Hannes Thiemann, Stephan Kindermann, Michael Lautenschlager

5.5 Beispiele für Data Sharing am Deutschen Klimarechenzentrum (DKRZ)

Abstract: Über den Erzeugerkreis hinaus werden Klimamodelldaten für unterschiedliche Zielgruppen bereitgestellt. Die Zielgruppen besitzen spezifische Anforderungen an Struktur, Qualität und Dokumentation der Daten. Diese Anforderungen werden in den verschiedenen Data Sharing Angeboten des DKRZ berücksichtigt, wobei die FAIR-Prinzipien bei deren Gestaltung eine Richtschnur bilden. Spezielle Herausforderungen für die Bearbeitung von Klimamodelldaten ergeben sich nicht nur aus den sehr großen Datenmengen, sondern auch der Vielzahl der Datenentitäten und der interdisziplinären Nachnutzung. Insbesondere die immensen Datenvolumina, die oft in Datenanalyseaktivitäten benötigt werden, bergen spezifische Anforderungen bezüglich der Integration von Prozessierungsdiensten und Data-Sharing-Diensten. Das Vorgehensmodell und die Angebotspalette des DKRZ lässt sich auf andere Diensteanbieter bzw. Forschungsdisziplinen, abhängig von deren spezifischen Anforderungen und zur Verfügung stehenden Ressourcen übertragen.

Einleitung

Das Deutsche Klimarechenzentrum (DKRZ)¹ stellt als zentrale Service-Einrichtung der Klimaforschung in Deutschland Rechenleistung, Speicherkapazitäten und Dienste zur Verfügung, um Klimamodelle zu rechnen und Klimamodelldaten zu speichern und auszuwerten. Darüber hinaus unterstützt das DKRZ seine Nutzerinnen und Nutzer organisatorisch und infrastrukturell bei der globalen Verteilung ihrer am DKRZ oder extern produzierten Forschungsdaten. Dadurch wird ein erheblicher Mehrwert der Daten generiert, da diese erst dann für die gesamte Klima- und Erdsystemforschungsgemeinschaft effektiv nachnutzbar sind.

Die globale Verteilung relevanter Datensätze ist ein Grundbedarf der Erdsystemforschung. Insbesondere in den Klima- und Erdsystemwissenschaften werden großvolumige Datensätze bis in den PetaByte (PB)-Bereich von hohem globalem Interesse an vereinzelt Institutionen erzeugt und teilweise zur Nachnutzung aufbereitet. Hierbei handelt es sich z. B. um Beiträge zu international koordinierten Modellvergleichsstudien wie etwa Coupled Model Intercomparison Projects (CMIP)² oder um

1 S. <http://www.dkrz.de>. Letztes Abrufdatum der Internet-Dokumente ist der 15.11.2020.

2 S. <https://www.wcrp-climate.org/wgcm-cmip>; Eyring et al. 2016.

Daten, welche aus einer institutionellen Projekt-Initiative heraus entstanden sind, zum Beispiel aus dem Max-Planck-Institut Grand-Ensemble.³

Ein großer Mehrwert dieser Datensätze wird erst durch deren Nachnutzung geschaffen. Diese wird durch effektives Teilen der Daten mit der Forschungsgemeinschaft und der vorausgehenden Verfügbarmachung ermöglicht. Die damit verbundene lokale Verwaltung und globale Datenverteilung erfordert institutionalisierte Kompetenzen im Bereich Datenmanagement, -aufbereitung, -dokumentation und Infrastruktur.

Am DKRZ werden für das Data Sharing verschiedene, anwendungsangepasste Dienste angeboten. Je nachdem, ob DKRZ-Nutzende ihre Daten im Rahmen eines Projektes, von internationalen Modellvergleichsstudien oder im Zuge einer wissenschaftlichen Publikation global (ver-)teilen möchten, bieten die am DKRZ bereitgestellten Dienste passende Lösungen:

- DKRZ-Cloud (Swiftbrowser)⁴ – der Cloud-Service zum einfachen und direkten Teilen großer Datenmengen.
- DKRZ CMIP Data Pool⁵ – auf einem zentralen, effizient zugreifbaren Speicherbereich verwaltete Datenkollektionen, die von übergreifender Bedeutung für die Datenverteilung und Datenanalyse sind.
- Earth System Grid Federation (ESGF)⁶ – die globale Dateninfrastruktur zur föderierten Verteilung von Modellvergleichsstudien-Daten (CMIPs) oder projektbezogener Daten von internationalem Interesse, z. B. das MPI-M Grand Ensemble.
- Digitales Langzeitarchiv World Data Center Climate (LTA WDCC)⁷ – das zertifizierte Langzeitarchiv mit DataCite DOI Vergabe,⁸ Erfüllung der FAIR-Data Leitlinien⁹ und Indizierung in globalen Suchdiensten, z. B. Google Data Search,¹⁰ European Open Science Cloud (EOSC)/B2FIND,¹¹ World Data System (WDS).¹²

Die Datennutzenden des DKRZ können verschiedenen Gruppen zugeordnet werden, die über unterschiedliches Fachwissen zu Struktur, Bearbeitung und Aussagekraft von Klimamodelldaten verfügen. Grob lassen sich vier Klassen unterscheiden:

3 S. <https://www.mpimet.mpg.de/en/grand-ensemble>; Maher et al. 2019.

4 S. <https://www.dkrz.de/up/de-systems/de-swift>.

5 S. <https://www.dkrz.de/up/de-services/de-data-management/de-cmip-data-pool>.

6 S. <https://www.dkrz.de/up/de-services/de-data-management/de-esgf-services-1>.

7 S. <https://www.dkrz.de/dienste/langzeitarchivierung-1/langzeitarchivierung>.

8 S. <https://www.dkrz.de/up/services/data-distribution/data-publication>.

9 S. <https://www.dkrz.de/up/de-services/de-data-management/de-LTA/de-fairness>.

10 S. <https://toolbox.google.com/datasetsearch>.

11 S. <http://b2find.eudat.eu>.

12 S. <https://www.icsu-wds.org/services/data-portal>.

1. Wissenschaftliche Arbeitsgruppe (Erforschung von Prozessen im Klimasystem, Quantifizierung von Klimaänderungen)
 - Produzierende und Nutzende von Klimamodellergebnissen sind Teil der Arbeitsgruppe. Detaillierte Kenntnisse zu Forschungsgegenstand sowie Struktur und Inhalt der einzelnen Datenentitäten sind vorhanden.
2. Nationale und internationale Wissenschaftsdisziplin (Erforschung des Klimasystems)
 - Die Datennutzenden verfügen über detaillierte Kenntnisse des Forschungsgegenstands, während die Kenntnisse zu Struktur und Inhalt der Klimamodelldaten breiter gefächert sind. Informationen zur Bearbeitung der Datenentitäten werden meist benötigt.
3. Verwandte Wissenschaftsdisziplinen (Erforschung der Auswirkungen von Klimaänderungen oder der Vermeidung von Klimaänderungen)
 - Nutzende aus diesen Disziplinen verfügen meist über keine Detailkenntnisse der Klimamodellierung. Ausführliche Informationen zu Interpretation, Inhalt und Bearbeitungsmöglichkeiten von Klimamodelldaten sind erforderlich.
4. Politische und gesellschaftliche Entscheidungsträger der Klimapolitik (Ableitung und Umsetzung von Handlungsoptionen zur Begrenzung von und Anpassung an Klimaänderungen)
 - Nutzende aus diesem Bereich arbeiten in den seltensten Fällen direkt mit den Klimamodelldaten. Sie stützen sich auf die wissenschaftlichen Aussagen der Klimaforschung zur Ableitung von Handlungsoptionen.

Diese Anwendungsgebiete sind unterschiedlich weit entfernt vom Forschungsgegenstand Klimasystem, in dem die Daten erhoben bzw. erzeugt und zuerst genutzt werden. Den abnehmenden Kenntnissen zu Daten, Kontext und Bearbeitungsmethoden muss im Datenmanagement und den verschiedenen Ebenen des Data Sharing Rechnung getragen werden.

Zugriffe auf Daten und Metadaten sind unabhängig vom Dienst generell kostenfrei, erfordern jedoch, wenn in den folgenden Kapiteln nicht anders erwähnt, eine (kostenfreie) Registrierung.

1 Die DKRZ-Cloud

Das DKRZ stellt seinen Nutzerinnen und Nutzern und den von ihnen vertretenen Forschungsprojekten eine Cloud-Umgebung in der Größenordnung von mehreren PB auf Basis von Openstack Object Storage (Swift)¹³ zur Verfügung. Swift ist sehr

¹³ S. <https://wiki.openstack.org/wiki/Swift>.

gut geeignet für die Speicherung und Verteilung unstrukturierter Daten. Die Open-stack-Software stellt sicher, dass Datenobjekte in mehreren Kopien gespeichert werden und bietet so einen rudimentären Schutz gegen Datenverlust. Der Zugriff auf die Daten ist von unterschiedlichsten Geräten, vom Smartphone, über Notebook bis, und das ist im Umfeld des DKRZ entscheidend, hin zum Supercomputer möglich. Für Nutzende noch ungewohnt ist, dass es sich bei Swift um ein sogenanntes Object System handelt, welches eine andere Datenhierarchie bietet als altbewährte POSIX-Dateisysteme,¹⁴ in denen Metadaten traditionell in Directory- und Filenamen kodiert werden.

Wie einleitend bemerkt, ist Swift insbesondere für die Verteilung unstrukturierter Daten geeignet. Als Beispiel für ein Projekt auf DKRZ-Swift sei hier das Radiative Convective Equilibrium Model Intercomparison Project (RCEMIP)¹⁵ erwähnt. RCEMIP ist eine Initiative, die vom World Climate Research Program (WCRP) angestoßen wurde. Vergleichbar zu anderen Model Intercomparison Projects (MIPs) nutzt die Forschung den Vergleich und die Analyse von Daten mehrerer verschiedener Modelle. Für RCEMIP von Vorteil ist die unkomplizierte Handhabung von Nutzeraccounts sowohl für den Datenupload als auch –download. Zugriffsrechte können nach initialer Einrichtung ohne weitere Rücksprache mit dem DKRZ von dem Projekt verwaltet werden. Auch in das eigentliche Datenhandling ist das DKRZ nicht involviert. RCEMIP verwendet die DKRZ-Cloud nach Einweisung selbstständig und greift auf Hilfe des DKRZ nur bei Schwierigkeiten zurück. Die Einweisung umfasst insbesondere die begrenzten Möglichkeiten zur Einrichtung von Datenhierarchien, welche herkömmlich von MIP-Projekten zur Datenorganisation verwendet werden. Gezwungenermaßen ergänzt RCEMIP das Datenmanagement mit einer ausführlichen Dokumentation zur Datenorganisation, die sowohl Providern als auch Nutzenden das Auffinden von Daten ermöglicht.

Die Cloud selbst ermöglicht im Sinne der FAIR-Prinzipien lediglich das „A“, die Accessibility. Alle anderen Aspekte müssen von Projekten gegenwärtig noch mit ergänzenden Dienstleistungen selbst geschaffen werden. Weitere Informationen hierzu unten in den Abschnitten 5 Datenanalyse bzw. Fazit.

2 Der DKRZ CMIP Datenpool

Das DKRZ beteiligt sich mit seinen Datenservices im Rahmen der Klimamodellierung an der internationalen Bereitstellung nationaler Ergebnisse und der Zusammenführung internationaler Beiträge. Im Hinblick auf eine nachfolgende effiziente

¹⁴ S. https://de.wikipedia.org/wiki/Portable_Operating_System_Interface.

¹⁵ S. <http://myweb.fsu.edu/awing/rcemip.html>.

Datenanalyse sind Homogenisierung und Qualitätskontrolle wichtig, um ein möglichst einheitliche Datenbasis bereitzustellen. Data Sharing ist im DKRZ-CMIP-Datenpool in zwei Richtungen von Bedeutung, zum Füllen und im Zugriff. Die ESGF Data-Sharing-Infrastruktur wird genutzt, um den Datenpool am DKRZ aufzubauen. Große Datenmengen werden von den anderen Datenknoten der ESGF Föderation heruntergeladen und am DKRZ zusammengeführt. Zum anderen wird Data Sharing genutzt für die DKRZ-interne, nationale und internationale Verteilung der Daten.

Das aktuell prominente Beispiel internationaler Klimamodellierung ist CMIP6,¹⁶ das Climate Model Intercomparison Project Generation 6. CMIP6 liefert nicht nur Informationen zur Erforschung des Klimasystems selbst, sondern bildet auch die Datengrundlage für Veröffentlichungen, die in dem nächsten Statusbericht des internationalen Klimarats (Intergovernmental Panel on Climate Change Sixth Assessment Report, IPCC-AR6¹⁷) eingehen. Aufbauend auf den Erfahrungen aus den CMIP5-Aktivitäten entwickelte sich in der deutschen Klimaforschungsgemeinschaft die Anforderung, die CMIP-Aktivitäten im CMIP6-Zyklus stärker zu koordinieren und daher u. a. ein gegenüber dem CMIP5-Prototyp weiter verbessertes CMIP-Datenarchiv zu etablieren. Unter Leitung des DKRZ und unter Beteiligung weiterer Einrichtungen unterstützte das BMBF die Koordinierung der CMIP6-Aktivitäten in Deutschland.

Im internationalen Rahmen fügen sich die deutschen Anstrengungen für die CMIP6-Dateninfrastruktur in die vom WGCM¹⁸ Infrastructure Panel (WIP) koordinierten Aktivitäten ein, deren Ziel es ist, im globalen Rahmen eine robuste und nachhaltige Dateninfrastruktur zur Unterstützung der wissenschaftlichen Ziele des WGCM zu etablieren. Wesentliches Element dieser Dateninfrastruktur ist die Earth System Grid Federation (s. Abschnitt 3). Das DKRZ betreibt einen zentralen Datenknoten in der internationalen Datenföderation ESGF.

¹⁶ S. <https://www.wcrp-climate.org/wgcm-cmip/wgcm-cmip6>.

¹⁷ S. <https://www.ipcc.ch/assessment-report/ar6/>.

¹⁸ Working Group on Climate Models.

Abb. 1: Der CMIP Data Life Cycle kann grob aufgeteilt werden in Produktionsphase, Datenanalyse in der ESGF (Project Phase) und die Bibliometrische Phase mit Langzeitarchivierung im WDCC am DKRZ und DataCite Datenpublikation. M1–M4 und D1–D3 bezeichnen die Prüfungen von Metadaten (MD) und Daten (D) in den Stufen des Life Cycle.

Das ESGF Datenarchiv speichert im November 2019 ca. fünf PB an CMIP6 Daten aus Berechnungen. Sie beinhalten die sogenannten Diagnostic, Evaluation and Characterization of Klima (DECK) Experimente, historische Simulationen und 23 CMIP6 „Endorsed MIPs“,¹⁹ die spezielle wissenschaftliche Fragestellungen betrachten. Derzeit liefern 65 Erdsystemmodelle CMIP6-Daten ins ESGF, die Deutsche Forschungsgemeinschaft (DFG) beteiligt sich mit vier Modellen. Für das vollständige CMIP6 Datenarchiv wird ein Volumen von über zehn PB plus erwartet.

Die großen Datenmengen für CMIP6 erfordern eine detaillierte Planung des Datenmanagements, um Datentransfer und doppelte Speicherung zu minimieren. Die Anforderungen der deutschen Klimaforschungsgemeinschaft an die Dateninfrastruktur im Rahmen von CMIP6 beinhalten Komponenten, von denen einige im Data Sharing am DKRZ adressiert werden.

1. Bereitstellung der Ergebnisse der deutschen Klimamodelle für die internationale Wissenschaftscommunity.
2. Schneller, direkter und komfortabler Zugriff auf die Ergebnisse der deutschen Klimamodelle für die deutsche Wissenschaftscommunity.
3. Schneller, direkter und frühzeitiger Zugriff auf relevante Ergebnisse der internationalen CMIP6-Klimamodellierungs-Community, wie sie für die deutschen Beiträge zum IPCC-AR6 Bericht benötigt werden.

¹⁹ S. <https://www.wcrp-climate.org/wgcm-cmip/wgcm-cmip6>.

4. Bereitstellung von Daten, die von der deutschen CMIP6 Community benötigt werden um die neu generierten CMIP6-Datenbestände auszuwerten und zu evaluieren (dies schließt CMIP5-Daten und Beobachtungsdaten mit ein).
5. Langfristige und sichere Verfügbarkeit der in 1.-4. bezeichneten Daten.

Die unter 1.-4. bezeichneten Daten werden im DKRZ-CMIP-Datenpool zusammengeführt. Neben den Klimamodelldaten aus CMIP6 enthält der DKRZ-CMIP-Datenpool weiteren Daten aus vergangenen Modellierungsprojekten wie CMIP5 (Vorgänger von CMIP6) und CODEX (regionale Klimamodellierung aufbauend auf CMIP5). Für alle Daten in diesem Pool gilt, dass diese zusammenhängend, einheitlich und zuverlässig zur Verfügung gestellt werden.

Abb. 2: Aufbau des CMIP6-Datenpools als Teil des DKRZ-CMIP-Datenpools

Die Gesamtheit der Komponenten, der CMIP-Datenpool, der ESGF Datenknoten, das Langzeitarchiv im World Data Center for Climate (s. Abschnitt 4) und die erforderlichen zusätzlichen Dienste wie etwa die Qualitätsprüfung und die DOI-Datenpublikation, bilden das nationale CMIP Datenarchiv, welches der nationalen Klimaforschung frühzeitig ein hochwertiges und umfangreiches Datenarchiv nicht nur mit den nationalen, sondern auch mit den wichtigsten internationalen Daten zentral zur Verfügung stellt. Die Langzeitarchivierung im zertifizierten WDCC garantiert die Nachhaltigkeit der Datenbasis.

Neben effektivem Datenzugriff und –transport ist für den Zugriff auch die Dokumentation der Daten für eine nutzerangepasste Datensuche erforderlich. Zunehmend stärker wird im Rahmen des Datenzugriffs auch Rechenkapazität am Archiv-

standort nachgefragt für eine Vorauswertung und Reduzierung der zu transferierenden Datenmenge (s. Abschnitt 5).

3 Die Earth System Grid Federation (ESGF)

Eine spezielle Herausforderung des Data Sharing stellen die international koordinierten Klimamodell-Vergleichsexperimente (Coupled Model Intercomparison Projects – CMIP) dar. Ergebnisse dieser Experimente werden einerseits von der Klimaforschungsgemeinschaft, andererseits auch intensiv in anderen Disziplinen wie der Klimafolgenforschung genutzt und fließen in die IPCC Sachstandsberichte ein (aktuell AR6). Neben dem Datenvolumen (z. B. > 20 PB CMIP6 Daten und > 4 PB CMIP5 Daten) ist hier besonders herausfordernd, dass Datenkollektionen von weltweit verteilten Datenproduzierenden (den Klimarechenzentren und Klimadatenzentren) einer weltweiten Forschungsgemeinschaft über einheitliche Such- und Zugriffsmechanismen zugänglich gemacht werden müssen.

Mit der Entwicklung der Earth System Grid Federation (ESGF), einem verteilten Peer-to-Peer-System zur Klimadatenverteilung, wurde auf diese Herausforderung zur Unterstützung von CMIP5 geantwortet. Eine Weiterentwicklung der ESGF Infrastruktur ist auch aktuell für CMIP6 im Produktionseinsatz.²⁰ Datenkollektionen werden hier in weltweit verteilten Datenknoten einheitlich abgelegt und über einen einheitlichen Bereitstellungsprozess der globalen Föderation bekannt gemacht (dies wird als „ESGF Datenpublikation“ bezeichnet). Der Bereitstellungsprozess schließt eine minimale Datenqualitätsprüfung mit ein und stellt die Metadaten in dedizierten „Index-Knoten“ bereit, die wiederum mit Portal-Diensten verknüpft sind (Login, facettierte Suche, shopping card, etc.). Dedizierte Daten-Knoten übernehmen hierbei Zusatzaufgaben wie z. B. die Replikation und Bereitstellung von Replikaten).

Das DKRZ ist Gründungsmitglied von ESGF und einer der weltweit größten ESGF Datenknoten. Aktuell sind am DKRZ ESGF Portal über 4 000 Nutzerinnen und Nutzer registriert und ca. 350 Nutzende nutzen das Portal pro Monat, um Daten zu suchen und Datendownloads anzustoßen. Aktuell (Stand 01.2020) werden ca. 70 000 getrennte Datendownloads pro Monat (korrespondierend zu ca. 100 TByte Datenvolumen) angestoßen, mit stark steigender Tendenz, da die aktuellen CMIP6-Daten erst teilweise im DKRZ-ESGF-Portal bereitgestellt wurden.

Neben der Aufgabe der Replikation von Daten und Replikat-Bereitstellung übernimmt das DKRZ die Zusatzaufgabe der Langzeitarchivierung von CMIP Datenkollektionen im WDCC des DKRZ. Hierzu wurde das DKRZ Langzeitarchiv als Datenknoten in die ESGF Föderation eingebunden und ist aktuell der einzige Knoten weltweit

²⁰ Vgl. Balaji 2018.

der ein Bandarchiv (mit assoziierter Metadatenbank) direkt in ESGF integriert. In einem weitgehend automatisierten Prozess werden die ESGF Metadaten mit zusätzlichen Archivierungsmetadaten verknüpft und in der WDCC Datenbank hinterlegt. Basierend auf diesen Metadaten wird dann ein „Filesystem in Userspace“ (FUSE) auf dem Langzeitarchiv simuliert, das es erlaubt die Datenbestände als ESGF-Datenknoten in der ESGF-Datenföderation suchbar und zugreifbar zu machen.

Eine vom DKRZ entwickelte Erweiterung der ESGF-Infrastruktur unterstützt zudem die Vergabe von persistenten Identifikatoren (PID, hier: Handle PID) als Teil des ESGF-Datenpublikationsprozesses. Die PID-Vergabe erfolgt auf File-Ebene und auf Kollektionsebene (Zeitreihe einer Variable), zudem wird die Datenversionierung und Datenreplikation in den PID-Metadaten mitverfolgt. PID bilden die Grundlage höherer Dienste, die das Data Sharing, z. B. die Zuordnung von Errata-Information für Daten oder die Möglichkeit der Bildung von nutzerdefinierten Datenkollektionen, wesentlich unterstützen. Auf technischer Ebene werden die PID-Registrierungsinformationen auf Basis einer verteilten ausfallsicheren Message-Queue an die zentrale Registrierungsstelle am DKRZ weitergeleitet.

Auf höherer Datenkollektionsebene (z. B. alle Daten eines konkreten Modell-Laufes) vergibt das DKRZ zudem Datacite DOI. Als Langzeitarchiv ist das DKRZ insbesondere an der Nachhaltigkeit der entwickelten PID Dienste interessiert. Auf operativer Ebene ist das DKRZ deshalb Teil des europäischen PID Consortiums (ePIC)²¹ und auch Mitglied des European Data (EUDAT)²² Konsortiums. Zudem engagiert sich das DKRZ in PID bezogenen Arbeitsgruppen der Resarch Data Alliance (RDA) und ist an der European Open Science Cloud (EOSC) beteiligt.

Grundvoraussetzung für die Datenbereitstellung im Rahmen der ESGF-Datenföderation ist die Einhaltung von Datenkonventionen (netcdf-CF²³), sowie spezifischen Modell-Vergleichsprojekt Festlegungen (siehe z. B. die CMIP6-Modelldatenvorgaben²⁴). Diese konstituieren zum einen kontrollierte Vokabulare, auf denen z. B. Metadata-Werte in den netcdf-CF files basieren, aber auch Auswahllisten in den facettierten Such-Schnittstellen in den Portalen. Zum anderen legen diese Namenskonventionen File-Namen fest, sowie Strukturvorgaben zur Abspeicherung auf Filesystemen.

²¹ S. <https://www.pidconsortium.net/>.

²² S. <https://eudat.eu/>.

²³ Climate and Forecast (CF) Conventions and Metadata, s. <http://cfconventions.org>.

²⁴ CMIP6 model output requirements, s. <https://pcmdi.llnl.gov/CMIP6/Guide/modelers.html>.

Bezug zum Datenpool

Der oben im Abschnitt 2 beschriebene DKRZ-CMIP-Datenpool ist wesentliche Grundlage der ESGF-Datenbereitstellung. Die in ihm enthaltenen CMIP5-, CMIP6- und CORDEX-Daten (CORDEX steht für COordinated Downscaling Experiment) bilden die Speicherbasis der DKRZ-ESGF-Datenknoten. Der Datenaufnahmeprozess erfolgt im Wesentlichen über zwei Wege: die Datenreplikation von anderen ESGF-Datenknoten weltweit, sowie die direkte Datenübernahme von Klimamodellierungszentren (deutschen und ausländischen). In jedem Falle erfolgt eine Datenqualitätsprüfung vor der „ESGF Datenpublikation“. Im Falle von Originaldaten mithilfe automatisierter, detaillierter Prüfung der Einhaltung der geforderten Standards und Konventionen (z. B. CF-Konventionen und CMIP6-Festlegungen) und im Falle der Replikate mithilfe einfacher Vollständigkeits-Prüfroutinen.

Da das Gesamtvolumen der CMIP6-Datenkollektionen (~20 PB) deutlich grösser ist als der zur Verfügung stehende DKRZ-Datenpool-Plattenplatz (~fünf PB) ist die Aufnahme (und ESGF-Publikation) von Replikaten zudem von den Bedürfnissen und Prioritäten der Nutzenden gesteuert. Wissenschaftlerinnen und Wissenschaftler, die für ihre Analysen spezifischen Bedarf von CMIP-Datenreplikaten im Onlinezugriff am DKRZ haben, können diese bei den DKRZ-Datenmanagern anmelden. Diese sorgen dann dafür, dass diese Anforderungen Priorität bei der Organisation der Datenreplikation bekommen.

Die Datenreplikation erfolgt prioritär über hoch performante Globus-Connect-Endpunkte,²⁵ die von zentralen ESGF-Datenknoten bereitgestellt werden. Eine weltweite Arbeitsgruppe, die mit den nationalen und internationalen Forschungsnetzbetreibern in engem Kontakt steht, sorgt für eine Überwachung und Optimierung der Übertragungswege. Im Allgemeinen wird folgende Strategie verfolgt, um die CMIP-Datenkollektionen möglichst schnell der internationalen Forschungsgemeinschaft bereitzustellen: Datenkollektionen von Datenknoten mit schlechter Netzanbindung werden möglichst schnell zu zentralen Datenknoten (z. B. auf den jeweiligen Kontinenten) repliziert, die dann als sogenannte „Tier1“-Datenknoten die Weiterverteilung über optimierte Übertragungswege übernehmen können. Die Datenreplikation gewährleistet damit zwei zentrale Aspekte des Data Sharings: schnelle Zugreifbarkeit und hohe Verfügbarkeit der Datenkollektionen. Ein weiterer zentraler Aspekt des Data Sharings, die Verfügbarmachung von hochvolumigen Datenkollektionen zum Zweck der effizienten Datenanalyse; wird in Abschnitt 5 beleuchtet.

²⁵ Globus research data management cloud, s. <http://docs.globus.org>.

4 Das World Data Center for Climate (WDCC)

Das World Data Center for Climate (WDCC)²⁶ ist ein vom DKRZ betriebenes Langzeitarchiv für großvolumige Simulationsdaten aus der Klimaforschung. Seit 2003 ist es reguläres Mitglied des World Data System of the International Science Council (WDS)²⁷, früher ICSU. Als solches unterstützt es die Vision des International Science (ISC)²⁸ durch die langfristige Bewahrung und des Angebots eines freien Zugangs zu qualitätsgesicherten wissenschaftlichen Daten aus dem Bereich der Klimaforschung für die Nutzung in der Klimaforschung und darüber hinaus.

The screenshot shows the WDCC website interface. At the top, there is a navigation bar with links: Home, FAIRness, Documentation, Utilities, Register, and Login. The main header features the WDC CLIMATE logo on the left and the DKRZ logo on the right. Below the header, a large banner displays the text "DKRZ long term archive" and "Find your data" with a search input field. The banner also includes a "CORE TRUST SEAL" logo. Below the banner, the "Services" section is visible, containing four columns: "News" (with links to DOI assigned for wind farm parameterization software and DOI assigned for MESSy Software), "Data access" (regarding FAIR principles and graphical web interface), "Data archiving" (describing LTA WDCC and LTA DOKU services), and "Data publication" (describing the publication process and DOI data). At the bottom, there is a footer with links for "Powered by CERA", "Terms of use", "WDCC service status", and "Member of WORLD DATA SYSTEM".

Abb. 3: Eingangsseite zum World Data Center for Climate (WDCC)

Verbunden mit der Mitgliedschaft im WDS ist eine Zertifizierung durch das Core-TrustSeal (CTS)²⁹. Technisch baut das WDCC auf unterschiedlichen Basisdiensten des DKRZ, wie z. B. dem High Performance Storage System (HPSS), das auch eine

²⁶ S. <https://cera-www.dkrz.de/WDCC/ui/cerasearch>.

²⁷ S. <https://www.icsu-wds.org>.

²⁸ S. <https://council.science>.

²⁹ S. <https://www.coretrustseal.org>.

geographisch verteilte Datenhaltung ermöglicht, oder einer Oracle-Datenbank als Persistenzschicht auf. Eingebettet ist es in etablierte Prozessketten zur Datenübernahme, Qualitätssicherung und Kuration.

Mit seiner Positionierung als Langzeitarchiv erfüllt das WDCC neben der Unterstützung der guten wissenschaftlichen Praxis mehrere Anforderungen der Wissenschaft im Bereich des Data Sharings: zum einen die Nachnutzung der Daten über einen Projektzeitraum hinaus und zum anderen die Nachnutzung der Daten im interdisziplinären Kontext. Hiermit unterscheidet sich das Profil des WDCC signifikant von den anderen Data-Sharing-Angeboten des DKRZ. Durch die Verlagerung der Nutzung in sowohl zeitlicher als auch „disziplinärer“ Hinsicht aus dem Projekt- bzw. Community-Kontext hinaus, steigen die Anforderungen an die Aufbereitung und Prüfung der Daten und Metadaten sowohl bei Datenproduzierenden als auch beim DKRZ als Betreiber. Können bei Nachnutzung im Projekt noch gewisse Kenntnisse über die Daten und ihre Erzeugung vorausgesetzt werden, so nehmen diese mit steigender „Entfernung“ vom Projekt ab. Diese Distanz muss bei der Verfügbarmachung der Daten mitberücksichtigt werden. Durch die Zuweisung von DataCite DOIs schafft das DKRZ für geprüfte Daten die Voraussetzungen zur fachgerechten Zitation und Förderung des wissenschaftlichen Records.

Auch die Auffindbarkeit der Daten selbst muss sich anderen Anforderungen stellen. Kann im Projektrahmen noch die Bekanntheit eines Repositoriums vorausgesetzt werden, so ist dies bei der Langzeitnutzung nicht mehr gegeben. Diese Anforderung nimmt das WDCC auf, indem es die Metadaten der qualitätsgeprüften und zitierbaren Datensätze gemäß ISO 19115 und anderer akzeptierter Standards über Schnittstellen wie OAI-PMH³⁰ anbietet und damit z. B. im Rahmen der EOSC über B2FIND,³¹ WIS,³² GEOS, ³³ WDS³⁴ oder Google Dataset Search³⁵ auffindbar macht.

Als Langzeitarchiv mit Datenmengen im PB-Bereich steht das WDCC vor unterschiedlichen Herausforderungen: Die Daten werden mit unterschiedlicher Intensität genutzt und Festplattenspeicher sind nicht nur in der Anschaffung, sondern auch im Betrieb kostenintensiv. Daher werden die Daten auf Magnetbändern gespeichert und nur auf Anfrage der Nutzenden hin in einen Cache geladen, in dem lediglich die meistgenutzten Daten vorgehalten werden. Bedingt durch die Speicherung auf Magnetband sind direkte Downloads bzw. Zugriffe auf die Daten nicht möglich und deren unmittelbare Auswertung wird, wie im folgenden Abschnitt beschrieben, erschwert. Jedoch hat das WDCC in den Download ergänzende Dienste integriert, die darauf abzielen, die an die Benutzenden zu transferierende Datenmenge zu reduzie-

³⁰ S. <https://www.openarchives.org/pmh>.

³¹ S. <http://b2find.eudat.eu>.

³² WIS-WMO, s. <https://www.wmo.int/pages/prog/www/WIS/overview.html>.

³³ S. <https://www.geoportal.org>.

³⁴ ICSU-WDS Data Portal, s. <https://www.icsu-wds.org/services/data-portal>.

³⁵ S. <https://toolbox.google.com/datasetsearch>.

ren. Hier sind im Wesentlichen die Dienste zum Ausschneiden von Daten in Raum und Zeit zu erwähnen.

Die allermeisten vom WDCC angebotenen Daten stehen unter der Lizenz CC-BY,³⁶ es können jedoch nach Anforderung der Datenproduzierenden auch andere Lizenzen oder auch Embargoregeln implementiert werden. Das WDCC unterstützt den Open-Science-Gedanken und wirkt bei den Datenproduzierenden auf die Umsetzung hin.

Als zertifiziertes Langzeitarchiv beherbergt das WDCC Domain- oder Projektarchive, wie z. B. IPCC-DDC³⁷ der CoastDat,³⁸ und bietet diesen somit eine langfristige Heimat. Es ermöglicht diesen ein Data-Sharing auf lange Sicht und trägt damit zu deren verlässlichen Einbindung in die Forschungslandschaft bei.

Abb. 4: Zahl der Datendownloads aus dem WDCC

5 Datenanalyse

In der Vergangenheit wurden die am DKRZ über Data-Sharing-Dienste (Swift, Datenpool, ESGF, WDCC) bereitgestellten Daten oft an den Heimatinstituten der Forschenden analysiert. Diese bisherige Datenanalysepraxis des „Herunterladens und Zu-

³⁶ Für Creative-Commons-Lizenzen (CC-Lizenzen) s. <https://creativecommons.org>.

³⁷ IPCC Data Distribution Center, s. <https://www.ipcc-data.org>.

³⁸ S. <https://www.coastdat.de>.

hausa Verarbeitens“ ist jedoch angesichts der auflaufenden Datenvolumina nicht mehr gangbar.

Die Bereitstellung und Verwaltung eines zentralisierten Datenpools erfüllt somit auch die wichtige Aufgabe eine effiziente Datenauswertung am DKRZ zu unterstützen. Wichtige Klimadaten-Kollektionen, die für Analysen benötigt werden, werden in Abstimmung mit dem DKRZ-Datenmanagement lokal im Datenpool zusammengeführt und sind dort effizient durch die eng angebundenen Rechenressourcen nutzbar. Die Bereitstellung von Diensten, die Forschende dabei unterstützen, ihre Analysen dort auszuführen, wo die Datenkollektionen gesammelt und verwaltet werden, ist aktuell eine zentrale Herausforderung der sich das DKRZ stellen muss. Im Folgenden wird eine kurze Übersicht der verschiedenen Dienste gegeben, die zurzeit am DKRZ hierfür bereitgestellt werden:

- *Interaktive Datenanalyse*: Wissenschaftlerinnen und Wissenschaftler können direkt (z. B. über secure shell sessions) über dedizierte Analyseknotten oder indirekt über Jupyter Notebooks³⁹ auf dem Datenpool arbeiten. Die Notebook-Umgebung wird in einer jupyterhub-Installation⁴⁰ bereitgestellt und erlaubt es HPC-Ressourcen für Analyseaktivitäten zu reservieren und exklusiv zu nutzen.
- *Dienstehosting*: Wissenschaftlerinnen und Wissenschaftler bzw. Projekte haben oft die Anforderung bestimmte Mehrwert-Dienste bereitzustellen, die auf der Datenpool-Datenbasis aufbauen (z. B. Onlinebereitstellung von offline regelmäßig generierten Analyseresultaten). Hierzu betreibt das DKRZ eine Openstack-Installation, die es ermöglicht, schnell dedizierte virtuelle Maschinen bereitzustellen, die dann projektspezifisch konfiguriert werden können.⁴¹
- *Webservices*: Oft genutzte Basis-Prozessierungsdienste (wie z. B. Datenvolumen reduzierende Ausschneideoperationen oder Basis-Statistik-Operationen) sind sinnvolle Kandidaten, über wohldefinierte Interfaces als Webservices bereitgestellt zu werden. Diese können direkt von Forschenden angesprochen werden oder (was ein häufigerer Use Case ist) über dedizierte Portale und Plattformen, die spezifische Nutzengruppen unterstützen (siehe z. B. das europäische Klima-Folgeforschungsportal⁴² oder die Anbindung von Klimamodelldaten an den Copernicus Data Store). Zur Unterstützung dieser Anforderung initiierte das DKRZ das „Birdhouse“-Open-Source-Projekt⁴³ zur Bereitstellung von OGC-WPS-Standard-konformen Prozessierungsdiensten.

³⁹ Jupyter Interactive Computing Services, s. <http://jupyter.org>.

⁴⁰ JupyterHub Computational Environment, s. <http://jupyter.org/hub>.

⁴¹ S. z. B. <https://cmip-esmvaltool.dkrz.de>, wo aktuelle CMIP6-Modell-Diagnostik-Resultate basierend auf ESMValTool veröffentlicht werden; s. a. Eyring 2019.

⁴² IS-ENES Climate4Impact Portal, s. <https://climate4impact.eu>.

⁴³ S. <https://birdhouse.readthedocs.io/en/latest>; s. a. Ehbrecht et al. 2018.

Wegen der engen internationalen Verflechtung von Klimadaten-Analyseaktivitäten (z. B. durch die IPCC-Arbeitsgruppen für die IPCC-Sachstandsberichte) ist auch in Bezug auf die Prozessierungsdienste eine enge Absprache und Anpassung der Dienstangebote an den verschiedenen Klimadatenzentren erforderlich. Auf europäischer Ebene erfolgt dies aktuell im Rahmen des IS-ENES3-H2020-Projektes, das Datenprozessierungsdienste an zentralen Klimadatenzentren in Europa (DKRZ/Deutschland, Institut Pierre Simon Laplace – IPSL/Frankreich, Centre for Environmental Data Analysis – CEDA/England und Centro Euro-Mediterraneo sui Cambiamenti Climatici – CMCC/Italien) etabliert und abgestimmt bereitstellt. Zudem werden aktuell Prozessierungsdienste etabliert, die an den COPERNICUS Climate Data Store (CDS) angebunden werden sollen. Auf internationaler Ebene wurde eine ESGF-Arbeitsgruppe etabliert, um Basis-Prozessierungsdienste an ESGF-Datenknoten zu definieren.

6 Die FAIR-Prinzipien

Die FAIR-Prinzipien⁴⁴ definieren Grundsätze für das nachhaltige Forschungsdatenmanagement, auf die Forschende zunehmend nicht nur von ihren jeweiligen Forschungsinstitutionen, sondern auch von Geldgebern verpflichtet werden. Als Infrastrukturprovider für die Erdsystemforschung unterstützt bzw. ergänzt das DKRZ mit seinen Dienstleistungen die Anstrengungen der Forschenden in dieser Hinsicht in unterschiedlichem Maße:

- Die DKRZ-Cloud unterstützt lediglich die Accessibility der Forschungsdaten.
- Durch die vielfältigen und international abgestimmten Vorarbeiten zur Datenauswahl und -standardisierung erfüllen die Daten im CMIP-Datenpool insbesondere die Bereiche „I“ und „R“, während die Bereiche „F“ und „A“ weniger stark ausgeprägt sind. Dies wird jedoch kompensiert durch die auf dem Pool aufbauenden weiteren Dienstleistungen.
- Die im internationalen Rahmen etablierte technologische Lösung des ESGF (insbesondere die durch das DKRZ initiierte PID-Vergabe für CMIP6-Daten als Teil des ESGF-Datenpublikationsprozesses), ergänzt durch die enge Interaktion mit den Wissenschaftlerinnen und Wissenschaftlern, schafft gute Voraussetzungen für die Erfüllung aller Bereiche von FAIR.
- Auch die Langzeitarchivierung im WDCC erfüllt in hohem Maße die FAIR-Richtlinien (im Besonderen erreicht durch die intensive Datenkuration und Zuweisung von DataCite DOIs für angemessene Datengranularitäten). Wie auch beim ESGF demonstrieren die Zugriffszahlen die starke Nachnutzung der Daten. Das

⁴⁴ Vgl. Wilkinson et al. 2016.

WDCC geht jedoch deutlich über FAIR hinaus, indem es die nachhaltige Nutzung der Daten ermöglicht.

Bedingt durch die hohen und weiter steigenden Datenvolumina ist die Ausgestaltung der DKRZ-Analysedienste nach den FAIR-Prinzipien von großer Bedeutung. Sowohl die Netzwerk-, die Speicherplatz- als auch die CPU-Anforderungen für die Datenanalyse können von vielen Nutzenden nicht mehr an ihren Heimatinstitutionen erfüllt werden. Daher steigen die Anfragen für die sogenannte Server-Side-Analyse.

Während die Tragweite der FAIR-Prinzipien für den Datenlebenszyklus mittlerweile gut verstanden ist und auch umgesetzt wird, ist dieses für den Bereich der Analysedienste jedoch noch nicht der Fall und muss in den kommenden Jahren weiterverfolgt werden.

Fazit

Für die unterschiedlichen Gruppen von Datennutzenden am DKRZ werden verschiedene, angepasste Data-Sharing-Dienste zur Verfügung gestellt,⁴⁵ deren Schaffung, Weiterentwicklung und Betrieb auf einer kontinuierlich durchgeführten Analyse der Nutzer- und Datenerzeugeranforderungen beruht und Entwicklungen außerhalb des DKRZ und der Klimaforschung berücksichtigt. Für die Arbeitsgruppen am DKRZ stehen die Festplatten auf dem Großrechner Mistral, die DKRZ Cloud und der CMIP-Datenpool mit ESGF zur Verfügung. Die Fachwissenschaftlerinnen und -wissenschaftler aus der Klimaforschung greifen eher auf Klimamodelldaten aus der DKRZ Cloud, dem CMIP-Datenpool mit ESGF und dem Langzeitdatenarchiv WDCC zu. Wissenschaftlerinnen und Wissenschaftler aus verwandten Fachdisziplinen entfernen sich weiter von der Hochleistungsrechnerumgebung des DKRZ und greifen eher auf Daten aus dem CMIP-Datenpool mit ESGF und dem WDCC zu. Noch weiter entfernt von der Rechnerumgebung des DKRZ und der Klimamodellierung sind Wissenschaftlerinnen und Wissenschaftler aus anderen Fachdisziplinen, Politikerinnen bzw. Politiker und Entscheidungsträgerinnen bzw. -träger. Diese Nutzendengruppe ist auf Informationsmaterial des DKRZ angewiesen⁴⁶ und nutzt für den Datenzugriff vornehmlich das WDCC am DKRZ.

Das DKRZ verfolgt in seinen Data-Sharing-Diensten das Ziel, die FAIR-Datenmanagementprinzipien (Findable, Accessible, Interoperable, Reusable)⁴⁷ und Open Access für den Datenzugriff umzusetzen. In der DKRZ Cloud und im CMIP-Datenpool sind Abstriche im Bereich „Reusable“ zu verzeichnen, da die Informationen zu Da-

⁴⁵ S. <https://www.dkrz.de/dienste/datenverteilung>.

⁴⁶ S. <https://www.dkrz.de/kommunikation/pub>.

⁴⁷ Vgl. Wilkinson et al. 2016.

teninterpretation und -bearbeitung nicht vollständig sind. Hier werden Kenntnisse zu Klimaforschung und Klimamodellierung vorausgesetzt. Die FAIR-Prinzipien sind weitgehend im WDCC umgesetzt. Die beschreibenden Metadaten zielen in ihrem Umfang auf Vollständigkeit für Datennachnutzung unabhängig vom Datenproduzierendem.

Herausforderungen bestehen insbesondere in den Bereichen Datenorganisation, Datenanalyse, Standards und Datendokumentation. In der Klimamodellierung wachsen sowohl die Datenmenge als auch die Anzahl der Datenentitäten. Klassische Filesysteme stoßen an ihre Grenzen und sind nicht flexibel genug, um unterschiedliche Datentopologien darzustellen. Lösungsansätze bieten hier die PID mit ihrem globalen Auflösungsmechanismus⁴⁸ und eine darauf aufbauende DOA (Digital Object Architecture).⁴⁹

Große Datenmengen in der Klimamodellierung, spezifische Datenformate und die Datennutzung in vielen Wissenschaftsdisziplinen sowie gesellschaftlichen Bereichen stellen Herausforderungen an transparente Datenbearbeitung und die Auswertung in einer globalen Datenföderation (Datenanalyse). Angefragt ist nicht der Zugriff auf Originaldatenentitäten, sondern auf die „on the fly“-Auswertung von klimatologischen Fragestellungen (z. B. Tropentage oder Starkniederschläge in einer Region).

Eine weitere Herausforderung in der Analyse großer Datenmengen, wie sie in der Klimamodellierung auftreten, ist die automatisierte Datenbearbeitung eingebunden in die Data-Sharing-Infrastruktur. Definition und Einhaltung von Standards für Daten und Metadaten sind für eine reibungsfreie, automatisierte Datenbearbeitung notwendig. Wesentliche Entwicklungen liegen hier nicht in der technischen Implementierung, sondern in der Vereinheitlichung der Standards und ihrer Anwendung in den Dateninfrastrukturen, für die das DKRZ mit vielen Partnern, z. B. in nationalen Infrastrukturen wie der Nationalen Forschungsdateninfrastruktur (NFDI)⁵⁰ oder der European Open Science Cloud (EOSC),⁵¹ zusammenarbeitet.

Auch die Datendokumentation gehört nach wie vor zu den Herausforderungen im Management wissenschaftlicher Daten. Der direkte Kontakt zum Datenproduzierenden verliert sich immer mehr, je weiter sich die Nutzergruppe vom Forschungsgegenstand Klimamodellierung entfernt: Klimaforschung – Klimafolgenforschung – Politik und Öffentlichkeit. Umso wichtiger ist eine vollständige, einfach zugängliche Dokumentation der Klimamodelldaten, die eine Beurteilung der Daten auf Basis des eigenen Kenntnisstands erlaubt. Forschungsgegenstand ist insbesondere ein objektiver und akzeptierter Qualitätsstandard für Klimadaten. Ansätze finden sich in den

⁴⁸ Handle Proxy Server, s. <http://hdl.handle.net>.

⁴⁹ Vgl. Weigel et al. 2013; Weigel et al. 2014.

⁵⁰ S. <https://www.bmbf.de/de/nationale-forschungsdateninfrastruktur-8299.html>.

⁵¹ S. <https://www.eosc-portal.eu>.

Entwicklungen und Anwendungen der Data Maturity Matrix.⁵² Auch die Bereitstellung von Provenance-Informationen gehört zu den Herausforderungen im Bereich der Datendokumentation. Transparenz der Bearbeitungsschritte und Entwicklungshistorie von wissenschaftlichen Daten hilft bei der Beurteilung und stärkt das Vertrauen.

Literatur

Letztes Abrufdatum der Internet-Dokumente ist der 15.11.2020.

- Eyring, Veronika et al. 2019. „ESMValTool v2.0 – Extended set of large-scale diagnostics for quasi-operational and comprehensive evaluation of Earth system models in CMIP.“ *Geosci. Model Dev. Discuss.* (in review). doi:10.5194/gmd-2019-291.
- Eyring, V., S. Bony, G. A. Meehl, C. A. Senior, B. Stevens, R. J. Stouffer und K. E. Taylor. 2016. „Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization.“ *Geosci. Model Dev.* 9: 1937–1958. doi:10.5194/gmd-9-1937-2016.
- Balaji, V., K. E. Taylor, M. Juckes, B. N. Lawrence, P. J. Durack, M. Lautenschlager, C. Blanton, L. Cinquini, S. Denvil, M. Elkinington, F. Guglielmo, E. Guilyardi, D. Hassell, S. Kharin, S. Kindermann, S. Nikonov, A. Radhakrishnan, M. Stockhause, T. Weigel, und D. Williams. 2018. „Requirements for a global data infrastructure in support of CMIP6.“ *Geosci. Model Dev.* 11: 3659–3680. doi:10.5194/gmd-11-3659-2018.
- Ehbrecht, Carsten, Tom Landry, Nils Hempelmann, David Huard und Stephan Kindermann. 2018. „Projects based on the web processing service framework birdhouse.“ *ISPRS – International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XLII-4/W8*: 43–47. doi:10.5194/isprs-archives-XLII-4-W8-43-2018.
- Maher, Nicola, Sebastian Milinski, Laura Suarez-Gutierrez, Michael Botzet, Mikhail Dobrynin, Luis Kornblueh, Jürgen Kröger, Yohei Takano, Rohit Ghosh, Christopher Hedemann, Chao Li, Hongmei Li, Elisa Manzini, Dirk Notz, Dian Putrasahan, Lena Boysen, Martin Claussen, Tatiana Ilyina, Dirk Olonscheck, Thomas Raddatz, Bjorn Stevens und Jochem Marotzke. 2019. „The Max Planck Institute Grand Ensemble: Enabling the Exploration of Climate System Variability.“ *JAMES – Journal of Advances in Modeling Earth Systems* 11 (7): 2050–2069. doi:10.1029/2019MS001639.
- Weigel, T., M. Lautenschlager, F. Toussaint und S. Kindermann. 2013. „A Framework for Extended Persistent Identification of Scientific Assets.“ *Data Science Journal* 12: 10–22. doi:10.2481/dsj.12-036.
- Weigel, T., S. Kindermann und M. Lautenschlager. 2014. „Actionable Persistent Identifier Collections.“ *Data Science Journal* 12: 191–206. doi:10.2481/dsj.12-058.
- Wilkinson, M., M. Dumontier, I. Aalbersberg et al. „The FAIR Guiding Principles for scientific data management and stewardship.“ *Sci Data* 3: 160018. doi:10.1038/sdata.2016.18.

52 S. https://doi.org/10.2312/WDCC/TR_QMM_Checklist.