sicher an einen zentralen Speicherort zu übermitteln (z.B. aus dem Feld, von Mobilgeräten, etc.) oder um Forschenden den Remote-Zugang zu ermöglichen?
2. Dokumentation	**Metadatenerstellung und Dokumentation des Forschungsprozesses**	Muss ein Metadatenschema erarbeitet werden, oder kann ein bestehendes Schema verwendet werden? Sind bereits Anforderungen des Repositoriums, Forschungsdatenzentrums oder Langzeitarchivs bekannt, in welchem die Daten nach Projektende archiviert werden sollen und können von Anfang an berücksichtigt werden? Müssen Dokumentation und Metadaten nachträglich erstellt oder bearbeitet werden, ist das kostenintensiver, als wenn diese direkt bei der Entstehung der Daten erstellt werden.
3. Analyse	**Kollaboratives Arbeiten**	Wird zusätzliche Software benötigt, um die Zusammenarbeit im Projekt zu unterstützen und möglichst effektiv und transparent zu gestalten (z.B. Kommunikationsplattformen, virtuelle Forschungsumgebungen, etc.)?
4. Publikation	**Publikationskosten für Datensätze**	Sollen die Daten in ein Repositorium überführt werden und dort öffentlich online verfügbar sein, so entstehen evtl. abhängig vom gewählten Repositorium zusätzliche Kosten. Das Verzeichnis der Forschungsdatenrepositorien www.re3data.org gibt hier eine Übersicht.
5. Archivierung	**Digitalisierung von nicht-digitalen Objekten**	Wird zusätzliche Hard-/Software benötigt? Mit welchem Zeitaufwand ist zu rechnen – für die Digitalisierung als solche sowie Qualitätskontrollen und mögliche manuelle Nachbearbeitung, Aufbereitung von Materialien (z.B. Feldnotizen, ...) für die Archivierung oder Zugänglichmachung?
	Nachbereitung für ein Archiv	Werden die Anforderungen des Archivs von Anfang an berücksichtigt, fallen die zusätzlichen Kosten geringer aus, als wenn zum Projektende Daten/Dokumentation nachbearbeitet werden müssen. Müssen Daten und Dokumentation für die Archivierung in einem Forschungsdatenzentrum/Archiv/Repository besonders aufbereitet werden?

Schritt im Lebenszyklus	Kostenfaktor	Erläuterungen/Kommentare
5. Archivierung	**Datenkonvertierung**	Müssen Daten und Dokumentation in bestimmte Formate konvertiert werden? Auch hier gilt: Die Kosten sind deutlich geringer, wenn die Anforderungen des Archivs von Anfang an berücksichtigt werden. Wird für die Konvertierung zusätzliche Soft-/Hardware benötigt? Wie groß ist der Zeitaufwand für die Durchführung der Konvertierung und eine Qualitätskontrolle der Ergebnisse?
	Datenübergabe an ein Langzeitarchiv	Welche Zeit-/Personalressourcen werden für die Kommunikation zur Übergabe der Daten, Ausgestaltung der Lizenzvereinbarungen, Ausfüllen von Übergabeformularen, etc. benötigt?
	Regelmäßige oder einmalige Gebühren	Werden vom Langzeitarchiv oder Repositorium Gebühren erhoben? Wenn ja, handelt es sich um einmalige Kosten (z.B. weil der Ingest besonders aufwendig ist) oder um Kosten, die regelmäßig über die gesamte Archivierungsdauer anfallen. Wenn ja, für welchen Zeitraum fallen die Kosten an?
	Aufbereitung von Software	Ist im Rahmen des Projekts Software entstanden, die für eine Archivierung und mögliche Nachnutzung aufbereitet werden muss?
6. Nachnutzung	**Klärung von Rechten**	Zeit-/Personalaufwand für die Klärung der Frage, wer die Rechte an den Daten hat (aufwendiger bei Projekten, an denen verschiedene Institutionen beteiligt sind, insbesondere wenn wirtschaftliche Interessen berührt werden). In komplexen Fällen ggf. Kosten für eine Rechtsberatung.
	Anonymisierung personenbezogener Daten	Werden Daten früh anonymisiert, entstehen geringere Kosten, als wenn nachträglich anonymisiert werden muss. Der Aufwand (und damit die Kosten) hängen wesentlich von der Art der Daten ab (quantitativ vs. qualitativ, textuell vs. audio-visuell, ...).
	Lizenzen	Auswahl passender Lizenzen. Können modulare Lizenzen wie z.B. Creative Commons nachgenutzt werden (geringe Kosten), oder müssen eigene Lizenztexte und Vereinbarungen erstellt werden (höhere Kosten)?
	Vertrieb der Daten	Sollen die Daten über ein projekteigenes Repositorium zur Nachnutzung verfügbar gemacht werden, fallen zusätzliche Kosten an – z.B. für die Entwicklung, für zusätzlich benötigte Hard-/Software sowie personelle Ressourcen. Je nach angestrebter Lösung können die Kosten sehr unterschiedlich ausfallen. Zu berücksichtigen ist hier aber insbesondere, dass ein solches Repositorium auch nach Projektende Kosten verursacht.
	Nutzerbetreuung und -beratung	Wenn Daten vom Projekt von der eigenen Institution vertrieben werden sollen, dann müssen u.U. Ressourcen für den Support nach Projektende eingeplant werden.

Schritt im Lebenszyklus	Kostenfaktor	Erläuterungen/Kommentare
Phasen übergreifend	**Implementierung des Datenmanagements**	Abhängig vom Projektumfang ist es sinnvoll, eine Person zu bestimmen, die das Datenmanagement verantwortlich koordiniert. Welche personellen Ressourcen werden hierfür benötigt? Entstehen zusätzliche Kosten für Datenmanagement-Aktivitäten z.B. durch zusätzliche Projekttreffen zur Planung des Datenmanagements oder weil eine kollaborative Arbeitsumgebung aufgesetzt werden muss? Müssen Schulungen durchgeführt werden?
	Speicherplatz	Welche Speicherkapazitäten werden über die gesamte Projektlaufzeit und ggf. darüber hinaus benötigt? Werden diese Kapazitäten von einer am Projekt beteiligten Institution bereit gestellt oder müssen zusätzliche Anschaffungen getätigt werden? In diesem Fall entstehen Sachkosten für Server und Speicher, aber auch Kosten für die Einrichtung und Administration.
	Backupsystem	Die entstehenden Kosten sind davon abhängig, wie viele Kopien auf welchen Medien an welchen Standorten vorgehalten werden sollen. Reichen die Standardprozeduren der Daten haltenden Institution oder muss zusätzlicher Aufwand (mit zusätzlichen Kosten) betrieben werden?
	Datensicherheit	Welche Maßnahmen müssen zum Schutz der Daten und Dokumentation vor unautorisiertem Zugriff, Veränderung, Korruption getroffen werden (z.B. Identifizierungs- und Autorisierungsverfahren, Verschlüsselung, etc.)? Kann die ‚lokale' IT entsprechende Maßnahmen anbieten oder müssen zusätzliche Tools/Software angeschafft und neue Verfahren implementiert werden?
	Schulungen	Müssen Projektbeteiligte zum Thema Datenmanagement und dessen konkreter Umsetzung im Projekt geschult werden und welche Kosten fallen hierfür an? (Gebühren für Schulungsanbieter, Zeitaufwand, etc.).
	Entwicklung von Werkzeugen	Entstehen Kosten für die Entwicklung von Werkzeugen oder Software, die das Datenmanagement in allen Phasen des Lebenszyklus unterstützt?

Wer übernimmt die Kosten für das Datenmanagement?

In der Regel verteilen sich die Kosten für das Management, die Langfristsicherung und die Nachnutzung von Forschungsdaten auf die unterschiedlichen beteiligten Akteure:

Forschungsförderer: Es empfiehlt sich in Förderanträgen alle zusätzlichen Kosten zu beziffern, die im Rahmen des Projekts für das Management, die Archivierung und Nachnutzung von Daten entstehen. Diese können vom Forschungsförderer übernommen werden. So ermutigt die DFG beispielsweise Antragstellende, projektspezifische Kosten zu beantragen, die für eine Nachnutzung von Forschungsdaten entstehen **[10]**. Hierunter fallen auch Kosten für das Datenmanagement, wie die unter **[11]** aufgeführten Beispiele verdeutlichen. Auch Gebühren, die für die Publikation und Langfristsicherung von Daten und Dokumentation (in einem Data Journal, in einem Archiv, etc.) anfallen, können von Forschungsförderern übernommen werden, wenn sie im Projektantrag berücksichtigt werden.
Datenerzeugerinnen oder -erzeuger: Bei vielen Maßnahmen des Datenmanagements handelt es sich um Aktivitäten, die Teil des regulären Forschungsprozesses sind. Diese Kosten werden in der Regel über die Projektmittel gedeckt. Sofern keine zusätzlichen Mittel für das Datenmanagement beantragt wurden, müssen u.U. auch zusätzlich anfallende Kosten hierfür aus Projektmitteln gedeckt werden.
Institution (Grundfinanzierung): Hier findet man Angebote zu virtuellen Forschungsumgebungen und anderen Verfahren, die die (internationale) Zusammenarbeit verbessern sowie meist ein rudimentäres Angebot zum Backup der laufenden Projekte.
Datenzentren, Archive, Repositorien: Häufig werden die Kosten für eine Langfristsicherung und Verfügbarmachung von Daten von den jeweiligen Archiven, Datenzentren und Repositorien getragen, d.h., oft werden keine Gebühren für die entsprechenden Dienstleistungen erhoben (insbesondere wenn diese aus öffentlichen Geldern finanziert werden).
Datennutzerinnen und -nutzer: Die Nachnutzung von Forschungsdaten ist oft kostenlos – insbesondere, wenn es sich um Ergebnisse öffentlich geförderter Forschung handelt. Gelegentlich werden aber für die Nachnutzung von Daten Gebühren erhoben, insbesondere dann, wenn mit der Bereitstellung der Daten ein besonderer Aufwand verbunden ist. Zum Beispiel könnten Gebühren dafür verlangt werden, dass Daten auf bestimmte Weise individuell zusammengestellt und auf einen Datenträger gebrannt werden, da hierfür Personalkosten aufgewendet werden müssen.

weiterführende Literatur [10]

5. WO KANN ICH DIE ANREGUNGEN VERTIEFEN?
ANHANG, TABELLEN, GLOSSAR

Weiterführende Literatur

Kapitel 1: Forschungsdaten in den Geowissenschaften

[1] Corti, L, Van den Eynden, V, Bishop, L & Woollard, M (2014) **Managing and Sharing Research Data. A Guide to Good Practice.** Sage Publications, Los Angeles.

[2] **DFG-Projekt RADIESCHEN – Rahmenbedingungen einer disziplinübergreifenden Forschungsdaten-infrastruktur** (Hrsg.) (2013) http://dx.doi.org/10.2312/RADIESCHEN_005

[3] Piwowar HA, Day RS, Fridsma DB (2007) **Sharing Detailed Research Data Is Associated with Increased Citation Rate.** PLoS ONE 2(3): e308. http://dx.doi.org/10.1371/journal.pone.0000308

[4] The Royal Society (2012) **Science as an open enterprise: open data for open science.** The Royal Society Science Policy Centre report 02/12. The Royal Society, London. https://royalsociety.org/policy/projects/science-public-enterprise/Report/

[5] EU (2013) **Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020,** Version 16 December 2013, http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf

Kapitel 2: Wie organisiere ich Daten?

[1] **Geo-Fortschrittsbericht der Bundesregierung** (Okt/2012) http://www.imagi.de/SharedDocs/Downloads/IMAGI/DE/Geofortschrittsberichte/3_Fortschrittsbericht.pdf?_blob=publicationFile

[2] **Interpretation of the „full and open" access to and use of (geographic) data:** existing approaches, Living paper of the GEO Data Sharing Working Group (Okt/2013) http://earthobservations.org/documents/dswg/08_Interpretation of the full and open access to and use of geographic data existing approaches.pdf

[3] **ICSU (International Council for Sciences) Strategic Plan II,** 2012-2017 (2011) http://www.icsu.org/publications/reports-and-reviewus/icsu-strategic-plan-2012-2017/

[4] **Global Earth Observation System of Systems (GEOSS)** 10-Year Implementation Plan Reference Document (Feb/2005) http://earthobservations.org/documents/10-Year Plan Reference Document.pdf

[5] **MARS (Meteorological Archival and Retrieval System)** ECMWF (2013) http://www.ecmwf.int/en/learning/education-material/ http://old.ecmwf.int/services/computing/training/material/mars.pdf

[6] **Langzeitarchivierung von Forschungsdaten – Eine Bestandsaufnahme** (2012) http://nestor.sub.uni-goettingen.de/bestandsaufnahme/nestor_lza_forschungsdaten_bestandsaufnahme.pdf

[7] **Empfehlungen zur gesicherten Aufbewahrung und Bereitstellung digitaler Forschungsprimärdaten,** DFG (Jan/2009) http://dfg.de/download/pdf/foerderung/programme/lis/ua_inf_empfehlungen_200901.pdf

[8] verschiedene Leitfäden zur Entwicklung von Policies http://www.dcc.ac.uk/resources/policy-and-legal/policy-tools-and-guidance/policy-tools-and-guidance

[9] nestor Handbuch – **Eine kleine Enzyklopädie der digitalen Langzeitarchivierung** (2010) http://nestor.sub.uni-goettingen.de/handbuch/nestor-handbuch_23.pdf http://nbn-resolving.de/urn/resolver.pl?urn:nbn:de:0008-2010071949

[10] Research Information Network: **Stewardship of digital research data:** a framework of principles and guidelines (Jan/2008) http://www.rin.ac.uk/system/files/attachments/Stewardship-data-guidelines.pdf

[11] **International Polar Year 2007-2008 Data Policy** (Apr/2008) http://classic.ipy.org/Subcommittees/final_ipy_data_policy.pdf

[12] **Zusammenstellung von verschiedenen Daten Policies** (ICSU, CODATA) http://www.codata.org/committees-and-groups/international-data-policy-committee

[13] **Datenmanagementplan des TERENO Projekts** (2011)
http://teodoor.icg.kfa-juelich.de/coordination-teams-en/ct-data-management

[14] **Data Observation Network for Earth,** Schulungsmaterial
https://www.dataone.org/data-management-planning

[15] **Handbuch Forschungsdatenmanagement** (2011)
http://opus4.kobv.de/opus4-fhpotsdam/frontdoor/index/index/docId/208urn:nbn:de:kobv:525-opus-2412

[16] **Konzeptstudie Vernetzte Primärdaten-Infrastruktur für den Wissenschaftler-Arbeitsplatz in der Chemie,** (2010)
http://www.tib-hannover.de/fileadmin/projekte/primaer-chemie/Konzeptstudie_Forschungsdaten_Chemie.pdf

[17] **Technology Watch Report – Preserving Geospatial Data** (2009)
http://www.dpconline.org/component/docman/doc_download/363-preserving-geospatioal-data-by-guy-mcgarva-steve-morris-and-gred-greg-janee

Kapitel 3: Hilfsmittel und Strategien im Datenworkflow

[1] Jensen (2012): **Leitlinien zum Management von Forschungsdaten,** GESIS – Leibniz-Institut für Sozialwissenschaften

[2] **Data Documentation Initiative** http://www.ddialliance.org/

[3] **Sicherheit in Cloud-Diensten:** https://www.sit.fraunhofer.de/de/cloud-security/

[4] **Kriterien zur Auswahl von Cloud Diensten** https://itservices.msu.edu/dept/cont-collab/assets/collaborative-tools-report.pdf

[5] **Literaturverwaltungssysteme** http://www.slub-dresden.de/fileadmin/groups/slubsite/Service/PDF_Service/Literartur verwaltungsprogramme_im_Überblick.pdf

[6] **Literaturverwaltungssysteme** http://mediatum.ub.tum.de/?id=1223124

[9] **re3data** http://www.re3data.org

[10] **Kompetenzzentrum Forschungsdaten** http://www.komfor.net

[11] **web2rights** http://www.web2rights.org.uk/

[12] **Creative Commons** http://creativecommons.org/

[13] **„Safe to be open" Study on the protection of research data and recommendations for access and usage,** Edited by Lucie Guibault and Andreas Wiebe, Universitätsverlag Göttingen 2013.

Kapitel 4: Warum muss ich mich um die Kosten kümmern?

[1] Kindling, M., 2013. **Qualitätssicherung im Umgang mit digitalen Forschungsdaten.** Information – Wissenschaft & Praxis, 64(2-3), pp.137–147.
http://www.degruyter.com/view/j/iwp.2013.64.issue-2-3/iwp-2013-0020/iwp-2013-0020.xml

[2] Lee, Y.W. et al., 2002. AIMQ: **a methodology for information quality assessment. Information & Management,** 40(2), pp.133–146. http://linkinghub.elsevier.com/retrieve/pii/S0378720602000435

[3] **Office of Management and Budget,** 2002. Guidelines for Ensuring and Maximizing the Quality, Objectivity, Utility, and Integrity of Information Disseminated by Federal Agencies; Republication,
http://www.whitehouse.gov/sites/default/files/omb/fedreg/reproducible2.pdf

[4] Waaijers, L. & van der Graaf, M., 2011. **Quality of Research Data, an Operational Approach.** D-Lib Magazine, 17(1/2), pp.1–8. http://www.dlib.org/dlib/january11/waaijers/01waaijers.html

[7] **DataONE – Data Observation Network for Earth,** Provide budget information for your data management plan. Available at: https://www.dataone.org/best-practices/provide-budget-information-your-data-management-plan

[8] DFG, 2013. **Sicherung guter wissenschaftlicher Praxis / Safeguarding Good Scientific Practice.** Denkschrift / Memorandum, Weinheim: Wiley-VCH Verlag. http://www.dfg.de/download/pdf/dfg_im_profil/reden_stellungnahmen/download/empfehlung_wiss_praxis_1310.pdf

[9] **UK Data Service, 2013.** Data management costing tool. UK Data Archive, University of Essex. http://www.data-archive.ac.uk/media/247429/costingtool.pdf bzw. http://www.data-archive.ac.uk/create-manage/planning-for-sharing/costing

[10] DFG, o.J., **Leitfaden für die Antragstellung. Projektanträge** (DFG-Vordruck 54.01 – 06/14), Bonn.
http://www.dfg.de/formulare/54_01/54_01_de.pdf

[11] DFG, o.J., **Nachnutzung von Forschungsdaten – Anregungen und Best-Practice Beispiele.** http://www.dfg.de/foerderung/antragstellung_begutachtung_entscheidung/antragstellende/antragstellung/nachnutzung_forschungsdaten/

Ausgewählte Internetquellen zu Datenportalen und Metadaten, Stand: September 2014

DATENPORTALE	
re3data, Portal zu Datenportalen	http://www.re3data.org/
Kompetenzzentrum Forschungsdaten	http://www.komfor.net
World Data System, Trusted Data Services for Global Science	http://www.icsu-wds.org/
British Atmospheric Data Centre (BADC)	http://badc.nerc.ac.uk/
Australian National Data Service	http://www.ands.org.au/
GEOFON und EIDA Datenarchiv	http://eida.gfz-potsdam.de
INSPIRE Datenportal	http://inspire-geoportal.ec.europa.eu/
GEOSS Datenportal	http://www.geoportal.org
WDCC (World Data Center for Climate)	http://www.dkrz.de/daten/wdcc/
Geoportal, Geodaten aus Deutschland	http://www.geoportal.de/
CERA (Climate and Environmental Data Retrieval and Archiving System)	http://cera-www.dkrz.de/
Earth System Grid	http://esgf.org/
ENES (European Network for Earth System Modelling), dort weiter unter Data Infrastrukture / Data Portals	https://verc.enes.org/
KNMI Climate Explorer	http://climexp.knmi.nl/
NASA Global Change Master Directory	http://gcmd.nasa.gov/
NCAR Community Data Portal	http://cdp.ucar.edu/
NOAA National Climatic Data Center	http://www.ncdc.noaa.gov/

PORTALE ZUR DATENPUBLIKATION	
Datenpublikation am DKRZ	http://www.dkrz.de/daten-en/Datapublication/
PANGAEA Data Publisher for Earth & Environmental Science	http://www.pangaea.de/submit/
Geodatenkatalog (im Aufbau begriffen)	https://wiki.gdi-de.org/display/gdk/Geodatenkatalog-DE/
Data Cite	http://www.datacite.org/

METADATENSCHEMATA	
NetCDF CF (Climate and Forecast) Conventions and Metadata	http://cfconventions.org/
GDI-DE Registry	http://repository.gdi-de.org/schemas/registry/
Marine Metadata Interoperability – Common Metadata Standards	https://marinemetadata.org/guides/mdatastandards/comstds/
Dublin Core Metadata Terms (2012)	http://dublincore.org/documents/dcmi-terms/
GEOSS Registry	http://geossregistries.info/
Digital preservation formats	http://www.digitalpreservation.gov/formats/

Ausgewählte Internetquellen zu Werkzeugen

Datenerfassung und Kollaboration – File-Sharing	
Dropbox	http://www.dropbox.com/
Powerfolder	https://www.powerfolder.com/
DFN-Cloud	https://www.dfn.de/dfn-cloud/
Owncloud	https://owncloud.org/
Seafile	http://seafile.com/
Google Drive	https://drive.google.com/
Teamdrive	http://www.teamdrive.com/

Datenerfassung und Kollaboration – Texte und Quellcode	
GITHUB	https://github.com/
Sourceforge	http://www.sf.net/
Mercurical	http://mercurial.selenic.com/
Datenerfassung und Kollaboration – Dokumentbearbeitung	
Etherpad	http://etherpad.org/
Titanpad	http://titanpad.com/
Google Docs	https://docs.google.com/
Wiki-Implementierungen	https://www.mediawiki.org/

Datenerfassung und Kollaboration – Projektmanagement	
Jira, Confluence	http://www.atlassian.com/
Trac	http://trac.edgewall.org/
Passwortgeneratoren (Pwgen)	http://pwgen-win.sourceforge.net/
	http://wiki.ubuntuusers.de/pwgen/

Sicherung der Datenqualität	
Karma Provenance Collection Tool	http://d2i.indiana.edu/provenance_karma/

Metadatenerstellung	
GEMET Thesaurus	http://www.eionet.europa.eu/gemet/en/themes/
GCMD Science Keywords der NASA	http://gcmd.nasa.gov/learn/keyword_list.html

Datensicherung und Reproduzierbarkeit	
Ntbackup	http://en.wikipedia.org/wiki/NTBackup
Rsync	http://de.wikipedia.org/wiki/Rsync
Tar	http://de.wikipedia.org/wiki/Tar
Revision Control System (RCS)	https://www.gnu.org/software/rcs/
Concurrent Versions System (CVS)	http://www.nongnu.org/cvs/
Subversion (SVN)	https://subversion.apache.org/
GIT	http://git-scm.com/
Mercurial	http://mercurial.selenic.com/

Kurzes Glossar

Begriff	Beschreibung
Checksumme (auch: Prüfsumme)	Eine Checksumme dient der Überprüfung der Integrität eines digitalen Objekts, zum Beispiel bei dessen Übertragung oder Sicherung. Mithilfe eines Algorithmus wird aus den Bits des Objekts eine Prüfsumme errechnet. Ändert sich nur ein einziges Bit, ändert sich auch diese Prüfsumme. So lässt sich überprüfen, ob ein digitales Objekt fehlerfrei und unverändert übermittelt oder gesichert wurde: sind die Prüfsummen vor und nach der Übermittlung identisch, sind es auch die beiden Objekte.
Datenmanagementplan	Dokument zur Beschreibung des Lebenszyklus von Daten von der Erhebung bis zur Archivierung, einschließlich aller Maßnahmen, die gewährleisten, dass die Daten verfügbar, nutzbar und nachvollziehbar (verständlich) bleiben. In Deutschland gibt es noch kein vorgeschriebenes standardisiertes Verfahren zur Erstellung eines Datenmanagementplans.
Datenpublikation	Zitierbare Dokumentation eines Datensatzes, über die der Datensatz allgemein zugänglich wird. Idealerweise hat eine Datenpublikation eine eindeutige Adresse im Internet (s. Persistent Identifier).
Datensicherung	Temporäre Duplizierung von Daten zur Vermeidung von Datenverlust aufgrund von Störungen.
Forschungsdaten	(Digitale) Daten, die je nach Fachkontext Gegenstand eines Forschungsprozesses sind, während eines Forschungsprozesses entstehen oder ihr Ergebnis sind und die von der wissenschaftlichen Community als notwendig für die Verifizierung von Forschungsergebnissen erachtet werden. Forschungsdaten werden unter Anwendung verschiedener Methoden – abhängig von der Forschungsfrage – erzeugt, z.B. durch Quellenforschungen, Experimente, Messungen, Beschreibungen, Erhebungen oder Befragungen.
Forschungsdatenmanagement	Ein Prozess, der alle Methoden und Verfahren umfasst, die zur Sicherung der langfristigen Nutzbarkeit von Forschungsdaten angewendet werden: die Generierung, die Bearbeitung, die Anreicherung, die Archivierung und die Veröffentlichung. Im Ergebnis entstehen selbstbeschreibende Forschungsdaten. Zu Projektbeginn empfiehlt es sich, die Methoden und Verfahren in einem Datenmanagementplan zu beschreiben.
Langzeitarchivierung	Verfahren, das Daten für einen unbestimmten Zeitraum, der über nicht vorhersehbare technologische und soziokulturelle Veränderungen hinausreicht, verfügbar und für Menschen interpretierbar hält. Im Unterschied zum Backup (Datensicherung), das lediglich die jetzige Arbeitsumgebung dupliziert, müssen Daten jederzeit decodierbar und lesbar erhalten werden, z.B. auch über Dateiformatänderungen hinweg.
Metadaten	Alle zusätzlichen Informationen, die zur Interpretation der eigentlichen Forschungsdaten notwendig oder sinnvoll sind und die eine (automatische) Verarbeitung der Forschungsdaten durch technische Systeme ermöglichen. Damit bspw. Messdaten interpretierbar und damit nachnutzbar sind, ist die vollständige und korrekte Angabe der jeweils verwendeten (SI) Einheit unabdingbar. Sinnvoll sind auch beschreibende Erläuterungen (etwa in Form eines Abstracts), die zusammen mit den Forschungsdaten aufbewahrt werden. Dazu zählen auch Hinweise auf Nutzungsrechte, eingesetztes Equipment, verwendete Standards, wenn keine dazugehörige Publikation vorhanden ist.

Begriff	Beschreibung
Persistent Identifier (PID)	Eindeutige Benennung einer digitalen Ressource (z.B. Zeitschriftenartikel oder Forschungsdaten) durch Vergabe eines Codes, der im Internet dauerhaft eindeutig referenziert werden kann. Dadurch wird verhindert, dass tote Links entstehen, wenn beispielsweise Verlage die Internetadresse eines Servers ändern. In den Naturwissenschaften ist die DOI (Digital Object Identifier) als ein Typ eines PID am weitesten verbreitet.
Qualitätsmanagement	Aktivitäten zur Identifikation und Beseitigung von Datenfehlern oder Datenlücken. Dazu gehört die Prüfung der Validität von Dateiformaten, soweit das möglich ist, sowie die Sicherung von Authentizität und Integrität bei Übertragung von Daten an ein Forschungsdatenzentrum oder Langzeitarchiv. Inhaltlich kann nur in Ausnahmefällen geprüft werden, ob Daten korrekt sind, weil dieser Vorgang sehr aufwendig ist und fachliche Kompetenz erfordert. PANGAEA bspw. prüft mit Hilfe von angestellten Daten-Kuratoren, ob übermittelte Daten plausibel sind (Einheiten-Check etc.).
Rechte	Aus Sicht der Forschenden Entscheidungsbefugnisse über die Daten, die sich aus der Erzeugung ergeben. Aus Sicht von Nutzerinnen und Nutzern die Rechte, welche bei Nachnutzung von Daten zu beachten sind. Für Letzteres gelten mindestens die Regeln der guten wissenschaftlichen Praxis, d.h. im Wesentlichen die Pflicht, Urheber korrekt zu zitieren. Mit Vergabe der Creative-Commons-Lizenz CC-BY lässt sich diese Regel durch den Datenerzeuger auch lizenzrechtlich weitgehend nachbilden. Datenschutz-, patent- und persönlichkeitsrechtliche Einschränkungen können die Nachnutzung erschweren. Rechte können in Form von Lizenzen und zugehörigen Lizenztexten sowie Vereinbarungen in rechtlich verbindlicher Form festgelegt und kommuniziert werden.
Richtlinien, Regeln, Policies	Leitlinien, die für alle Mitarbeiter einer Institution (z.B. Hochschulinstitut) festschreiben, welche Verfahren beim Forschungsdatenmanagement eingesetzt werden sollen. In Deutschland gibt es fast keine Policies mit detaillierten Vorgaben sondern zumeist nur grundlegende Eigenverpflichtungen etwa zu den Prinzipien des Open Access.
Repository (Repositorium)	(Online zugängliche) Datenbank zur Verzeichnung und Publikation von Forschungsdaten, Hochschulschriften und anderen digitalen Objekten. Zumeist der wesentliche Service eines Forschungsdatenzentrums.
Versionierung	Buchführung über die Bearbeitungsschritte von Daten auf der Ebene der persönlichen und Gruppen-Domäne. Ein Versionierungssystem unterstützt dies effektiv, transparent und standardisiert.

Nachwort

Diese Broschüre entstand im Sommer 2014 auf dem Workshop „Wege in die Köpfe“ des DFG geförderten Projekts EWIG (Entwicklung von Workflowkomponenten für die Langzeitarchivierung von Forschungsdaten in den Geowissenschaften). Während der Laufzeit des Projekts wurden Workflows von Forschungsdaten in den Geowissenschaften im Hinblick auf die Sicherung der Langzeitverfügbarkeit untersucht. Zur Identifikation möglicher wiederkehrend auftretender Probleme wurden über 20 Gespräche mit verschiedenen Experten aus allen Bereichen des wissenschaftlichen Forschungsprozesses geführt. Nachdem zunächst an den Übergang in die Langfristige Domäne gedacht wurde, stellte sich schnell heraus, dass das ein viel zu später Zeitpunkt im Wissenschaftsbetrieb ist. Wir haben uns daher entschlossen, den Prozess der Langzeitverfügbarkeit ganz nach vorne zu verlagern: in die Ausbildung von wissenschaftlichem Nachwuchs selbst. Nur wenn hier eine Kultur der Nachnutzung und des freien Teilens von Wissen entsteht, werden auch Maßnahmen ergriffen werden, die eigenen Forschungsdaten für die Langzeitverfügbarkeit fit zu machen.

Zu diesem Zweck haben wir diese Handreichung für die Ausbildung und das erste Herangehen von Forschenden an das Thema Datenmanagement entwickelt. Um die Empfehlungen auf eine möglichst breite Grundlage zu stellen, haben wir über die Grenzen der Fachdisziplin hinweg Vertreter aus Infrastruktur, Forschungsbibliotheken und der Wissenschaft selbst eingeladen, um entlang eines vorgegebenen Rahmens aus Anreizen, Definitionen, persönlicher Organisation und dem Aspekt der Kosten, Inhalte für Module im geowiossenschaftlichen Curriculum zu entwickeln. Unser Dank geht an die Teilnehmer und Teilnehmerinnen: Thomas Bergmann, Roland Bertelmann, Christoph Bruch, Kirsten Elger, Petra Gebauer, Tim Hasler, Maurice Heinrich, Mirjam Hirt, Christopher Kadow, Ingo Kirchner, Monika Kuberek, Nora Mettig, Heinz Pampel, Wolfgang Peters-Kottig, Thorsten Rathmann, Matthias Razum, Astrid Recker, Ulrike Schenk, Katharina Schütze, Elena Simukovic, Juliane Steckel, Hannes Thiemann, Damian Ulbricht, Stephan van Gasselt, Johanna Vompras, Sebastian Walter und Martin Wattenbach.

Teilnehmer des EWIG Workshops „Wege in die Köpfe“ am 3. und 4. Juli 2014

Impressum

Herausgeber: Helmholtz-Zentrum Potsdam – Deutsches GeoForschungsZentrum GFZ, Bibliothek und Informationsdienste, Institut für Meteorologie der Freien Universität Berlin, Konrad-Zuse-Zentrum für Informationstechnik Berlin

Autoren: Roland Bertelmann, Petra Gebauer, Tim Hasler, Ingo Kirchner, Wolfgang Peters-Kottig, Matthias Razum, Astrid Recker, Damian Ulbricht, Stephan van Gasselt

Gefördert durch die Deutsche Forschungsgemeinschaft

DFG Deutsche Forschungsgemeinschaft

Version September 2014

Gestaltung: Beate Autering / beworx Berlin

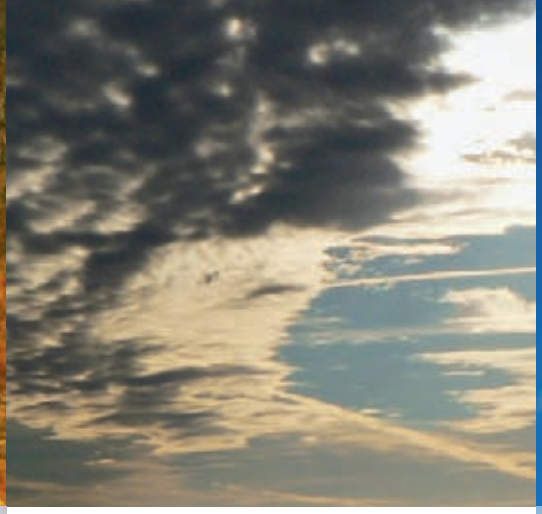
Bildnachweise: Titelseite (v.o.l.n.u.r.): CC0 Public Domain/pixabay.com, Elmar Söllner, Dan Mirica/Fotolia.com, CC0 Public Domain/pixabay.com (2), Beate Autering, creative commons (CC-PD-Mark/PD Old), pict rider/Fotolia.com, CC0 Public Domain/pixabay.com (3), Alma/creative commons (CC-BY-SA-2.5)suze/photocase.com; Rückseite (v.o.l.n.u.r.): HSuepfle/creative commons (CC-PD-Mark/PD US Government), Lachsy/pixelio.de, rangizzz/Fotolia.com, inkje/photocase.com, oversnap/istockphoto.com, berean/Fotolia.com, Zacarias da Mata/Fotolia.com, Rapture/Fotolia.com, Klaus-G. Hinzen (et al.), Universität Köln/creativ commons (CC-BY-3.0), Wolfgang Beyer/creativ commons (CC-BY-SA-3.0), Oleg Kozlov/Fotolia.com; Illustration: godruma/Fotolia.com (Umschlag), Beate Autering (S.2, S.3)

Dauerhaft zitierbar über Digital Object Identifier (DOI):
DOI: 10.2312/lis.14.01

ZIB

Freie Universität Berlin

GFZ Helmholtz-Zentrum Potsdam



Nur wenn eine Kultur der Nachnutzung und des freien Teilens von Wissen entsteht, werden Wissenschaftler auch die eigenen Forschungsdaten für die Langzeitverfügbarkeit aufbereiten und freigeben. Zu diesem Zweck ist diese Handreichung als Einstieg in das Thema Forschungsdaten-management entstanden.