Bericht über initialen Kernstandard und Kurationskriterien

Version 2.4

Anette Ganske, Angelina Kraft,
Technische Informationsbibliothek (TIB)

Daniel Neumann, Andrea Lammert, Heinke Höck, Hannes Thiemann,
Deutsches Klimarechenzentrum (DKRZ)

Vivien Voss, David Grawe, Bernd Leitl, K. Heinke Schlünzen,
Universität Hamburg, Meteorologisches Institut

Johannes Quaas
Universität Leipzig, Institut für Meteorologie

26. März 2020

Inhaltsverzeichnis

1 Einleitung

Hochwertige Ergebnisse von Simulationen mit Atmosphärischen Modellen sollen so veröffentlicht werden, dass sie von anderen Wissenschaftlern für weitere Forschungen genutzt werden können. Um dies zu erreichen, müssen die Daten auffindbar, in nutzbarer Form abgespeichert und mit ausreichenden Metadaten beschrieben sein. Handlungsanweisungen um diese Bedingungen zu erfüllen bilden die Regeln der FAIR-Prinzipien: **F**indable, **A**ccessible, **I**nteroperable, **R**euseable [Wilkinson et al., 2016]. Daten, die in einem Repositorium archiviert und mit einem PID (Persistent Identifier) versehen sind, erfüllen schon die *F* und *A* Prinzipien aber nicht unbedingt die *I* und *R* Prinzipien.

Ein Ziel des Projektes AtMoDat[1] (**At**mospheric **Mo**del **Dat**a) ist es, einen Kernstandard und Kurationskriterien für Atmosphärenmodelldaten auf Basis existierender Standards und bisheriger Erfahrungen bei der Kuratierung von Daten im World Data Center for Climate (WDCC) zu erstellen. Die Kombination aus Kernstandard und Kurationskriterien sollen die Interoperabilität und Nachnutzbarkeit (Reusability) von Atmosphärenmodelldaten deutlich verbessern. Der Kernstandard kann von Teildisziplinen der Atmosphärenforschung für ihre jeweiligen Belange verfeinert werden. Da viele Atmosphärendaten mit einem DataCite DOI (Digital Object Identifier) [DataCite Metadata Working Group, 2019] archiviert werden, beschränken wir uns im folgenden Bericht auf diesen PID. Die Ergebnisse des Berichts können jedoch leicht auf andere PIDs übertragen werden.

Im Laufe des AtMoDat Projekts, das Mai 2022 endet, wird geprüft werden, ob der Kernstandard sinnvoll und für alle atmosphärischen Modelldaten ausreichend ist. Diese Erfahrungen werden dann in die neueren Versionen des Berichts einfliessen. Zudem ist eine Erweiterung des Kernstandards auf Windkanaldaten geplant. Neben der Festlegung des Kernstandards wird zusätzlich ein AtMoDat-DOI entwickelt, der Daten auszeichnet, die unter diesem Standard veröffentlicht werden. Dieser AtMoDat-DOI ist als eine Art Qualitätsindikator anzusehen, der *einfach* nachnutzbare Daten kennzeichnet.

Im folgenden Bericht werden zuerst die FAIR Prinzipien beschrieben (Kapitel 2). Anschliessend wird der Begriff Granularität eingeführt (Kapitel 3). In den folgenden Kapiteln 4 und 5 wird erläutert, wie man die DataCite Metadaten ausfüllen und die Landing Page gestalten muss, um die Bedingungen für einen AtMoDat-DOI zu erfüllen. Zudem müssen für die FAIRness der Daten bestimmte Kriterien für die Datenformate eingehalten werden (Kapitel 6). Anschliessend wird erläutert, wie man die so erzielte Qualität sichtbar

[1] https://www.atmodat.de

machen kann (Kapitel 7). Der Bericht endet mit einer Zusammenfassung.

2 Die FAIR Prinzipien

Um den FAIR-Prinzipien zu folgen, müssen Forschungs- und Metadaten die Eigenschaften Findable, Accessible, Interoperable und Reusable haben, siehe Wilkinson et al. [2016]. Für die Atmosphärischen Forschungsdaten bedeutet dies:

Findable: die Forschungsdaten werden mit Metadaten ausführlich beschrieben und ihnen wird eine eindeutige und persistente Kennung (PID) zugeordnet - für Atmosphärische Forschungsdaten wird immer ein DOI empfohlen. Die Metadaten enthalten den DOI und werden so abgespeichert, dass sie von Suchmaschinen gefunden werden können.
Zudem werden – soweit vorhanden – für die Metadaten standardisierte Begriffe verwendet, so dass Forschungsdaten gleichen Typs leichter gefunden werden können und automatisierte Verknüpfungen zwischen verschiedenen Datensätzen möglich sind.

Accessible: Sowohl auf die Forschungsdaten als auch auf die Metadaten kann mit einem standardisierten, maschinenlesbaren Protokoll zugegriffen werden. Ein solches Protokoll ist offen, frei und universell implementierbar. Die Metadaten werden in von Suchmaschinen lesbaren Datenformaten geschrieben. Um viele verschiedene Suchmaschinen bedienen zu können, die jeweils Metadaten in unterschiedlichen Datenformaten auslesen, werden die Metadaten im maschinenlesbaren Teil der Landing Page in mehreren Formaten bereitgestellt. Zusätzlich sind die Metadaten aller DataCite DOIs bei DataCite gespeichert und können über eine öffentlich verfügbare Schnittstelle (API = Automated Programming Interface) von Suchmaschinen durchsucht werden.
DataCite erlaubt den Repositorien die Setzung verschiedener Zugriffsrechte für die Forschungsdaten sowie die Möglichkeit eines Genehmigungsverfahrens für den Datenzugriff. Zum Beispiel können bestimmte Daten nur nach Anfrage für bestimmte Nutzer, wie z.B. Forschende aus einer bestimmten Disziplin, freigeschaltet werden.
Unabhängig von den Nutzugsrechten muss auf die Metadaten immer und auch dann noch zugegriffen werden können, wenn die Daten nicht (mehr) zugänglich sind. Dies ist auch eine Bedingung von DataCite für die Vergabe eines DOI.

Interoperable: Meta- und Forschungsdaten liegen in einem Datenformat vor, das von anderen mit möglichst frei verfügbarer Software gelesen

und verarbeitet werden kann. Für Metadaten werden diese Kriterien z.B. bei den Datenformaten von DataCite[2], DCAT[3] und schema.org[4] eingehalten. Beispielsweise erfüllt das netCDF-Format[5], das oft für atmosphärische Modell-Daten verwendet wird, ebenfalls diese Bedingungen.
Zudem haben Meta- und Forschungsdaten qualifizierte Referenzen auf andere (Meta)daten. Dies kann z.B. für den AtMoDat-DOI erfolgen, indem im Metadatensatz eine Verlinkung auf die Randdaten enthalten ist, die bei der Berechnung der Forschungsdaten verwendet wurden. Falls es eine Publikation gibt, die die Datenerstellung erläutert, so wird darauf verwiesen. Zusätzlich werden, falls möglich, die Datensätze in einen wissenschaftlichen Kontext gesetzt. So soll z.B. bei Datensätzen, die im Rahmen von einem Model Intercomparison Project (MIP, siehe z.B. das Climate Model Intercomparison Project CMIP6 Eyring et al. [2016]) erstellt wurden, immer ein persistenter Link zu anderen Datensätzen aus dem MIP in den Metadaten enthalten sein. Zusätzlich wird auf die Homepage des MIPs verwiesen werden, falls diese persistent ist und nicht z.B. nach Ende des Projekts abgeschaltet wird. Ebenso werden Links zu verwandten Datensätzen, die sich mit dem gleichen Forschungsproblem beschäftigen, in die Metadaten aufgenommen.

Reusable: Meta- und Forschungsdaten werden ausführlich, möglichst akkurat und mit relevanten Attributen beschrieben. Sie werden mit eindeutigen und frei zugänglichen Rechten veröffentlicht.
Die Herkunft der Daten ist klar beschrieben und sowohl die Forschungs- als auch die Metadaten entsprechen den Standards, die in der Forschungsgemeinschaft üblich sind. Eine Übertragung von Standards, die für die CMIP-Klimamodelldaten gelten (siehe z.B. Stockhause et al. [2012]), auf andere Modelldaten, wie z.B. die Daten von Stadtklimamodellen, wird für den AtMoDat-DOI angestrebt.

3 Granularität von Atmosphären-Daten

Der erste Unterpunkt von Findable in den FAIR-Prinzipien beinhaltet, dass den Daten ein PID zugewiesen wird, siehe Wilkinson et al. [2016]. Ergebnis-

[2]https://doi.org/10.5438/BMJT-BX77
[3]https://www.w3.org/TR/2019/WD-vocab-dcat-2-20190528/
[4]https://schema.org/Dataset
[5]https://doi.org/10.5065/D6H70CW6

se einer atmosphärischen Simulation werden jedoch meist in größeren Datensätzen mit mehreren Dateien abgespeichert. Hierbei stellt sich die Frage, wem der PID genau zugewiesen wird - jeder einzelnen Datei oder dem gesamten Datensatz? Dies entscheidet darüber, ob die Metadaten des PIDs entweder den gesamten Datensatz oder eine einzelne Datei beschreiben.

Diese unterschiedlichen Datenebenen bezeichnet man auch als Granularität (= Verdichtungsgrad). Ergebnisse von Atmosphären-Modellen liegen in unterschiedlicher Granularität vor. Da die Atmosphäre ein komplexes System mit sehr vielen Freiheitsgraden ist, enthalten die Modellergebnisse räumlich ein-, zwei- und dreidimensionale Variable, die sowohl zeitabhängig als auch konstant sein können. Zudem können die Variablen in mehreren räumlichen Aggregationen wie z.B. als Punktwerte oder Flächenmittel, abgespeichert werden. Ebenso können die Variablen in mehreren zeitlichen Aggregationen vorliegen, z.B. als Instantan-, Stundenmittel-, Monatsmittel- oder Jahresmittelwerte. Die Entscheidung, welche der Variablen in welcher räumlichen und zeitlichen Aggregation für die Nachnutzung abgespeichert werden, liegt entweder beim Erzeuger der Daten oder wird bei einem Modellvergleichs-Projekt wie CMIP von der Projektleitung festgelegt.

Die Ergebnisse werden vom Modell meist in mehrere große Dateien herausgeschrieben. So kann eine Datei z.B. alle zeitunabhängigen Felder enthalten, eine andere enthält alle zeitabhängigen Felder mit jeweils allen 3 Raumdimensionen. Diese können nach der Modellsimulation mit einem Post-Processing in Dateien umgespeichert werden, die beispielsweise jeweils nur Zeitreihen einer Variablen enthalten. Für Modellergebnisse, die im Rahmen von CMIP produziert wurden, ist z.B. festgelegt, dass immer nur eine Variable in einer netCDF-Datei steht. Andere Forschergruppen oder MIPs schreiben mehrere Variablen in eine gemeinsame netCDF-Datei.

In jedem Fall kann man zwischen einer grob- und einer feingranularen Ebene der Daten aus einer Simulation unterscheiden:

1. **Ein Daten-Paket mit mehreren Datensätzen:**
   Die *grobgranulare Ebene* bildet eine ganze Simulation bzw. ein Daten-Paket. Sie kann mehrere Dateien mit Ergebnissen in unterschiedlichen Dimensionen und zeitlichen Auflösungen enthalten.
   Die *feingranulare Ebene* bilden die einzelnen Dateien im Paket, die wiederum mehrere Variablen enthalten können.

2. **Eine Datei, die mehrere Variablen enthält:**
   Die *grobgranulare Ebene* bildet die Datei.
   Die *feingranulare Ebene* bilden die einzelnen Variablen in der Datei.

Von DataCite ist nicht festgelegt, für welche Daten ein DOI vergeben wird. Deshalb werden DOIs sowohl für einzelne Dateien mit einer einzigen Variablen als auch für ganze Daten-Pakete mit mehreren Dateien vergeben. Da die Kuratierung von Daten mit DOI aufwendig ist und die Vergabe eines DOIs in Zukunft kostenpflichtig wird, bekommt oft die grobgranulare Ebene der Datenstruktur – meistens alle Ergebnisse einer ganzen Simulation – einen einzigen DOI. Beispielsweise wurde beim WDCC bei den CMIP5-Daten je ein DOI für je eine sogenannte Dataset-group vergeben, die mehrere sogenannte Datasets (Datensätze) enthält (siehe Abbildung 1). Ein Dataset besteht jeweils aus Ergebnissen für eine einzige Variable, z. B. einer Zeitreihe einer einzigen Variablen. Sind es Daten von einer Simulation, die nicht in einem MIP gemacht wurde, kann der DOI ebenfalls für die gesamte Simulation (beim WDCC Experiment genannt) vergeben werden. Diese kann eine oder mehrere Datasets enthalten, in denen wiederum mehrere Variablen abgespeichert sein können. Andere Datenzentren vergeben auch für einzelne Dateien einer Simulation DOIs, so dass die Ergebnisse dieser Simulation auf mehrere Dateien mit jeweils einem DOI verteilt sein können (Parent- and Child-DOI). Dabei wird in den Metadaten der Child-DOIs auf den Parent-DOI verwiesen und umgekehrt.

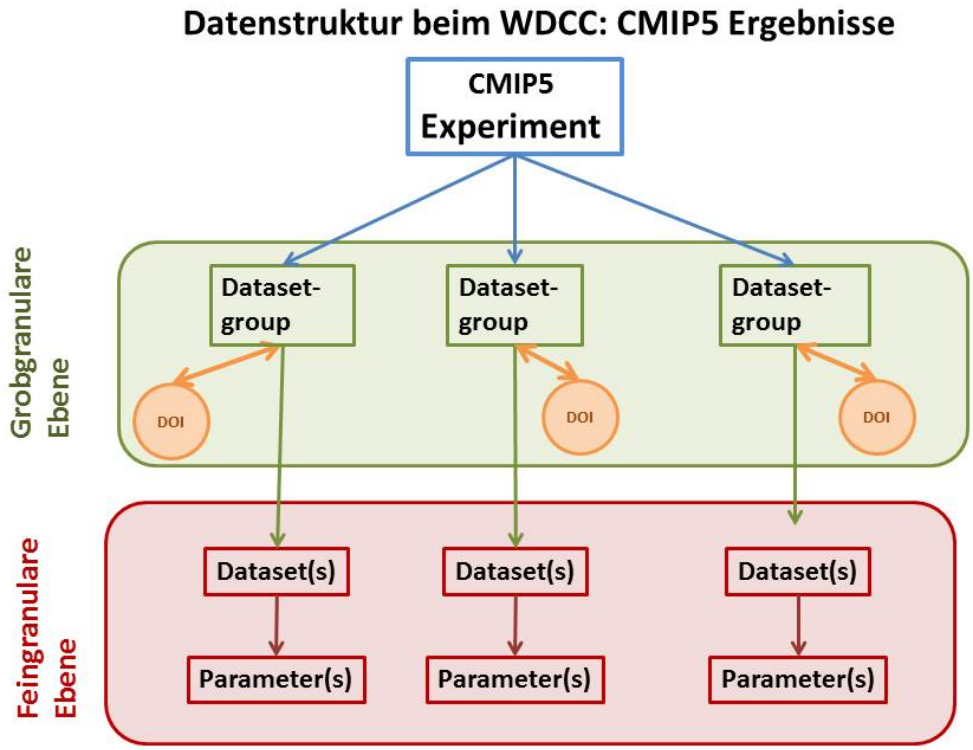

Abbildung 1: Dateistruktur bei CMIP5-Daten in der WDCC Datenbank und ihre Granularitätsebenen. Dabei repräsentieren die blauen, grünen und roten Pfeile die Verweise, die z.B. durch Links auf den Webseiten von der Beschreibung des CMIP5 Experiments zu den einzelnen Simulationen (Dataset-groups), von den Beschreibungen der Simulationen zu denen der Datensätzen (Dataset) und von den Beschreibungen der Datensätze zu denen der einzelnen Parameter führen. Die orangenen Pfeile weisen von der Datset-group auf ihren DOI, dessen Metadaten die jeweilige Dataset-group beschreiben.

4 Die Metadaten

Metadaten sind ein wichtiger Bestandteil des FAIR-Prinzips. So fordert z.B. der Unterpunkt F2 in der Definition von FAIR bei Wilkinson et al. [2016], dass die Daten mit reichhaltigen Metadaten beschrieben werden sollen.

Um FAIRness zu erreichen, sollte in den Metadaten einerseits der Ursprung der Daten dokumentiert werden: die Institution und die Namen der Personen, die die Forschungsdaten berechneten, und der Prozess der Erzeugung der Daten. Andererseits sollen die möglichen Nachnutzer mithilfe der Metadaten entscheiden können, ob diese Daten für sie wirklich nützlich sind. Dabei kann die Beschreibung der Daten und der Modelle, mit denen sie berechnet wurden, dadurch strukturiert werden, dass die Metadaten Verweise sowohl auf Modelldokumentationen als auch auf die Dokumentationen der Qualitätskontrolle und der Randbedingungen enthalten. Diejenigen Metadaten , die erst bei der Weiterverarbeitung der Daten gebraucht werden, wie z.B. die Beschreibung eines Gitters oder die Einheiten der jeweiligen Parameter, stehen nicht unbedingt in den DataCite Metadatenfeldern. Statt dessen stehen sie direkt bei den Daten, z.B. im Header einer netCDF-Datei.

Metadaten werden jedoch nicht nur von Menschen gelesen. Zum einen dienen die Metadaten als Stichwortliste, die von Suchmaschinen ausgewertet wird. Zum anderen können sie verwendet werden, um automatisierte Listen zu erstellen: So können für Evaluierungen von Institutionen Übersichten über deren Veröffentlichungen erzeugt werden indem die Metadatenfelder für Institutionen und Autoren ausgewertet werden. Ebenso kann beispielsweise mit der Angabe des Förderes abgefragt werden, wie viele geförderte Projekte eine Institution hatte und wie viele Ergebnisse erzielt wurden. Zudem können die Metadaten verschiedener Datensätze miteinander verknüpft werden, um verwandte Datensätze aufzuzeigen und eine Datensatz übergreifende Auswertung zu ermöglichen. Für alle diese Anwendungsszenarien ist es wichtig, dass möglichst umfassende Metadaten menschen- und maschinenlesbar vorhanden sind.

4.1 AtMoDat Metadaten der grobgranularen Ebene

Die Metadaten der grobgranularen Ebene sind diejenigen Metadaten, die dem DOI mitgegeben werden. Diese Metadaten werden bei DataCite abgespeichert und können von Suchmaschinen abgefragt werden. Sie stehen auf der obersten Ebene der Landing Page, auf die über den DOI hin verlinkt wird.

Dabei gibt es bei DataCite verpflichtende, empfohlene und optionale Metadaten. Für den AtMoDat-DOI sollen **alle** Metadatenfelder ausgefüllt wer-

den, soweit möglich und sinnvoll:

Identifier: der DOI selbst – entspricht einer Forderung aus dem Findable von FAIR

Creator: alle Erzeuger der Simulation oder die Institution, bei der die Daten erzeugt wurden– Name und PID für Personen, wie z.B. ORCID oder Scopus ID, bei Institutionen PID wie z.B. ROR[6] oder GRID[7]. Dieses Feld wird auch beim automatischen Erstellen von Evaluierungsberichten verwendet.

Title: Name des Datensatzes/der Simulation

Publisher: Wo wurden die Daten publiziert – in der Regel ein Datenzentrum. Falls vorhanden, PID des Datenzentrums angeben, z.B. ROR oder GRID.

Publication Year: Wann wurden die Daten publiziert

Resource Type: bei AtMoDat immer anzugeben *resourceTypeGeneral*=Dataset. Bei Daten, die auf einem Gitter vorliegen: *ResourceType*=grid – erscheint dann in den Metadaten als Dataset-grid.

Subject: Mehrere Schlagworte sind anzugeben. Sie sollen, soweit vorhanden, aus einem kontrollierten Vokabular stammen und bei MIPs z.T. vorgegeben werden. Das erste Subject ist immer "*AtMoDat*". Darauf folgt eine Forschungsdisziplin, entweder aus der offiziellen DFG-Liste[8] oder aus der Wikipedia-Liste der Forschungsdiziplinen[9]. Anschließend wird das sogenannte *Realm* des Modellsystem angegeben, was in etwa dem Kompartiment entspricht. Der Wert bzw. die Werte für Realm sind dem entsprechenden kontrollierten Vokabular von CMIP6 zu entnehmen[10], das im CMIP6 GitHub Repository[11] oder im ES-DOC[12] zu finden ist. Weitere Subjects können frei gewählt werden, wie z.B. Name des MIP, in dem die Simulation gemacht wurde. Dabei sind möglichst nicht Begriffe aus dem Titel zu verwenden, sondern es sollen zusätzliche Keywords oder übergeordnete Begriffe angegeben werden.

[6] https://ror.org/

[7] https://grid.ac/

[8] http://gepris.dfg.de

[9] `https://en.wikipedia.org/wiki/List_of_academic_fields`

[10] Aktuell möglich sind aerosol, atmos, atmosChem, land, landIce, ocean, ocnBgchem und seaIce.

[11] `https://github.com/WCRP-CMIP/CMIP6_CVs/blob/master/CMIP6_realm.json`

[12] https://es-doc.org/

Contributor: alle Personen oder Institutionen, die am Entstehungsprozess beteiligt waren, wie z.B. Ansprechpartner oder Projektleiter. Dabei müssen die PIDs der einzelnen Personen angegeben werden – falls vorhanden.

Date: *Created:* wann wurden die Daten erzeugt
Updated: falls es eine neue Version gibt, Zeitpunkt angeben
Issued: Datum, an dem die Daten veröffentlicht wurden.
Available: falls Daten nicht sofort zugänglich gemacht werden, dann hier angeben, ab wann auf die Daten zugegriffen werden kann (max. 2 Jahre nach Veröffentlichung).
Valid: falls eine Zeitreihe vorliegt, dann Anfang und Ende angeben – immer DIN ISO 8601 (ISO Central Secretary [2004]) und ISO 19108 (ISO Central Secretary [2002] berücksichtigen. Beispielsweise schreibt man den Zeitraum vom 1.6.2019-31.8.2019 als 20190601/20190831. Falls nur die Länge einer Zeitreihe bekannt ist, wird eine Dauer von mehreren Jahren mit PxxX angegeben. So entspricht z.B. P1Y der Dauer von einem Jahr. Falls ein Datum für den Anfang der Zeitreihe existiert, so entspricht z.B. 19710101/P10Y einer 10 Jahre langen Zeitreihe, die am 1.1.1971 beginnt. Vor Angaben zu Zeitdauern kürzer als 1 Tag ein T stellen. So wird eine Dauer von fünf Tagen, vier Stunden, 30 Minuten und 10 Sekunden dargestellt als P5DT4h30M10S.

Language: Sprache der Metadaten (in der Regel bei AtMoDat: en)

Alternate Identifier: gibt es noch einen weiteren PID?

Related Identifier: mehrere Möglichkeiten, um frühere Versionen zu zitieren und um Verbindungen zu Randdaten, Modellbeschreibung, Auswertungen der Daten oder anderen Veröffentlichungen herzustellen. Falls vorhanden, sollte immer genannt werden:
relationType=IsNewVersionOf: Link oder DOI zu vorhergehenden Versionen des Datensatzes.
relationType=IsPreviousVersionOf: Link oder DOI zu nachfolgenden Versionen des Datensatzes (Metadaten werden aktualisiert).
relationType=IsDerivedFrom: Link oder DOI zu Beschreibung der Randdaten.
relationType=IsReviewedBy: Ein Link zur Beschreibung der Dokumentation der Qualitätsprüfung des Datenzentrums sollte immer angegeben werden. Falls vorhanden, sollte auch auf die Validation der Daten (z.B. PID der entsprechenden Veröffentlichung oder des Berichts) verwiesen werden.

relationType=IsDescribedBy: entweder PID des Modells oder persistenter Link zur Beschreibung des Modells, mit dem die Daten erzeugt wurden.
relationType=IsCitedBy: PID der Veröffentlichung, für die die Daten verwendet wurden und in der der Datensatz zitiert wurde.
relationType=IsReferencedBy: PID der Veröffentlichung, in dem das Experiment oder MIP beschrieben wird, für das die Daten produziert wurden.
relationType=IsSupplementTo: Datensatz ist Supplement von einer Veröffentlichung (PID angeben).
relationType=IsPartOf: bei Daten von Simulationen, die in einem Model Intercomparison Project (MIP) gemacht wurden: Verweis auf einen PID mit der Beschreibung des MIP.
relationType=IsVariantFormOf: bei Daten von Simulationen, die in einem Model Intercomparison Project (MIP) gemacht wurden: Verweise auf PIDs mit der Beschreibung der Daten der anderen Simulationen, die zum Vergleich gemacht wurden (z.B. bei CMIP6 Verweis auf die anderen Simulationen in einem gemeinsamen MIP wie z.B. das AerChemMIP))

Size: Größe der gesamten Datensatzes, für den der DOI vergeben wird (Speicherplatz).

Format: Ausgabeformat(e) der Daten – es können verschiedene Ausgabeformate in einem Daten-Paket sein.

Version: Versionsnummer des Datensatzes, für den der DOI vergeben wird. Falls mehrere Dateien enthalten sind (Daten-Paket), können die einzelnen Dateien eigene Versionsnummern haben. Diese werden dann z.B. im Header des netCDF-files angegeben.

Rights: Rechte für die Nutzung der Daten. Für den AtMoDat-DOI sollten immer die CC-Lizenzen[13] verwendet werden, um die Rechte eindeutig festzulegen. Hierbei wird CC-BY 4.0 empfohlen. Falls ein DOI für mehrere Dateien (Daten-Paket) vergeben wird, müssen alle Dateien die gleichen Rechte besitzen! Eine Beschränkung der Rechte auf einzelne Nutzergruppen ist für den AtMoDat-DOI nicht zulässig.

Description: Beschreibung der Simulation – hier Schlagworte aus kontrolliertem Vokabular verwenden soweit möglich, z.B aus den Climate and

[13] https://creativecommons.org/licenses/by/4.0/legalcode

Forecast Conventions (CF Conventions[14]).

Geo Location: grobe Beschreibung der Region oder des Ortes (z.B. lon/lat Box und/oder Names des Orts bzw. der Region), kann auf der feingranularen Ebene genauer werden. Regionsbezeichnungen sollten aus einem kontrollierten Vokabular ausgewählt werden, z.B. aus geonames[15].

Funding Reference: Angabe der Förderer. Falls der/die Förderer in der Crossref Funder Registry[16] gelistet ist/sind, sollte zusätzlich die im Funder Registry angegebene URL verwendet werden, z.B. für die Deutsche Forschungsgemeinschaft https://doi.org/10.13039/501100001659.

4.1.1 Weitere Angaben auf der grobgranularen Ebene

Zusätzliche Angaben auf der grobgranularen Ebene, die, falls anwendbar, für den AtMODat-DOI gemacht werden müssen, aber keine eigenen DataCite Metadatenfelder haben. Damit diese Angaben trotzdem in den DataCite Metadaten vorhanden sind, werden sie im Feld 'Description' von DataCite erwähnt. Zusätzlich werden sie im maschinenlesbaren Teil der Landing Page aufgelistet.

Name des Modells, mit dem Simulation gemacht wurde: kann auch im Title oder im Subject stehen

Version des Modells, mit dem Simulation gemacht wurde: kann auch als Teil des Modellnamens genannt werden

Räumliche horizontale Auflösung der Daten: falls anwendbar, dann CMIP6 Vokabular *nominal resolution* verwenden[17]

Verwendete räumliche Projektion: falls vorhanden, möglichst Vokabular aus Proj[18] verwenden

Art des Gitters: falls vorhanden, dann Vokabular aus ES-DOC verwenden

Basic Approximations: z.B. hydrostatisch, nicht-hydrostatisch, ...

Anwendungsbereich oder Grenzen der Anwendung: Wofür wurden die Daten gerechnet und für welche Untersuchungen kann man die Daten nicht verwenden?

[14] http://cfconventions.org/
[15] https://www.geonames.org/
[16] https://www.crossref.org/services/funder-registry/
[17] `https://github.com/WCRP-CMIP/CMIP6_CVs/blob/master/CMIP6_nominal_resolution.json`
[18] https://proj.org/

4.2 Metadaten der feingranularen Ebene

Metadaten der feingranularen Ebene enthalten alle Informationen zu den einzelnen Dateien (beim Daten-Paket mit mehreren Dateien, z.B. Dataset-Group) oder zu den einzelnen Variablen. Es werden immer zusätzliche Angaben zu den einzelnen Variablen gemacht, die z.T. über die Metadaten der grobgranularen Ebene hinaus gehen und u.U. von den DataCite Metadaten abweichen können. Dies sind, soweit anwendbar, wie folgt:

Title: Name des Datensatzes/der Variablen. Bei Variablennamen an die Climate and Forecast Metadata Conventions (CF Conventions[19]) halten, soweit möglich.

zeitliche Aggregation der Daten: Bezeichnungen aus CF Conventions (Tabelle "*Cell Methods*"[20]) für instantane Werte, Mittelwerte, tägliches Maximum, ...

räumliche Aggregation der Daten: Bezeichnungen aus CF Conventions (Tabelle "*Cell Methods*"[21]) für Mittelwerte, Gitterpunktswerte, ...

räumliche Dimension der Daten: 1D, 2D oder 3D

Art der vertikalen Koordinate: Höhe, Sigma-Koordinate, Luftdruck, ...; Begriffe aus ES-DOC[22] verwenden, soweit möglich.

Anzahl der vertikalen Levels: angeben, falls Daten in mehreren vertikalen Modellschichten vorliegen.

GeoLocation: entweder gleich mit DataCite Metadatenfeld GeoLocation oder genauere Angabe, falls das Gebiet sehr klein ist

Wertebereiche (Valid Ranges): z.B. nur positive Werte möglich, da Windgeschwindigkeit oder Temperatur in Kelvin?

Size: Größe der Datei (Speicherplatz)

Darüber hinaus müssen die Variablennamen bzw. die Bezeichnungen der dahinter stehenden physikalischen oder chemischen Größen in den menschen-

[19] http://cfconventions.org/

[20] `http://cfconventions.org/Data/cf-conventions/cf-conventions-1.8/cf-conventions.html#appendix-cell-methods`

[21] `http://cfconventions.org/Data/cf-conventions/cf-conventions-1.8/cf-conventions.html#appendix-cell-methods`

[22] https://es-doc.org/

und maschinenlesbaren Metadaten enthalten sein, damit menschliche Nutzer eindeutig wissen, was sie in den Datei zu erwarten haben, und um eine Auffindbarkeit der Variablen durch Suchmaschinen zu ermöglichen.

5 Metadaten auf der Landing Page

Wird eine DOI in einem konventionellen Webbrowser aufgerufen, führt diese zu einer menschenlesbaren HTML-Internetseite, die immer auch einen maschinenlesbaren Text, den Seitenquelltext, enthält. Diese Seite wird Landing Page genannt. Sie wird vom Repositorium, das die Daten archiviert hat, bereitgestellt und enthält grundlegende Metadaten. Dies gilt für DOIs im Allgemeinen und nicht nur für DataCite DOIs[23]. Über die sogenannte HTTP Content Negotiation ist es möglich, die Landing Page in anderen Formaten wie beispielsweise schema.org JSON-LD anzufordern. Diese alternativen Formate müssen allerdings nicht zwangsläufig vom Repositorium bereit gestellt werden.

Die Vorgaben von DataCite für die Landing Page sind in *Best Practices for DOI Landing Pages*[24] beschrieben. Auf der Landing Page muss im menschenlesbaren Format immer ein vollständiges Zitat des Datensatzes incl. des DOI selbst stehen, so dass der Datensatz eindeutig vom Menschen identifiziert werden kann. Zudem soll die DOI im maschinenlesbaren Teil der Landing Page so gekennzeichnet werden, dass Suchmaschinen sie finden können. Zusätzlich muss eine Landing Page immer Informationen darüber enthalten, wie man auf die Daten zugreifen kann. Falls der Datensatz selbst nicht mehr existiert, muss dies auf der Landing Page vermerkt werden (Tombstone Page).

Sollen die mit dem DOI versehenen Datensätze den FAIR-Prinzipien genügen, müssen zusätzlich bestimmte Anforderungen an die Gestaltung der Landing-Page gestellt werden. Deshalb müssen für den AtMoDat-DOI zusätzlich alle Metadatenfelder, die für den DOI angegeben werden, auch auf der Landing Page aufgelistet werden (siehe Tabellen 1 - 5). Hierbei können auf der Landing Page teilweise andere Begriffe verwendet werden als in dem Metadatenschema des DOI. So kann auf der Landing Page z.B. die Angabe zur Qualitätssicherung, die im DataCite Metadatenschema unter *Related Identifier:IsDerivedFrom* steht, direkt als Feld *Quality Check* bezeichnet werden (siehe Tabelle 4). Zusätzlich sollen bei Daten-Paketen mit mehreren Dateien oder bei Dateien mit mehreren Variablen auch die Beschreibungen

[23]DOI Handbook, Chapter 5: `https://www.doi.org/doi_handbook/5_Applications.html`

[24]https://support.datacite.org/docs/landing-pages

der einzelnen Dateien bzw. Variablen auf der Landing Page stehen, siehe Tabelle 6.

Tabelle 1: Beschreibung des Datensatzes

Landing Page	**Kommentar**	**DataCite Metadatenfeld**
DOI	der DOI selbst	Identifier
Alternate Identifier	gibt es noch einen PID?	Alternate Identifier
Title	Name des Datensatzes/der Simulation	Title
Dataset	immer angeben, dass es sich um Daten handelt. Bei Daten, die auf einem Gitter vorliegen: gridded Data	ResourceType="grid", ResourceTypeGeneral="Dataset"
Keywords	mehrere Keywords angeben	Subject
Summary	Beschreibung der Simulation	Description
Size	Größe des gesamten Datensatzes, für den der DOI vergeben wird (Speicherplatz)	Size
Format	Ausgabeformat(e) der Daten - es können verschiedene Ausgabeformate in einem Daten-Pakt sein	Format
Version	Versionsnummer des Datensatzes, für den DOI vergeben wird	Version
Licence	immer die CC-Lizenzen	Rights

Zur besseren Übersichtlichkeit kann die Landing Page auf mehrere Ebenen aufgespalten werden, indem z.B. die Metadaten in den einzelnen Tabellen jeweils auf eigene Unterseiten geschrieben werden. Die oberste Ebene der Landing Page soll auf jeden Fall die Metadaten der grobgranularen Ebene in Tabelle 1 und eine Liste aller verfügbarer Dateien/Variablen enthalten. Zusätzlich muss für die Metadaten jeder einzelnen Datei/Variablen (feingranulare Ebene) eine weitere Webseite angelegt und von der Landing Page darauf verlinkt werden. Auf diesen Unterseiten können einzelne Felder der grobgranularen Ebene – wie z.B. Rights, Contributor,. . . – wiederholt werden. Sowohl die Landing Page als auch die untergeordneten Webseiten sind persistent anzulegen.

Tabelle 2: Angaben zu Personen, Funding etc.

Landing Page	**Kommentar**	**DataCite Metadatenfeld**
Creator	alle, die an der Erzeugung der Simulation beteiligt waren - Namen und ORCIDs	Creator mit Untertypen
Publisher	Wo wurden die Daten publiziert - in der Regel ein Datenzentrum - Name und ROR oder GRID	Publisher
Contributor	alle Personen oder Institutionen, die am Entstehungsprozess beteiligt waren - Namen und ORCIDs	Contributor
Funding	Angabe der Förderer - Namen und die in der Funder Registry angegebenen URLs	FundingReference

Tabelle 3: Zeitangaben zu Veröffentlichung etc.

Landing Page	**Kommentar**	**DataCite Metadatenfeld**
Publication Year	Wann wurden die Daten publiziert	PublicationYear
Creation Date	wann wurden die Daten erzeugt	Date DateType="Created"
Erzeugung der neuen Version	falls neue Version vorhanden	Date DateType="Updated"
Issued	Datum, an dem die Daten veröffentlicht wurden	Date DateType="Issued"
Available	falls Daten nicht sofort zugänglich gemacht werden	Date DateType="Available"

Damit Suchmaschinen wie Google Dataset Search[25] die einzelnen Webseiten durchsuchen können, soll im maschinenlesbaren Teil der Landing Page das Dataset-Markup von schema.org [Schema.org Steering Group, 2019] oder eine gleichwertige Struktur im DCAT-Format [W3C Dataset Exchange Wor-

[25]https://toolbox.google.com/datasetsearch

Tabelle 4: Verweise auf Modellbeschreibung etc.

Landing Page	**Kommentar**	**DataCite Metadatenfeld**
Boundary Conditions	Link oder DOI zu Beschreibung der Randdaten	RelatedIdentifier relationType="IsDerivedFrom"
Quality Check	Link zur Beschreibung der Dokumentation der Qualitätsprüfung	RelatedIdentifier relationType="IsReviewedBy"
Model Documentation	entweder PID des Modells oder persistenter Link zur Beschreibung des Modells, mit dem die Daten erzeugt wurden	RelatedIdentifier relationType="IsDescribedBy"
References	Zitat und PID der Veröffentlichung, für die die Daten verwendet wurden	RelatedIdentifier relationType="IsCitedBy"
Simulation is part of	bei Simulationen innerhalb eines MIPs	RelatedIdentifier relationType="IsPartOf"
Related Simulations	bei Daten von Simulationen, die in einem MIP gemacht wurden: Verweise auf PIDs mit der Beschreibung der Daten der anderen Simulationen, die zum Vergleich gemacht wurden	RelatedIdentifier relationType="IsVariantFormOf"

king Group, 2019] des W3C verwendet werden – siehe z.B. die Datensatzbeschreibung von Google[26]. DataCite bietet bereits ein Matching der Metadaten in schema.org an. Dabei müssen alle verfügbaren Metadaten in einem der beiden Datenformate DCAT oder schema.org in den maschinenlesbaren Teil der Landing Page übertragen werden. Zusätzlich werden im maschinenlesbaren Teil Sitemaps angelegt, so dass die Webcrawler der Suchmachinen auch die verlinkten Webseiten finden.

Bei einer Suche mit Google oder Bing nach einem Datensatz werden als Ergebnis der Suche der Titel der Landing Page und ein Teil der Description ausgegeben. Um das Suchergebnis möglichst aussagekräftig zu gestalten, sollte der Text im Titel der Landing Page und der verlinkten Webseiten 65 Zeichen nicht überschreiten. Vom Text der Description werden im Su-

[26]https://developers.google.com/search/docs/data-types/dataset

Tabelle 5: Wichtige Angaben zum Datenpaket oder Datensatz, die nicht alle in den DataCite Metadaten stehen, die der Nutzer aber schnell finden soll

Landing Page	**Kommentar**	**DataCite Metadatenfeld**
Model	Name des Modells, mit dem Simulation gemacht wurde	–
Model version	Versionsnummer des Modells, mit dem Simulation gemacht wurde	–
Horizontal Resolution	Räumliche horizontale Auflösung der Daten	–
grid	Art des Gitters	–
Projection	verwendete geographische Projektion	–
Vertical Coordinate	Vertikales Koordinatensystem des Modells, z.B. Höhe, Sigma, Luftdruck	–
Temporal Coverage	Zeitangaben bei Zeitreihen	Date dateType="Valid"
Spatial Coverage	lon/lat Box und/oder Name des Orts bzw. der Region	GeoLocation
Basic Approximations	z.B. hydrostatisch, nichthydrostatisch,...	–

chergebnis bei vielen Suchmaschinen nur die ersten 1-2 Zeilen ausgegeben. Beispielsweise gibt Bing an, dass auf der Ergebnisseite der Suche max. 160 Zeichen ausgegeben werden. Zudem sollten im maschinenlesbaren Teil der Landing Page bei der Beschreibung mit schema.org alle textbasierten Felder, wie z.B. die Description, maximal 5000 Zeichen lang sein, da Google nur in den ersten 5000 Zeichen sucht[27].

Zusätzlich wird empfohlen, dass im schema.org Markup des maschinenlesbaren Teils keine Informationen stehen[28], die nicht auch im menschenlesbaren Teil enthalten sind, da sonst der Verdacht entsteht, dass die Seiten Spam enthalten. Dieses sogenannte Cloaking kann z.B. bei Bing dazu führen[29], dass diese Seiten aus den Suchverzeichnissen gestrichen werden.

[27]https://developers.google.com/search/docs/data-types/dataset

[28]z.B. von Google: https://developers.google.com/search/docs/guides/intro-structured-data

[29]https://www.bing.com/webmaster/help/webmaster-guidelines-30fba23a

Tabelle 6: Zusätzliche Angaben zu den einzelnen Dateien oder Variablen - für jeden einzelnen Datensatz im Daten-Paket oder jede einzelne Variable anlegen.

Landing Page	**Kommentar**
Variable/Dataset Name	Name des Datensatzes oder der Variablen
Temporal Aggregation	Daten sind Monatsmittel, Tagesmittel, etc.
Spatial Aggregation	Daten sind räumliche Mittelwerte, z.B. über Regionen (Nordsee, Europa) etc.
Dimension	Datendimension: 1D, 2D, 3D, 4D
Valid Range	Wertebereich- nur bei einzelnen Variablen pro Datei und falls anwendbar (z.B. bei Windgeschwindigkeiten oder Temperaturen in K)
Size	Größe der jeweiligen Datei (Speicherplatz)
Spatial Coverage	bei sehr kleinen Modellgebieten genaue Angaben, falls sie oben nur grob gemacht werden konnten

6 Dateiformate und Dateistandards für Forschungsdaten

Die Interoperabilität in FAIR soll sowohl für die Metadaten als auch für die Forschungsdaten gelten (siehe Kapitel 2). Die Erfüllung beider Aspekte unterscheidet sich stark voneinander. Auf den ersten Aspekt, bei dem es hauptsächlich auf die Menschen- und Maschinenlesbarkeit von Metadata auf der Landingpage ankommt, wird vertieft in den Kapitel 4 und 5 eingegangen. Der zweite Aspekt wird in diesem Kapitel behandelt.

6.1 AtMoDat Anforderungen an das Dateiformat

Eine Bedingung für die Interoperabilität von Forschungsdaten ist, dass sie in einem selbstbeschreibenden offenen Dateiformat gespeichert sind und die Struktur innerhalb der Datei einer von Menschen und *Maschinen* verstandenen/interpretierbaren Konvention folgt. Um dies zu erreichen, werden neben den Daten auch einige Metadaten in der Datei abgespeichert – beispielsweise Bezeichnungen von Variablen, Einheiten, Linzenzinformationen oder Kontaktdaten. Ausführlichere Zusatzinformationen, die zum Verständnis, der Prüfung oder zur Auswertung der abgespeicherten Daten benötigt werden, werden durch URIs (Uniform Resource Identifier) oder im besten Fall durch PIDs (Persistent Identifier) verlinkt. Das Vorgehen soll dem Verlinken von Vokabularen, Schemata oder ähnlichem in RDF (Resource Description Framework) [W3C RDF Working Group, 2014] ähneln, um einer semantischen Verknüpfung von Daten näher zu kommen.

6.2 In Frage kommende Dateiformate und Standards

6.2.1 netCDF

Das Network Common Data Format (netCDF) erfüllt einige dieser Anforderungen. Es ist selbstbeschreibend und offen dokumentiert. Die Standard-netCDF-Software ist open source und kostenfrei verfügbar für Linux, macOS und Windows. Es ist ein binär-Format und seit der Einführung von netCDF Version 4 im Jahr 2008 ist eine Art Komprimierung der Daten möglich (*deflate* genannt). Im Bereich der Atmosphärenmodellierung und generell in der Erdsystemmodellierung ist netCDF weit verbreitet, z.B. bei CMIP6 (Juckes et al. [2020]) und für das Marine Rapid Environmental Assessment (MREA, Signell [2008]).

Eine netCDF Datei enthält Dimensionen, Variablen und Attribute. Variablen enthalten die eigentlichen Daten und Informationen zu den zeitlichen

und räumlichen Koordinaten. Dimensionen sind die Zeit- und Raumachsen, an denen sich Variablen aufspannen. Eine modellierte Lufttemperatur in mehreren Höhenschichten, räumlich horizontal gegittert und in stündlicher Auflösung über Europa hätte vier Dimensionen: zwei horizontale, eine vertikale und eine zeitliche. Attribute enthalten Zusatzinformationen zu einzelnen Variablen oder zur gesamten Datei (*global*). Eine Variable *temperature* kann beispielsweise das Attribut *units* mit dem Wert *K* (für Kelvin) haben. Jede Datei enthält das globale Attribut *creator* mit Informationen zum Erzeuger der Datei.

NetCDF ist selbstbeschreibend. Allerdings kann der Nutzer frei entscheiden, wie er Variablen, Dimensionen und Attribute bezeichnet und was in ihnen steht. Dies schränkt die Menschen- und Maschinenlesbarkeit deutlich ein. Daher gibt es mehrere Standards, die Vorgaben für Bezeichnungen, Einheiten etc. machen, mit denen die Dateiinhalte formalisiert und automatisiert verarbeitbar gemacht werden. In der Atmosphärenmodellierung haben sich die *Climate and Forecast Metadata Conventions* (CF Conventions[30]) in den letzten Jahr als der meistgenutzte Standard etabliert. Sie werden aktiv weiter entwickelt und jede neue Version ist weitestgehend zu alten Versionen abwärtskompatibel. Zusätzlich bieten die *Attribute Conventions for Data Discovery* (ACDD Conventions [31]) eine standardisierte Liste an globalen Attributen zur Angabe detaillierter Metadaten. Die ACDD Conventions sind hauptsächlich für Anwendungsfälle gedacht, bei denen alle Metadaten für die Datensatzsuche automatisiert aus dem Header von netCDF Dateien extrahiert und nicht gesondert erfasst werden. Die Veröffentlichung von Daten über einen THREDDS[32] Data Server (TDS)[33] wäre ein solcher Anwendungsfall. Im Fall dieses Kernstandards wird eine separate Landing Page gefordert (siehe Kapitel 5), weshalb eine Nutzung der ACDD Conventions nicht nötig ist. Trotzdem bieten sie einige nützliche Attribute, die einem Nutzer die Nachnutzung von Daten vereinfachen (siehe Tabelle 7).

Die Modellergebnisse von CMIP5 und CMIP6 wurden in netCDF gespeichert, siehe [Eyring et al., 2016]. Die CMIP5- und CMIP6-Standards setzen auf den CF Conventions auf ([Juckes et al., 2020]), machen aber striktere Vorgaben. Eine Reihe globaler Attribute sind verpflichtend vorgegeben und ihre Werte sind aus kontrollierten Vokabularen zu wählen. Verpflichtende globale Attribute sind beispielsweise die Modellbezeichnung, die Gitterauflösung

[30] `http://cfconventions.org/`

[31] `http://wiki.esipfed.org/index.php/Attribute_Convention_for_Data_Discovery_1-3`

[32] Thematic Real-time Environmental Distributed Data Services

[33] `https://docs.unidata.ucar.edu/tds/5.0/userguide/index.html` or `https://github.com/Unidata/tds`

oder die verantwortliche Forschungsinstitution. Zusätzlich bestehen Vorgaben an die räumlichen Koordinaten. Dieses Vorgehen vereinfacht eine automatische Auswertung und Qualitätskontrolle deutlich. Gleichzeitig ist der Aufwand CMIP6-konforme netCDF Dateien zu erstellen relativ hoch, was eine deutliche Hürde für die Anwendung des CMIP6-Standards auf weitere Teildisziplinen ist.

6.2.2 GRIB

Das *general Regularly-distributed Information in Binary form* (GRIB) Format wurde für gegitterte meteorologische Daten entwickelt und durch die Kommission für Basissysteme der Weltorganisation für Meteorologie (WMO) standardisiert. GRIB ist ein komprimiertes binäres Dateiformat. Werden Modelldaten im GRIB Format gespeichert, verbrauchen sie deutlich weniger Speicherplatz verglichen mit einer Speicherung im komprimiertem (*deflated*) netCDF4. Daher findet das GRIB Format gerade bei der internen Archivierung operationeller regionaler Wettervorhersagemodelldaten und hochaufgelöster globaler Modelldaten häufig Anwendung. Über die Meteorologie hinaus ist GRIB wenig verbreitet. Darüber hinaus sind GRIB Daten nicht selbstbeschreibend wie netCDF Dateien. Stattdessen werden Informationen zu Variablen in externen Tabellen abgelegt. Dadurch ist die Verarbeitung von GRIB Daten fehleranfälliger falls sich der Ersteller nicht an die Vorgaben der WMO gehalten, keine GRIB Tabellen bereitgestellt oder Dateien unpassend benannt hat.

Das GRIB Format eignet sich sehr gut zur institutsinternen Archivierung und Übertragung von Modell- oder Satellitendaten auf Grund geringerer Dateigrößen als netCDF. Für die Veröffentlichung von Modelldaten ist das GRIB Format gegenüber netCDF aus den oben beschriebenen Gründen eher unpraktisch. Dies gilt insbesondere, wenn Modelldaten für Zielgruppen außerhalb der Meteorologie bereitgestellt werden.

6.2.3 Weitere Formate

Neben netCDF existieren weitere Formate, die selbstbeschreibend sind und zum Teil in Teildisziplinen der Atmosphärenmodellierung genutzt werden. Beispielsweise sind hochaufgelöste Gelände- und Gebäudedaten in Deutschland weiträumig im sogenannten CityGML Format verfügbar[34]. CityGML baut auf GML – einem XML-Standard für geographische Daten – auf [ISO Central Secretary, 2007][35]. Dateien im CityGML Format werden hauptsächlich

[34] `https://www.opengeospatial.org/standards/citygml`

[35] siehe auch `https://www.opengeospatial.org/standards/gml`

als Eingangsdaten für hindernisauflösende Modellsimulationen mit Fokus auf städtische Räume genutzt. Wir gehen davon aus, dass netCDF aktuell das am meisten genutzte Dateiformat für die Veröffentlichung und Weitergabe von Ergebnissen der Atmosphärenmodellierung ist.

6.3 Vorgaben für diesen Kernstandard

Modelldaten, die entsprechend dem AtMoDat Kernstandard veröffentlicht werden, sollen in netCDF abgespeichert werden, weil es etabliert und standardisiert ist und viele Anforderungen erfüllt. Es darf auf alternative Formate ausgewichen werden, solange diese in der entsprechenden Teildisziplin etabliert, selbstbeschreibend und mit quelloffener Software nutzbar sind. Es wird von der Veröffentlichung von Daten im GRIB Format abgeraten (Details in Kapitel 6.2.2).

Eine netCDF Datei, die mit dem AtMoDat Standard veröffentlicht wird, muss den CF Conventions[30] folgend strukturiert sein und zusätzliche globale Attribute enthalten, die vom CMIP6 Standard und von den ACDD Conventions[31] abgeleitet sind. Eine AtMoDat-netCDF Datei ist somit strikter strukturiert als eine pure CF-netCDF Datei aber weniger strikt als eine CMIP6-netCDF Datei. Tabelle 7 enthält eine Liste an verpflichtenden (*mandatory*), empfohlenen (*recommended*) und optionalen (*optional*) globalen Attributen.

Tabelle 7: Für AtMoDat Standard verpflichtende (*status:* M), empfohlene (R), optionale (O) und spezielle (S, siehe Fußnote) globale Attribute für netCDF Dateien. Die Spalte *vocabulary* gibt an, aus welchem Standard-Vokabular dieses Attribut ursprünglich stammt.

name	**type**	**status**	**vocabulary**
comment	string	O	CF-v1.8
contact	string	M	CMIP6
Conventions	CV	M	CF-v1.8
creator	string	M	neu
crs[a]	string	R	neu
featureType	string	S[b]	CF-v1.8
frequency	CV	R	CMIP6
further_info_url	CV	R	CMIP6
geospatial_lat_resolution	number	R	ACDD-v1.3
geospatial_lat_units	string	O[c]	ACDD-v1.3
geospatial_lon_resolution	number	R	ACDD-v1.3
geospatial_lon_units	string	O[c]	ACDD-v1.3
geospatial_vertical_resolution	number	R	ACDD-v1.3
geospatial_vertical_units	string	O[c]	ACDD-v1.3
history	string	R	CF-v1.8
institution	CV	M	CF-v1.8
institution_id	CV	R	CMIP6
keywords	string	R	ACDD-v1.3
keywords_vocabulary	string	O	ACDD-v1.3
license	string	M	CMIP6
metadata_link	string	O	ACDD-v1.3
nominal_resolution	CV	R	CMIP6
processing_level	string	O	ACDD-v1.3
product_version		R	ACDD-v1.3
program	string	O	ACDD-v1.3
project	string	O	ACDD-v1.3
references	string	O	CF-v1.8
source	registered content	M	CF-v1.8
source_type	CV	M	CMIP6
standard_name_vocabulary	string	R	ACDD-v1.3
summary	string	R	ACDD-v1.3
title	string	M	CF-v1.8

[a] *crs* (coordinate reference system) enthält eine Angabe zum Koordinaten Referenzsystem, z.B. "*wgs84*"
[b] Das globale Attribut *featureType* ist verpflichtend, wenn die Daten vom Type point, time series (at one spatial location), trajectory, profile, time series profile or trajectory profile sind. Das globale Attribut *featureType* darf nicht gesetzt werden, wenn es sich um räumlich gegitterte Daten handelt. Details sind im Kapitel *Discrete Sampling Geometries* der CF Conventions zu finden[30].
[c] Gibt die Einheit zu *geospatial_*_resolution* an. Wenn *geospatial_*_units* nicht gesetzt, wird für die Werte in *geospatial_*_resolution* angenommen, dass sie in *Grad Nord/Ost* (°N/°) angegeben sind. Details in den ACDD Conventions[31].

7 Sichtbarmachung von Qualitätsstandards

Die FAIR-Prinzipien sind Kriterien für eine qualitativ hochwertige Veröffentlichung von Daten. Ihre Einhaltung wird jedoch gegenwärtig nicht direkt dokumentiert. Es gibt jedoch mehrere Verfahren, um die FAIRness eines Datensatzes zu bewerten. Vergleiche dieser Verfahren sind in Wilkinson et al. [2018] und Bahim et al. [2019] dargestellt. Einige dieser Verfahren berücksichtigen zusätzliche Qualitätskriterien, die über FAIR hinaus gehen.

Sowohl für den Datenproduzenten, den Datennutzer als auch für das Repositorium ist es hilfreich, wenn die Metadaten auch eine Information über die Qualität der Daten enthalten. Dabei wird die Datenqualität durch folgende Maßnahmen erhöht:

- Die Daten und ihre Metadaten sind FAIR.
- Bei Daten und Metadaten werden disziplinspezifische Standards eingehalten.
- Das veröffentlichende Datenrepositorium hat die Metadaten auf Korrektheit und Vollständigkeit geprüft.
- Das veröffentlichende Datenrepositorium hat die Daten auf Vollständigkeit geprüft.
- Unabhängige Prüfung der Daten auf Plausibilität (ähnlich einem Peer-Review). Eine Prüfung der Daten auf Korrektheit ist schwierig, da nicht immer eindeutig festlegbar ist, wann die Daten inhaltlich korrekt sind.[36]

Dadurch findet ein Datennutzer eher die Daten, die er sucht, und kann sie leichter weiternutzen als z.B. Daten in einem proprietären Format ohne disziplinspezifische Standardisierung. Die Arbeit eines Datenproduzenten, der der wissenschaftlichen Community und darüber hinaus qualitative hochwertige Datensätze zur Verfügung stellt, erfährt auf diese Weise Anerkennung. Datenrepositorien, die eine hochwertige Prüfung und Datenkuration durchführen und hochwertige Datensätze anbieten, können leichter gewürdigt werden.

Dieser Qualitätsstandard kann auf zwei Wegen umgesetzt werden:

1. Es wird ein rein disziplinspezifischer Qualitätsstandard erstellt (z.B. der hier vorgestellte Standard für den AtMoDat-DOI).
2. wie 1., aber zusätzlich wird ein generischer Qualitätsstandard definiert, dessen Umsetzung eine disziplinspezifische Erweiterung erfordert.

[36] Fragen zur Korrektheit von Modelldaten: Wann sind Ausreißer als korrekt anzusehen und wann nicht? Wann ist ein Modell korrekt und wann nicht?

7.1 Rein disziplinspezifischer Qualitätsstandard

In den vorhergehenden Kapiteln wurde ein disziplinspezifischer Qualitätsstandard für Daten aus der Atmosphärenmodellierung beschrieben, der sich auf andere nahestehende Disziplinen wie beispielsweise Ozeanmodellierung übertragen lässt. Der Standard (bzw. die Einhaltung dieses Standards) könnte durch folgende Maßnahmen kenntlich gemacht werden:

- eigenes Logo bzw. eigener Schriftzug für den Standard: AtMoDat DOI, AtMoDatQ oder ähnliches;
- Logo oder Schriftzug stehen auf der Landing Page der standardisierten Datensätze und sind mit einer Information zum Standard verlinkt.
- Die standardisierten Datensätze sind durch ein Branding der DOIs markiert, indem der DOI-Suffix immer mit AtMoDat beginnt.

Die konkreten Maßnahmen werden bis Ende des Projekts (Mai 2022) festgelegt.

7.2 Generischer Qualitätsstandard

Da die Qualität der Daten für alle Fachdisziplinen von hoher Bedeutung ist, wäre die Einführung eines allgemeinen Qualitätsstandards sinnvoll, der die Nutzung von disziplinspezifischen Metadatenstandards und Qualitätsprüfungen fördert.

Kern dieses generischen Standards wäre, dass veröffentlichte Datensätze anhand der vorhanden Metadaten, genutzter Dateiformate und erfolgter Prüfungen durch das Datenrepositorium in unterschiedliche Kategorien/Levels eingeteilt werden. Anhand eines solchen Qualitätslevelsystems kann ein Datennutzer erkennen, wie „gut“ ein Datensatz durch Metadaten beschrieben ist, ob die Korrektheit der Daten und Metadaten geprüft wurde und ob die Daten in einem für ihn nutzbaren Format vorliegen.

Eine solche Einteilung **könnte** am Beispiel der bei DataCite registrierten Datensätze wie folgt aussehen (*Beispiel*):

- **Level 0:** minimaler Satz an DataCite Metadaten gefüllt
- **Level 1:** DataCite Metadaten soweit wie möglich gefüllt; Level 0 ist erfüllt
- **Level 2:** DataCite Metadatenschema um disziplinspezifische Metadatenfelder erweitert; Level 1 ist erfüllt

- **Level 3:** FAIR Kriterien erfüllt; offenes Dateiformat wurde genutzt und disziplinspezifischer Standard wurde für Dateiinhalte angewandt; Level 2 ist erfüllt
- **Level 4:** Landing Page wir vorgeschrieben; Level 3 ist erfüllt
- **Level 5:** Metadaten vom Datenrepositorium auf korrekte Befüllung getestet wurden; Level 4 ist erfüllt
- **Level 6:** Eine unabhängige inhaltliche Prüfung der Daten fand statt; Level 5 ist erfüllt

Die Definition von weiteren Leveln ist möglich. Die Level 2, 3, 4 und 5 werden aktuell schon von manchen disziplin-spezifischen Repositorien eingehalten. Level 6 wird vermutlich aktuell von keinem Repositorium eingehalten und ist eher als Ausblick zu sehen, wie eine Erweiterung dieser Level in Zukunft aussehen könnte.

Das DataCite Metadatenschema müsste um ein Metadatenfeld (property) erweitert werden, um diesen Qualitätsstandard umzusetzen. Dieses neue Feld würde in einem Unterfeld (sub-property) die Informationen zum Qualitätslevel enthalten. In einem weiteren Unterfeld könnten die Informationen (PID / URL) zu einem Dokument mit einer disziplinspezifischen Dokumentation der Qualitätslevel stehen.

Die Einführung eines solchen Qualitätsfeldes wird von AtMoDat empfohlen und in Fachvorträgen beworben. Dadurch soll eine Initiative zur Einführung dieses Feldes ins DataCite Metadatenschema eine breitere Unterstützung finden. Da dieser Prozess der Einführung von neuen Feldern meist mehrere Jahre dauert, wird für den AtMoDat-DOI zuerst der Ansatz für den disziplinspezifischen Qualitätsstandard verfolgt.

8 Zusammenfassung

Datensätze, die den AtMoDat-DOI bekommen, sollen FAIR sein. Dies beinhaltet, dass nicht nur die von DataCite geforderten (mandatory) Metadaten mitgeliefert werden, sondern dass die Liste mit verpflichtenden Metadaten deutlich erweitert wird. Dabei können die meisten Metadaten, die notwendig sind, um einen Datensatz so zu beschreiben, dass er von anderen Forschenden weiter verwendet werden kann, in existierenden Metadatenfeldern von DataCite zusätzlich eingetragen werden. Dadurch werden die Datensätze von den Suchmaschinen gefunden, die die Datenbank von DataCite durchsuchen. Für alle Suchmaschinen, die dies nicht tun, müssen diese Informationen auf die Landing Page geschrieben werden.

Da die Kuratierung der Metadaten zu einem DOI aufwendig ist und in Zukunft die Vergabe eines DOI kostenpflichtig sein wird, vergeben viele Datenzentren für eine gesamte Simulation mit z.T. mehreren Datensätzen nur einen DOI. Dies bildet die grobgranulare Ebene eines Datensatzes, der auf der feingranularen Ebene entweder mehrere Dateien oder eine Datei mit mehreren Variablen enthält. Die Granularität der Daten spiegelt sich dann auch in den Metadaten wieder. Beispielsweise beschreiben die Metadaten der grobgranularen Ebene die Simulation und die Metadaten der feingranularen Ebene die einzelnen Datensätze bzw. die einzelnen Parameter/Variablen. Dabei müssen manche Metadatenfelder auf der grob- und feingranularen Ebene identisch sein, insbesondere die Vergabe der Rechte für die Nutzung der Daten. Andere Metadatenfelder sollten verschieden sein, wie sich z.B. der Titel einer Simulation sich von der Titeln der einzelnen Variablen unterscheiden sollte.

Die zu einem Datensatz gehörende Landing Page besteht immer aus einen menschen- und einem maschinenlesbaren Teil, die sich z. T. in ihren Inhalten unterscheiden können. Die Granularität der Datensätze bestimmt den Aufbau der Landing Pages. So wird auf der obersten Ebene der Landing Page die grobgranulare Ebene der Daten beschrieben. Die Metadaten der feingranularen Ebene können zwar ebenfalls auf der obersten Ebene der Landing Page stehen. Da man auf der feingranularen Ebene aber meist mehrere verschiedene Parameter oder Dateien beschreiben muss, wird meist durch ein Auswahlmenü auf eine tiefere Ebene der Landing Page verlinkt. Eine so aufgebaute Kette von Landing Pages kann auch im maschinenlesbaren Teil die Metadaten der jeweiligen Granularitätsebene enthalten. Suchmaschinen wie Google Data Search, die die Landing Pages auswerten, durchsuchen auch die verlinkten Seiten, so dass in diesem Fall auch eine Variable oder ein Teildatensatz einer Simulation gefunden werden kann. Bei Suchmaschinen, die nur die DataCite Datenbank durchsuchen, können nur die Metadaten zur grobgranularen Ebene ausgewertet werden. In diesem Fall muss der Suchende erst auf die Landing Page gehen, um weitere Informationen zu verfügbaren Variablen zu erhalten.

Zusätzlich zum vollständigen und korrekten Ausfüllen der Metadaten müssen die Daten selbst in einem offenen selbstbeschreibenden Format abgespeichert sein. Sofern nicht durch das Format bereits definiert, muss die Darstellung innerhalb der Datei einem allgemein bekannten oder in der Datei verlinkten Standard entsprechen – z.B. müssen Variablen einheitlich benannt, standardisierte Einheiten genutzt und zusätzliche zur Weiterverarbeitung der Daten benötigte Informationen in der Datei hinterlegt werden. Im Fall der Atmosphärenmodelldaten bieten sich netCDF-Dateien an, die nach den CF Conventions standardisiert sind. Solche netCDF-CF-Dateien entsprechen dem oben genannten Kriterien und sind in vielen Teildisziplinen

der Atmosphärenmodellierung weit verbreitet. Hinzu werden weitere globale Attribute zusammen mit kontrollierten Vokabularen aus dem CMIP6 Standard übernommen.

Damit Forschungsdaten die FAIR-Kriterien einhalten, müssen sie also nicht nur einen DOI bekommen. Es reicht nicht, nur die von DataCite verpflichtenden Metadaten anzugeben, sondern die gesamte Metadatenliste von DataCite muss ausgefüllt werden. Zudem müssen die Landing Pages im maschinenlesbaren Teil so gestaltet werden, dass sie von Suchmaschinen ausgewertet werden können. Zuletzt müssen die Daten in einem passenden Format gespeichert sein.

Literatur

Christophe Bahim, Makx Dekkers, and Brecht Wyns. Results of an analysis of existing fair assessment tools, 2019.

DataCite Metadata Working Group. Datacite metadata schema documentation for the publication and citation of research data. Metadataschema v4.3, DataCite e.V., 2019.

V. Eyring, S. Bony, G. A. Meehl, C. A. Senior, B. Stevens, R. J. Stouffer, and K. E. Taylor. Overview of the coupled model intercomparison project phase 6 (cmip6) experimental design and organization. *Geosci. Model Dev.*, 9(5):1937–1958, 2016. doi: https://doi.org/10.5194/gmd-9-1937-2016.

ISO Central Secretary. Geographic information – temporal schema. Standard ISO/TC 211 19108:2002, International Organization for Standardization, Geneva, CH, 2002. URL `https://www.iso.org/standard/26013.html`.

ISO Central Secretary. Data elements and interchange formats – information interchange – representation of dates and times. Standard ISO/TC 154 8601:2004, International Organization for Standardization, Geneva, CH, 2004. URL `https://www.iso.org/standard/40874.htmll`.

ISO Central Secretary. Geographic information – geography markup language (gml). Standard ISO/TC 211 19136:2007, International Organization for Standardization, Geneva, CH, 2007. URL `https://www.iso.org/standard/32554.html`.

M. Juckes, K. E. Taylor, P. J. Durack, B. Lawrence, M. S. Mizielinski, A. Pamment, J.-Y. Peterschmitt, M. Rixen, and S. Sénési. The cmip6 data request (dreq, version 01.00.31). *Geoscientific Model Development*, 13(1):201–224, 2020. doi: https://doi.org/10.5194/gmd-13-201-2020.

Schema.org Steering Group. Schema.org, 2019. URL `http://schema.org/`. accessed 2019-12-12.

Sandro; Chiggiato Jacopo; Janekovic Ivica; Pullen Julie; Sherwood Christopher R. Signell, Richard P.; Carniel. Collaboration tools and techniques for large model datasets. *Journal of Marine Systems*, 69(1):154 —- 161, 2008. doi: https://doi.org/10.1016/j.jmarsys.2007.02.013.

M. Stockhause, H. Höck, F. Toussaint, and M. Lautenschlager. Quality assessment concept of the World Data Center for Climate and its application to CMIP5 data. *Geoscientific Model Development*, 5(4):1023–1032, 2012. doi: https://doi.org/10.5194/gmd-5-1023-2012.

W3C Dataset Exchange Working Group. Data catalog vocabulary, 2019. URL `https://www.w3.org/TR/vocab-dcat-2/`. accessed 2019-12-12.

W3C RDF Working Group. Resource description framework, 2014. URL `https://www.w3.org/RDF/`. accessed 2019-12-12.

Mark D. Wilkinson, Micheland Dumontier, I. Jsbrand Jan Aalbersberg, Gabrielle Appleton, Arie Axton, Mylesand Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Jildau Bourne, Philip E.and Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J. G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A. C. 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, 3:160018, 2016. doi: https://doi.org/10.1038/sdata.2016.18.

Mark D. Wilkinson, Michel Dumontier, Susanna-Assunta Sansone, Luiz Olavo Bonino da Silva Santos, Peter Prieto, Mario; McQuilton, Julian Gautier, Derek Murphy, Mercé Crosas, and Erik Schultes. Evaluating fair-compliance through an objective, automated, community-governed framework, 2018.