Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Ausarbeitung

Wissenschaftliche Datenformate

vorgelegt von

Lukas Ciepielowski

Fakultät für Mathematik, Informatik und Naturwissenschaften
Fachbereich Informatik
Arbeitsbereich Wissenschaftliches Rechnen

Studiengang: Wirtschaftsinformatik
Matrikelnummer: 7028745

Betreuer: Dr. Hermann Lenhart

Hamburg, 16.03.2019

Inhaltsverzeichnis

1 Einleitung

Daten sind ein wichtiger Bestandteil der Wissenschaft. Damit diese vorteilhaft gespeichert und verwendet werden können, gibt es eine zahlreiche Menge an verschiedenen Datenformaten. Datenformate entstehen durch die unterschiedlichen Anforderungen aus den Bereichen der Wissenschaft. Standards, vor allem in der Wissenschaft, sind sehr wichtig, da somit eine reibungslose Kommunikation zwischen Wissenschaftlern und Instituten ermöglicht wird. Die Datenformate enthalten meistens Metadaten. Damit sind strukturierte Daten gemeint, welche Informationen über die eigentlichen Daten enthalten und zudem oft maschinell lesbar sind. [1] Dies ist bei besonders großen Datenmengen von großer Bedeutung. Die Daten werden in N-Dimensionalen Arrays gespeichert. Ein Array besteht aus Elementen eines gleichen Typs, welche mit einem Index versehen werden, wodurch dann mithilfe einer Indizierung effizient auf die Informationen zugegriffen werden kann.

1.1 Voraussetzungen

Es gibt einige Voraussetzungen, welche durch das Datenformat gegeben sein sollten, damit es sich für die Verwendung in der Wissenschaft eignet. Ersteres wäre zunächst die kompakte Speicherung bzw. die Möglichkeit die Daten zu komprimieren. Daraus ergibt sich auch die Notwendigkeit der Portabilität. In Bereichen der Wissenschaft, in denen Daten oft ausgetauscht werden, hat dieser Aspekt eine große Bedeutung. Weitergehend ist ein schneller Zugriff wünschenswert, da dieser die Arbeit umfassend erleichtern kann. Damit die Daten auch gezielt bearbeitet und ausgelesen werden können, sollte es ein Angebot an Software geben, die diese Aspekte der Bearbeitung unterstützt. Um dies zu gewährleisten, ist es von Vorteil, wenn es eine Programmierschnittstelle (API) für ein möglichst weites Spektrum an verschiedenen Programmiersprachen gibt.[2]

[1] Vgl. Andreas Pfund: "Metadaten", `http://andreas-pfund.de/definition/metadaten/metadaten.php`, 09.03.2019 [1]

[2] Meelis Kull: "Scientific Data Formats", `http://kodu.ut.ee/~varmo/tday-kaariku/kull-slides.pdf`, 03.01.2019 [2]

1.2 Beispiele für Datenformate

Ein Beispiel aus dem Bereich der Klimatologie ist NetCDF (Network Common Data Format). Ich werde NetCDF später in dieser Ausarbeitung als genaueres Beispiel verwenden. Im Bereich der Biologie gibt es beispielsweise das PDB (Protein Data Bank) file format. Die PDB ist dabei zunächst eine Datenbank, die Informationen zu Molekülen speichert. Es handelt sich beim PDB file format um ein textuelles Format. In der PDB werden beispielsweise Informationen zu Proteinen und Atomkoordinaten in Form von einer Tabelle gespeichert. [3] Aus dem Bereich der Medizin ist DICOM (Digital Imaging and Communications in Medicine) ein weit verbreitetes Datenformat. Es ist ein offener Standard zur Speicherung und zum Austausch von Bildinformationen. Fast alle Hersteller von bildgebenden Systemen, wie zum Beispiel Röntgengeräte, implementieren den DICOM-Standard. Der DICOM Datensatz dient als Container, in welchem dann die Informationen der Behandlung eingefügt werden können. Selbstverständlich müssen auch Metadaten eingepflegt werden, wie der Patientenname, das Aufnahmedatum oder der Arztname. Es wird in einem Real World Information Model gearbeitet, das in die Stufen Patient, Studie, Serie und Instanz unterteilt ist. Jede Instanz eines DICOM Datensatzes enthält somit alle Informationen, welche zu einer Serie (Bild-Serie), einer Studie, und schließlich zu einem Patienten zugeordnet werden können.[4]

1.3 Notwendigkeit von Datenformaten

In der Klimatologie werden selbstverständlich mit der Zeit, durch fortschrittlichere Technologie, immer mehr Klimadaten gesammelt. Dies ist notwendig, um das Klima bzw. den Klimawandel besser zu verstehen. Klimadaten können durch in situ Beobachtungen, durch Satelliten oder aus Modellen gewonnen werden. Es steigt jedoch nicht nur die Menge an Rohdaten an sich, vielmehr erhöht sich die Bedeutung der Daten, welche aus Klimamodellen erlangt werden (siehe Figure 1.1). Um Klimamodelle aufstellen zu können und anschließend daraus Erkenntnisse zu erlangen, werden Datenformate benötigt. Diese sorgen für eine stabile Grundlage, um unsere Daten speichern und analysieren zu können. Aus dem Zuwachs der Modelldaten, ergibt sich auch die steigende Notwendigkeit für Datenformate, welche eine effiziente und schnelle Arbeit mit den Daten ermöglichen. Aus diesem Grund kommen wir nun zu NetCDF, dem Datenformat aus der Klimatologie.

[3] Wikipedia: "Protein Data Bank (file format)", `https://en.wikipedia.org/wiki/Protein_Data_Bank_(file_format)`, 10.03.2019 [4]

[4] Wikipedia: "Digital Imaging and Communications in Medicine", `https://de.wikipedia.org/wiki/Digital_Imaging_and_Communications_in_Medicine`, 10.03.2019 [5]

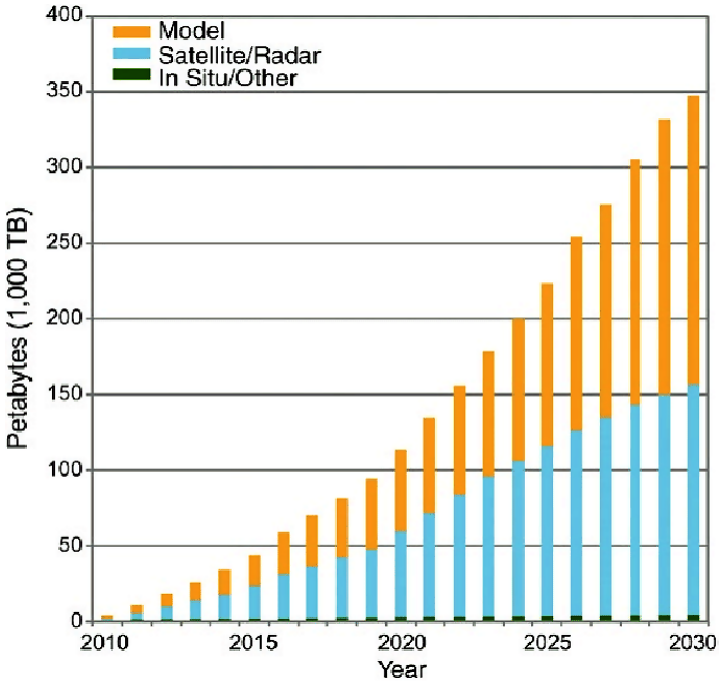

Figure 1.1: Zuwachs der Datenmenge in der Klimatologie
(Quelle: Jonathan T.Overpeck: "Climate Data Challenges in the 21st Century", `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.465.7115&rep=rep1&type=pdf`, 11.03.2019)

2 NetCDF

NetCDF ist ein selbstbeschreibendes Datenformat für den Austausch von wissenschaftlichen Daten. Ein Header in der NetCDF Datei beschreibt das Layout der Datei. Metadaten sind dort auch enthalten. NetCDF ist außerdem plattformübergreifend, somit kann eine erstellte Datei unabhängig vom Betriebssystem geöffnet werden. Außerdem ist es möglich, eine Teilmenge einer großen Datei effizient aufzurufen. Eine NetCDF Datei besteht aus Dimensionen, Variablen und Attributen (siehe Figure 2.1).

2.1 Dimensionen

Eine Dimension enthält einen Namen und eine Größe. Diese werden beim Erstellen der Dimension definiert. Damit ist die Größe fest vorgegeben. Dimensionen werden genutzt, um physikalische Größen, wie etwa Zeit, Breite, Länge oder Höhe, darzustellen. Es handelt sich um einen ganzzahligen positiven Wert. Innerhalb einer Datei kann es jedoch genau eine Dimension der Größe UNLIMITED geben. Die Größe wird also nicht vorgegeben. Bei NetCDF-4, einer späteren Version, ist es möglich mehrere solcher Dimensionen zu erstellen.

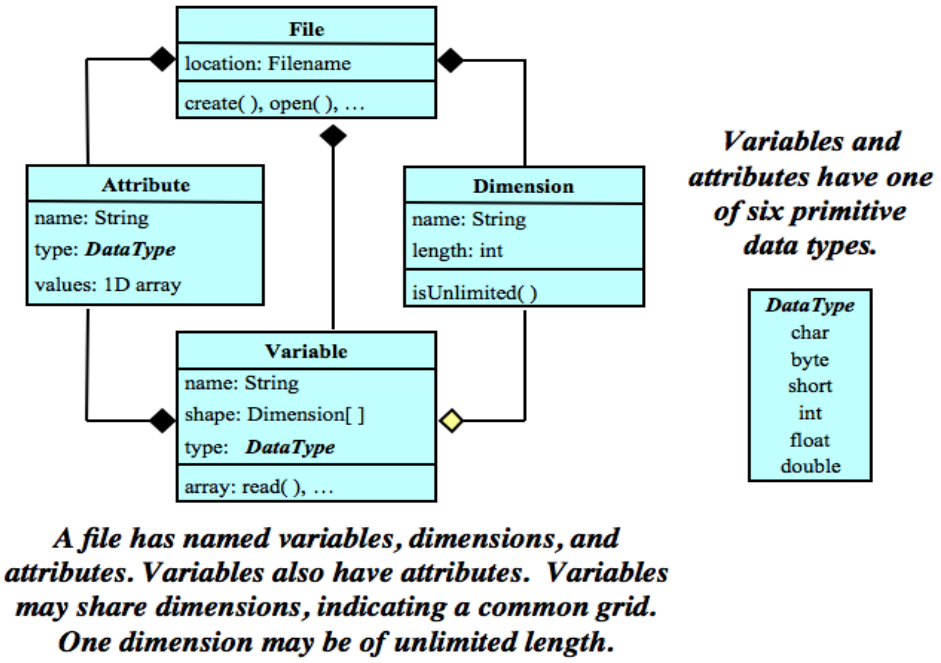

Figure 2.1: NetCDF Classic Data Model
(Quelle:"The Classic Model", `https://www.unidata.ucar.edu/software/netcdf/docs/netcdf_data_model.html`, 11.03.2019)

2.2 Variablen

Die Daten in NetCDF werden in Variablen gespeichert. Eine Variable ist ein Array mit Werten desselben Datentyps. Sie besteht aus Namen, Typ und Form. Eine Variable bezieht sich auf Dimensionen. Dementsprechend wird die Form durch die festgelegte Dimension gegeben. Eine "record variable" ist eine Variable mit einer unlimitierten Dimension. Enthält die Variable keine Dimension, so kann nur ein einziger Wert gespeichert werden. Solche Variablen nennt man Skalar. Variablen mit einer Dimension werden als Vektor und Variablen mit zwei Dimensionen als Matrix bezeichnet. Die Anzahl der Dimensionen einer Variable wird als Rang bezeichnet.

2.3 Koordinatenvariablen

Bei den Koordinatenvariablen handelt es sich um eine spezielle Form einer Variable. Eine Koordinatenvariable hat den gleichen Namen wie eine Dimension. Es handelt sich um eine Vektor Variable mit der identischen Form zu ihrer Dimension. Koordinatenvariablen werden genutzt, um physikalische Koordinaten in einem Koordinatensystem abzubilden. Sei also beispielsweise eine Dimension **Längengrad** der Größe 5 gegeben. Die dazugehörige Koordinatenvariable hat folglich auch den gleichen Namen **Längengrad**. Die Koordinatenvariable enthält also 5 Werte. Sei nun eine Variable **Temperatur** mit der Dimension **Längengrad** gegeben. Jeder Wert beschreibt dabei die Temperatur an

einem Längengrad. Nun könnte eine Visualisierungssoftware die Temperaturwerte für die Längengrade anzeigen, anstatt nur die Position im Array.

2.4 Attribute

Attribute werden genutzt, um Informationen über gespeicherte Daten bereitzustellen. Sie stellen also die Metadaten in NetCDF dar. Die Attribute enthalten meistens Informationen über eine Variable. Sie werden mit dem Namen der Variable und einem eigenen Namen versehen. Anderseits gibt es auch globale Attribute, welche Informationen über die gesamte NetCDF Datei enthalten. Diese werden mit dem Attributennamen sowie mit einem leeren Variablennamen definiert. [6] [7]

3 NetCDF File Format

Die erste Version von NetCDF war das NetCDF Classic Format. Dort war die Größe zunächst auf 2 GiB beschränkt. Ziel war dabei aber die Portabilität der Dateien. Es wird immer noch als Default Format angesehen und wenn möglich auch empfohlen. Das NetCDF 64-bit Offset Format entlastet die Beschränkung der Größe. Die einzelnen Variablen sind jedoch trotzdem auf 4 GiB beschränkt. [8]

3.1 NetCDF-4

NetCDF-4 stellt eine besondere Version dar, da diese auf HDF5 (Hierachical Data Format) aufsetzt. HDF5 ist ein Datenformat, welches vor allem in der Wissenschaft zur Speicherung großer Datenmengen genutzt wird. [9] Bei NetCDF-4 ist die Größe der Dateien nicht mehr beschränkt, es können somit deutlich größere Dateien erstellt werden. Außerdem ist es möglich, mehrere unlimitierte Dimensionen zu erstellen. Zusätzlich sind nun Gruppen möglich, wodurch die Arbeit mit großen Dateien überschaubarer wird. Die Vorteile von HDF5 werden also direkt genutzt. Es sind dennoch nicht alle Funktionen von HDF5 mit NetCDF kompatibel. [10]

[6] Simon Kostede: Performanceanalyse der Ein-/Ausgabe des Ökologiemodells ECOHAM5, Masterarbeit, Universität Hamburg, S.10, S. 18 ff [6]

[7] Environmental Systems Research Institute: "Was sind netCDF-Daten?", `http://desktop.arcgis.com/de/arcmap/10.3/manage-data/netcdf/what-is-netcdf-data.htm`, 04.03.2019 [8]

[8] Unidata : "An Introduction to NetCDF", `https://www.unidata.ucar.edu/software/netcdf/docs/netcdf_introduction.html`, 04.03.2019 [7]

[9] Wikipedia: "Hierachical Data Format", `https://de.wikipedia.org/wiki/Hierarchical_Data_Format`, 10.03.2019 [9]

[10] Simon Kostede : siehe Fußnote 6 [6]

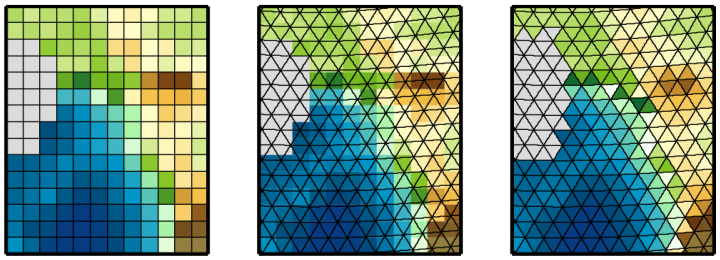

Figure 4.1: Beispiel einer Interpolation
(Quelle: Uwe Schulzweida: "CDO - Project Management Service",
`https://code.mpimet.mpg.de/projects/cdo/embedded/index.html#x1-5870002.12.1`, 11.03.2019)

4 Bearbeiten der Daten mit CDOs

CDOs (Climate Data Operators) stellen eine Sammlung von Werkzeugen zur Bearbeitung und zum Auslesen von wissenschaftlichen Daten dar. Informationen über Inhalte können aufgerufen werden, Daten können nachträglich modifiziert werden und diese lassen sich schließlich auch auswerten.
Mit Hilfe von `cdo showname infile` kann man sich die Namen aller Variablen einer Datei anzeigen lassen. Es ist außerdem möglich eine Grid description zu erhalten: `cdo griddes infile`. Diese zeigt einem diverse Information, wie die Namen der Achsen und die Größe, vom Koordinatensystem an.
Auch das Ändern der Variablennamen ist möglich : `cdo chaname, temp, t2m infile outfile`, hier wird der Name von temp zu t2m geändert. Eine neue Zeitachse wird mit `cdo settaxis,1987-01-16,12:00:00,1mon infile outfile` definiert, welche am 16.Januar 1987 um 12 Uhr beginnt und in 1 Monat Schritten wächst. Ein sehr wichtiger Aspekt ist die Interpolation. Dabei wird ein Gitter in ein anderes Gitter überführt (siehe Figure 4.1). Durchgeführt wird dies mit beispielsweise `cdo remapbil,n32 infile outfile`.
Zum Auswerten einer Datei gibt es viele verschiedene Möglichkeiten. Möchte man beispielsweise den Durchschnittswert pro Stunde einer Datei berechnen, so genügt der Befehl `cdo hourmean infile outfile`. Dies ist auch für mehrere Dateien gleichzeitig möglich: `cdo ensmean infile[1-6] outfile`. Hier wird der Durchschnitt aller Variablen berechnet. Es ist jedoch darauf zu achten, dass die 6 Inputfiles von der Struktur identisch sind. Mit `cdo enspctl, 50 infile[1-6] outfile` lässt sich beispielsweise auch der Median mehrerer Dateien berechnen. [11]

[11] Uwe Schulzweida: "CDO User's Guide", März 2013, S. 22 f., S. 48 f., S. 50, S.131 f., S. 80, S. 96 [10]

5 Visualisierung der Daten

Ähnlich wie beim Bearbeiten der Daten, gibt es auch diverse Softwaretools, die es einem ermöglichen, Daten visuell darzustellen. Ein Beispiel dafür wäre NCView. NCView bietet eine Steuerungskonsole, in der diverse Einstellungen vorgenommen werden können (siehe Figure 5.1). Handelt es sich beispielsweise um Daten in einem Zeitverlauf, so kann man jeden Tag einzeln darstellen oder einen Film abspielen lassen, welcher die fortlaufenden Tage in einer Animation abspielt. Nehmen wir an, bei unseren Daten handelt es sich um den Sauerstoffgehalt im Wasser im Bereich der Nordsee. Unsere Variable wäre somit der Sauerstoffgehalt. Diese wird auf einer Karte der Nordsee dargestellt. Wir können mit NCView nun die Farbskala für die Darstellung bestimmen. Zusätzlich können wir auch die Tiefen einstellen, in welcher wir den Sauerstoff anzeigen lassen wollen. Es ist außerdem möglich, einen bestimmten Punkt auf der Karte auszuwählen. Wir erhalten dann eine graphische Darstellung des zeitlichen Verlaufs einer Variable über die Zeit an einem Gitterpunkt (siehe Figure 5.2). Durch diese Möglichkeit der visuellen Darstellung ist es möglich, neue Schlüsse aus den Daten zu ziehen. Es ist wesentlich einfacher, die Daten zu verstehen, wenn diese einem visuell vorliegen und durchlaufen werden können.

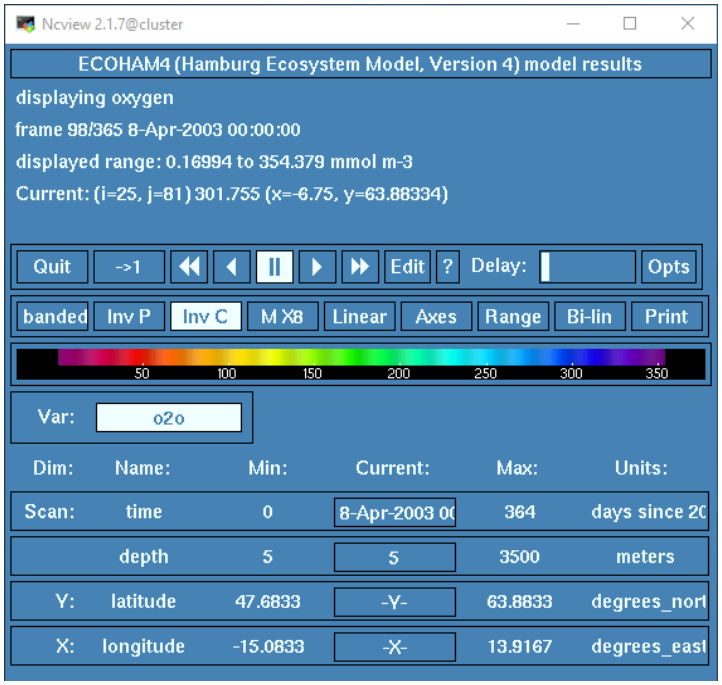

Figure 5.1: NCView Steuerungskonsole
(Quelle: H. Lenhart pers. comm.)

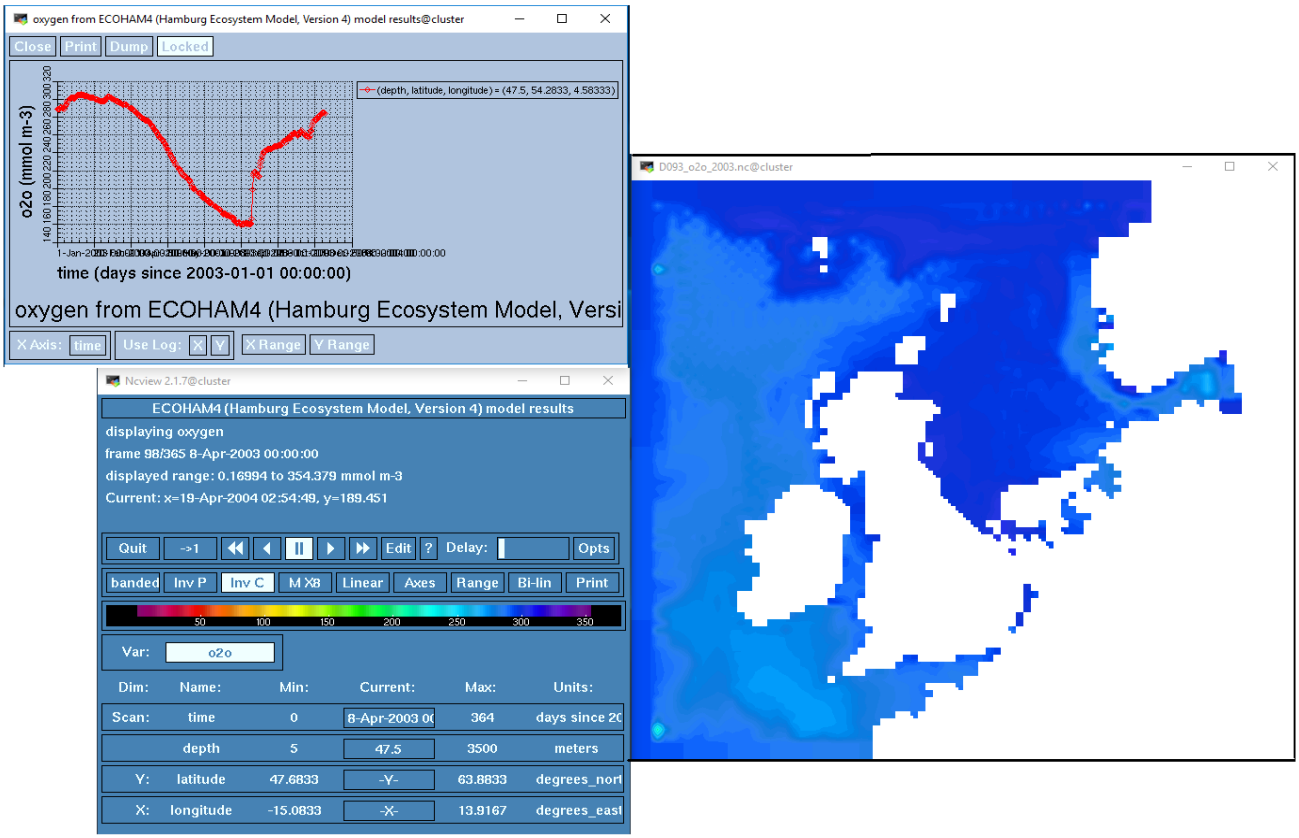

Figure 5.2: Darstellung von Daten in NCView
(Quelle: H. Lenhart pers. comm.)

6 Zusammenfassung

Der Grund dafür, dass es so viele Datenformate in der Wissenschaft gibt, sind die verschiedenen Anforderungen aus den Bereichen der Wissenschaft. Für jeden Bereich sind andere Eigenschaften von Bedeutung und diese müssen vom Datenformat gegeben sein. Ein Datenformat entsteht also aus den Anforderungen. Es gibt jedoch Voraussetzungen, welche für alle Datenformate gelten und somit auch vorhanden sein sollten. NetCDF ist ein weit verbreitetes Datenformat aus dem Bereich der Klimatologie. Es setzt sich aus Dimensionen, Variablen und Attributen zusammen. NetCDF-4 war dabei eine besondere Version, welche HDF5 implementiert. Möchte man seine Daten bearbeiten, so bieten die CDOs eine vorgefertigte Sammlung von Werkzeugen. Auch eine visuelle Darstellung ist durch Tools, wie NCView, möglich. Diese ermöglichen einem, die Daten besser zu verstehen.

Literaturverzeichnis

[1] Andreas Pfund: "Metadaten", http://andreas-pfund.de/definition/metadaten/metadaten.php, 09.03.2019

[2] Meelis Kull: "Scientific Data Formats", http://kodu.ut.ee/~varmo/tday-kaariku/kull-slides.pdf, 03.01.2019

[3] Wikipedia: "List of file formats", https://en.wikipedia.org/wiki/List_of_file_formats#Scientific_data_(data_exchange), 10.03.2019

[4] Wikipedia: "Protein Data Bank (file format)", https://en.wikipedia.org/wiki/Protein_Data_Bank_(file_format), 10.03.2019

[5] Wikipedia: "Digital Imaging and Communications in Medicine", https://de.wikipedia.org/wiki/Digital_Imaging_and_Communications_in_Medicine, 10.03.2019

[6] Simon Kostede: Performanceanalyse der Ein-/Ausgabe des Ökologiemodells ECO-HAM5, Masterarbeit, Universität Hamburg, S.10, S. 18 ff.

[7] Unidata : "An Introduction to NetCDF", https://www.unidata.ucar.edu/software/netcdf/docs/netcdf_introduction.html, 04.03.2019

[8] Environmental Systems Research Institute: "Was sind netCDF-Daten?", http://desktop.arcgis.com/de/arcmap/10.3/manage-data/netcdf/what-is-netcdf-data.htm, 04.03.2019

[9] Wikipedia: "Hierachical Data Format", https://de.wikipedia.org/wiki/Hierarchical_Data_Format, 10.03.2019

[10] Uwe Schulzweida: "CDO User's Guide", März 2013, S. 22 f., S. 48 f., S. 50, S.131 f., S. 80, S. 96