# Projekt RADIESCHEN

**Ra**hmenbedingungen einer **di**sziplinübergr**e**ifenden
For**sch**ungsdat**en**infrastruktur

## Report „Organisation und Struktur“

Jochen Klar, Harry Enke

Gefördert von:
DFG Deutsche Forschungsgemeinschaft

# Inhaltsverzeichnis

# Einleitung

Als Folge der anhaltenden Diskussion über den Umgang mit Forschungsdaten, entwickelte sich in den letzten zehn Jahren eine Vielzahl von nationalen und internationalen Projekten. Diese unterscheiden sich zum Teil erheblich in den gewählten Ansätzen und Strategien. Ziel des Arbeitspaketes 3 ist eine Analyse einer Auswahl dieser Projekte und eine darauf aufbauende Untersuchung, inwieweit die bereits vorhandenen oder im Entstehen begriffenen Strukturen weiterentwickelt werden können, um eine Etablierung des nachhaltigen Forschungsdatenmanagements zu erreichen.

Zunächst ist es also nötig, eine Bestandsaufnahme der bereits existierenden Organisationsstrukturen zu erstellen. Als Grundlage dient hierbei eine Zusammenstellung von verschiedenen durch die Deutschen Forschungsgemeinschaft (DFG) geförderten Projekten. Obwohl europäische und internationale Kontexte selbstverständlich großen Einfluss auf die Forschungslandschaft in Deutschland haben, dient eine solche Beschränkung dazu den Besonderheiten der deutschen Forschungslandschaft bezüglich des Forschungsdatenmanagements besser gerecht zu werden. Oben genannte Kontexte müssen jedoch immer als Randbedingungen mitgeführt werden.

Das Thema des Forschungsdatenmanagements ist in den letzten Jahren auf verschiedenen Ebenen diskutiert worden und es sind diverse Konzepte zur Umsetzung vorgeschlagen worden. Dies umfasst beispielsweise die Handlungsempfehlungen *Kommission Zukunft der Informationsinfrastruktur (KII)* der Leibniz-Gemeinschaft[1] und die *Schwerpunktinitiative „Digitale Information“* der Allianz der deutschen Wissenschaftsorganisationen[2]. Auch in den Forschungsorganisationen haben sich Arbeitskreise und Kommissionen zu Themen in diesem Umfeld gebildet. Es ist jedoch festzustellen, das noch keine generelle Umsetzung soweit gediehen ist, dass konkrete Resultate vorliegen. Da auch Fördermittel in verschiedener Form an Projekte in diesem Bereich vergeben worden sind, kann die Frage gestellt werden, ob die beabsichtigten Resultate erreicht wurden. Insbesondere ist es interessant, inwieweit diese Prozesse die Gesamtsituation des Forschungsdatenmanagements geändert haben und in einem vom unmittelbaren Projektzusammenhang unabhängigem Sinn das Forschungsdatenmanagement für die jeweilige Community (oder sogar darüber hinaus) verbessert haben. Es stellen sich weiter Fragen wie:

- Welche notwendigen (Infra-) Strukturen und Strategien sind in den Communities vorhanden bzw. bereits weitgehend ausgearbeitet?
- Wie muss das Verhältnis von disziplinären und interdisziplinären (Infra-) Strukturen beschaffen sein?
- Ist nachhaltiges Forschungsdatenmanagement vor allem eine Frage der Skalierbarkeit hin zu zentralen Institutionen?

[1] Webseite http://www.leibniz-gemeinschaft.de/infrastrukturen/kii
[2] Webseite http://www.allianzinitiative.de/de/handlungsfelder/forschungsdaten

- Wie kann das Forschungsdatenmanagement die internationale Kollaboration innerhalb der Disziplinen fördern?
- Ist das Forschungsdatenmanagement vor allem der Forschungstätigkeit zuzuordnen oder handelt es sich eher um eine Infrastruktur?

## Erweitertes Domänenmodell

In der Diskussion zum Thema Forschungsdaten hat es sich als zweckmäßig erwiesen, zunächst Strukturen zu entwickeln, welche die Vielzahl von verschiedenen Aspekten zusammenfassen und so eine strukturierte Auseinandersetzung mit der Thematik ermöglichen. Naturgemäß ist dies eine schwierige Aufgabe, da einerseits die Komplexität der Thematik in eine überschaubare Zahl von Faktoren vereinfacht werden muss, jedoch andererseits die für den Kontext der jeweiligen Bertachtung entscheidenden Faktoren nicht ausgeblendet werden dürfen. Es ist daher nicht verwunderlich, dass in den letzten Jahren hierzu eine Vielzahl verschiedener Arbeiten erschienen ist. Die für diese Arbeit wichtigsten Beiträge werden im Folgenden diskutiert.

Treloar und Harboe-Ree[3] identifizieren zunächst verschiedene, thematische Kontinua, welche beispielsweise die Existenz von Metadaten beschreiben, oder den Aufwand, der für Datenmanagement getrieben wird. Durch Vergleich der einzelnen Kontinua extrahieren sie drei **Domänen**: die private Domäne, in der der individuelle Forscher bzw. die Forscherin für die von ihm/ihr verantwortete Forschungsdaten verantwortlich ist, die Gruppendomäne, in der mehrere Forscher/Forscherinnen Zugang zu einem gemeinschaftlichen Satz von Forschungsdaten haben und die öffentliche Domäne, in der (meist nach abgeschlossener Forschung) die Daten für die Öffentlichkeit und insbesondere andere Forscher zugänglich sind.

Eine mehr auf die Organisationsstrukturen fokussierte Sichtweise wird im JISC-Report *Dealing with Data: Roles, Rights, Responsibilities and Relationships*[4] vorgestellt. In der Arbeit werden, aufgrund von Interviews und einem Workshop, Handlungsempfehlungen für die Weiterentwicklung der Forschungsdateninfrastruktur des Vereinigten Königreiches entwickelt. Hierzu wird die sog. **4R Matrix** verwendet, welche für die relevanten Akteure die Rollen, Rechte, Pflichten und Beziehungen (Roles, Rights, Responsibilities, Relations) beschreibt.

Eigens um die Fähigkeit einer Wissenschaftsdisziplin zu bewerten, datenintensive Forschung zu betreiben, wurde das **Community Capability Model Framework**[5] entwickelt. Dieses Framework besteht aus acht Leistungsfaktoren, welche wiederum Kennwerte beinhalten, die die einzelnen Aspekte des Forschungsdatenmanagements abbilden.

---

[3] Vgl. Treloar, A. und Harboe-Ree, C. (2008) *Data management and the curation continuum: how the Monash experience is informing repository relationships*, Proceedings of VALA 2008, Melbourne, February.

[4] Vgl. Lyon, L. (2007) *Dealing with Data: Roles, Rights, Responsibilities and Relationships - Consultancy Report,* http://www.ukoln.ac.uk/ukoln/staff/e.j.lyon/reports/dealing_with_data_report-final.pdf, online.

[5] Vgl. Lyon et al. (2012) *Community Capability Model Framework,* http://communitymodel.sharepoint.com/Documents/CCMDIRWhitePaper-24042012.pdf, online.

Im Rahmen des WissGrid Projektes wurde, basierend auf dem auf Langzeitarchivierung ausgerichteten *Leitfaden zum Forschungsdaten-Management*[6] ein Lebenszyklus von Forschungsdaten vorgeschlagen. Dieser benennt die Aufgaben Planung und Erstellung, Auswahl, Ingest und Übernahme, Speicherung und Infrastruktur, Erhaltungsmaßnahmen und Zugriff und Nutzung. Neben diesen eindeutig im Lebenszyklus verortbaren Aufgaben werden auch **übergreifenden Aufgaben des Forschungsdaten-Managements** identifiziert. Hierbei handelt es sich um die Komplexe Kosten, Recht, Metadaten, Identifikatoren und Organisation.

In der vorliegenden Untersuchung liegt der Schwerpunkt auf den Akteuren und Organisationsstrukturen im Bereich von Forschungsdateninfrastruktur. Die von uns in der weiteren Diskussion gewählte Systematik soll den Umgang mit Forschungsdaten in den einzelnen Wissenschaftsdisziplinen adäquat abbilden, muss aber auch die Perspektiven von Forschern, Forschungseinrichtungen und Förderinstitutionen berücksichtigen. Insbesondere soll sie auf die von uns untersuchten Projekte zur Weiterentwicklung der Forschungsdateninfrastruktur anwendbar sein.

Als Grundlage hierbei dient das oben beschriebene Domänenmodell von Treloar und Harboe-Ree. Der dort vorgeschlagene Weg welchen die Daten durch die Domänen nehmen, ist jedoch lediglich geeignet zur Modellierung des Datenworkflows in einer einzelnen Institution, nicht ausreichend, um auf verschiedene Disziplinen angewendet zu werden. Vielmehr müssen die einzelnen Domänen nebeneinander stehend als mit einander verknüpfte Bereiche eines komplexen Forschungsdatenworkflows gesehen werden. Um auch nicht-öffentliche dauerhafte Datenhaltung, beispielsweise in Archiven, zu berücksichtigen, muss eine zusätzliche Domäne hinzugefügt werden. Auch ist die Veröffentlichung von Forschungsdaten möglicherweise auf eine bestimmte Zielgruppe beschränkt und es ist daher eher von Zugang als Veröffentlichung zu reden. Es werden daher die vier Domänen:

- Private Domäne
- Gruppendomäne
- Dauerhafte Domäne
- Zugangsdomäne

zur weiteren Bertachtung herangezogen. Um auch die Organisationsstrukturen, welche im Domänenmodell noch nicht ausreichend berücksichtigt sind, in die Betrachtung einzubeziehen, werden einzelne Aspekte der 4R Matrix übernommen. Die vier Domänen zeigen schon Überschneidungen mit den Rollen *Forscher*, *Institution*, *Datenzentrum* und *Nutzer* aus der 4R-Matrix. Durch das Hinzunehmen der externen Aspekte:

- Finanzierung
- Publikationen/Verlage

[6] Vgl. WissGrid (2012) *Leitfaden zum Forschungsdaten-Management*, http://www.wissgrid.de/publikationen/Leitfaden_Data-Management-WissGrid.pdf, online.

ist es daher möglich, mit diesem erweiterten Domänenmodell alle Akteure der 4R-Matrix abzudecken. Zusätzlich werden, angelehnt an die in WissGrid identifizierten übergreifenden Aufgaben des Forschungsdaten-Managements, weitere Aspekte hinzugefügt, welche nicht direkt mit einzelnen Domänen bzw. Akteuren zusammenhängen. Diese umfassen

- Art und Menge der anfallenden Daten
- Formate
- Metadaten
- Identifikatoren
- rechtliche Fragen

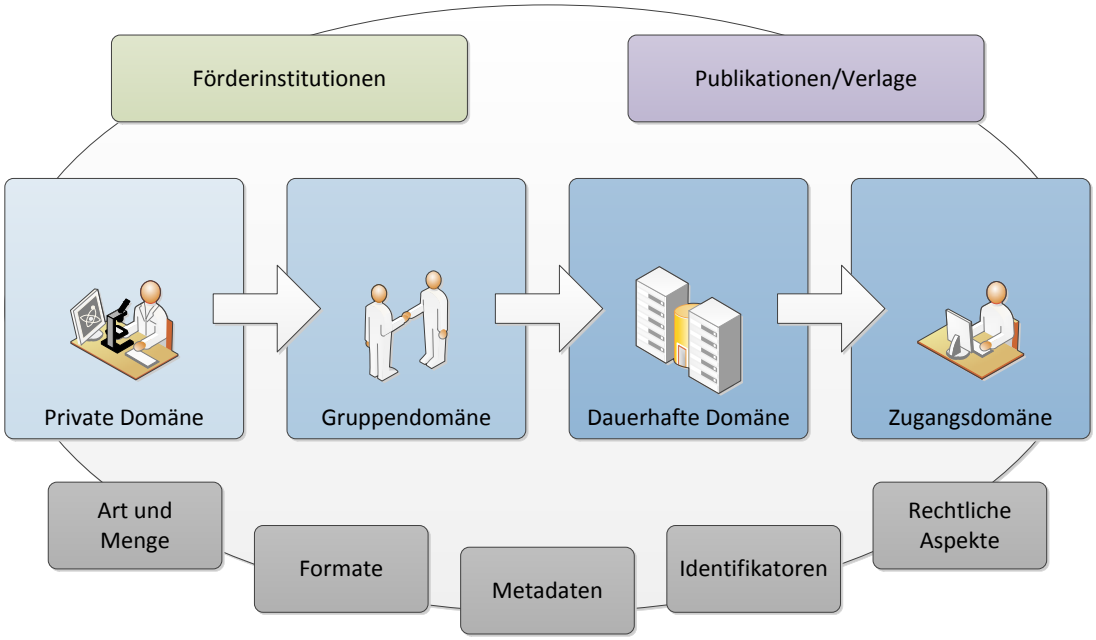


**Abbildung 1:** Erweitertes Domänenmodell

Die Erweiterung erlaubt auch, den Fluss der Daten (oder *scientific data flow*) und den wissenschaftlichen Workflow in und zwischen den Domänen systematisch zu unterscheiden und so deren Zusammenhänge genau darzustellen. Das komplette Erweiterte Domänenmodell ist in Abb. 1 dargestellt. Es dient als Systematik für die folgende Bestandsaufnahme und wird auch in den weiteren Betrachtungen als Grundlage dienen.

## Auswahl der untersuchten Projekte

Wie schon in der Einleitung besprochen sind in den letzten Jahren verschiedenste Projekte zum Forschungsdatenmanagement auf nationaler wie auch auf europäischer Ebene gefördert worden. Eine vollständige Bestandsaufnahme, welche all diese Projekte umfasst, würde den Rahmen dieser Arbeit sprengen. Wir beschränken uns daher auf die beiden wichtigsten

Förderungsinstrumente der DFG zum Thema Forschungsdatenmanagement der letzten Jahre, das Programmelement **Informationsmanagement und Informationsinfrastruktur in Sonderforschungsbereichen**, die sogenannten **INF-Projekte**, und die Ausschreibung **Informationsinfrastrukturen für Forschungsdaten** vom April 2010. Im Folgenden werden diese Förderungsinstrumente kurz vorgestellt und die geförderten Projekte aufgelistet. In der folgenden Bestandsaufnahme werden die Projekte dann ausführlicher diskutiert. Zusätzlich werden dort auch diverse andere Projekte, welche eine besondere Relevanz in Bezug auf den Umgang mit Forschungsdaten haben, besprochen.

Innerhalb der Sonderforschungsbereiche (SFB) und der Transregio (TRR), die von der DFG gefördert werden, dienen die **INF-Projekte** dazu, gezielt den nachhaltigen Umgang mit den in dem SFB bzw. TRR anfallenden Daten zu fördern. Hierzu wird die Entwicklung und Umsetzung eines Datenmanagementkonzeptes gefördert, welches dann in den INF-Projekten umgesetzt wird. Zu den Aufgaben dieser Projekte gehören neben dem Forschungsdatenmanagement auch der Aufbau und der Betrieb der dafür notwendigen Infrastruktur für die Dauer des SFB/TRR.[7] Darüber hinausgehende Nachhaltigkeit wird von der DFG bei den Hochschulen verortet.

Ein Schwerpunkt unserer Arbeit konzentriert sich daher auf die als INF-Projekt geförderten Teilprojekte der SFB und TRR die zurzeit von der DFG gefördert werden. Die einzelnen Projekte sind in Tabelle 1 zusammengefasst. Hierbei beschränken wir uns auf die SFB/TRR die Zwischen 2009 und 2012 begonnen haben. Die Eingruppierung in die diversen Fachgebiete ist hierbei der GEPRIS Datenbank[8] der DFG entnommen. Eine ausführliche Beschreibung der Projekte folgt dann im späteren Abschnitt *Bestandsaufnahme in den einzelnen Wissenschaftsdisziplinen*.

| Typ | Nr. | Titel | Projekt | Projekttitel | Fachgebiete |
|---|---|---|---|---|---|
| SFB | 649 | Ökonomisches Risiko | INF | Research Data Center (RDC) | Sozial- und Verhaltenswissenschaften<br>Mathematik |
| TRR | 62 | Eine Companion-Technologie für kognitive technische Systeme | Z03 | Data Management and Systems Integration | Medizin<br>Elektrotechnik, Informatik und Systemtechnik<br>Bauwesen und Architektur |
| SFB | 806 | Unser Weg nach Europa: Kultur-Umwelt Interaktion und menschliche Mobilität im Späten Quartär | INF | Data Management and Data Service | Geisteswissenschaften<br>Agrar-, Forstwissenschaften, Gartenbau und Tiermedizin<br>Geowissenschaften (einschl. Geographie) |
| SFB | 833 | Bedeutungskonstitution - Dynamik und Adaptivität sprachlicher Strukturen | INF | The Storage of Diverse Data Types - Representation and Processing | Geisteswissenschaften<br>Sozial- und Verhaltenswissenschaften |
| SFB | 850 | Kontrolle der Zellmotilität bei Morphogenese, Tumorinvasion und Metastasierung | Z01 | Validation of Mechanisms Involved in Invasion and Metastasis in Human Cancer | Biologie<br>Medizin |
| SFB | 852 | Ernährung, intestinale Mikrobiota und Wirtsinteraktionen beim Schwein | INF | Central technique and bioinformatic toolbox | Medizin<br>Agrar-, Forstwissenschaften, Gartenbau und Tiermedizin |
| SFB | 884 | Politische Ökonomie von Reformen | Z01 | Data Centre / Internet Panel | Geisteswissenschaften<br>Sozial- und Verhaltenswissenschaften |

[7] Vgl. http://www.dfg.de/foerderung/programme/koordinierte_programme/sfb/programmelemente/programmelement_inf/index.html, online.
[8] Webportal http://gepris.dfg.de

| | | | | | |
|---|---|---|---|---|---|
| TRR | 51 | Ökologie, Physiologie und Molekularbiologie der Roseobacter-Gruppe: Aufbruch zu einem systembiologischen Verständnis einer global wichtigen Gruppe mariner Bakterien | INF | Information infrastructure, database and bioinformatics tool development | Biologie<br>Medizin<br>Chemie |
| TRR | 77 | Leberkrebs - von der molekularen Pathogenese zur zielgerichteten Therapie | Z02 | Integrated information platform and cross-project biostatistical analyses | Biologie<br>Medizin |
| SFB | 673 | Ausrichtung in der Kommunikation | INF | Multimodal alignment corpora: statistical modeling and information management | Geisteswissenschaften<br>Sozial- und Verhaltenswissenschaften<br>Elektrotechnik, Informatik und Systemtechnik |
| SFB | 881 | Das Milchstraßensystem | INF | Computer and Information Infrastructure | Physik |
| TRR | 32 | Muster und Strukturen in Boden-Pflanzen-Atmosphären-Systemen: Erfassung, Modellierung und Datenassimilation | INF | Project Database and Data Management | Agrar-, Forstwissenschaften, Gartenbau und Tiermedizin<br>Mathematik<br>Geowissenschaften (einschl. Geographie) |
| SFB | 632 | Informationsstruktur: Die sprachlichen Mittel der Gliederung von Äußerung, Satz und Text | D01 | Linguistic database for information structure | Geisteswissenschaften |
| SFB | 882 | Von Heterogenitäten zu Ungleichheiten | INF | Information- and Datamanagement | Geisteswissenschaften<br>Sozial- und Verhaltenswissenschaften |
| SFB | 933 | Materiale Textkulturen. Materialität und Präsenz des Geschriebenen in non-typographischen Gesellschaften | INF | Service-Project on Information Management | Geisteswissenschaften |
| SFB | 950 | Manuskriptkulturen in Asien, Afrika und Europa | INF | Data repository manuscript cultures | Geisteswissenschaften<br>Materialwissenschaft und Werkstofftechnik<br>Elektrotechnik, Informatik und Systemtechnik |
| SFB | 991 | Die Struktur von Repräsentationen in Sprache, Kognition und Wissenschaft | INF | Service Project for Information Infrastrukture | Geisteswissenschaften<br>Medizin |
| SFB | 963 | Astrophysikalische Strömungsinstabilität und Turbulenz | INF | ADIR: Das AstroFIT Daten-InfRastrukturprojekt | Physik |
| SFB | 990 | Ökologische und sozioökonomische Funktionen tropischer Tieflandregenwald-Transformationssysteme (Sumatra, Indonesien) | INF | Web-GIS based information system and research data management | Geisteswissenschaften<br>Sozial- und Verhaltenswissenschaften<br>Biologie<br>Medizin<br>Agrar-, Forstwissenschaften, Gartenbau und Tiermedizin<br>Geowissenschaften (einschl. Geographie) |
| SFB | 1002 | Modulatorische Einheiten bei Herzinsuffizienz | INF | Vom Laborbuch zur Datenbank - Flexible Erhebung von Metadaten als Herzstück eines modernen Datenmanagements | Biologie<br>Medizin |
| SFB | 1026 | Sustainable Manufacturing - Globale Wertschöpfung nachhaltig gestalten | INF | Informationsinfrastruktur | Maschinenbau und Produktionstechnik |

**Tabelle 1:** Übersicht über die INF-Projekte in den SFB/Transregio der DFG.

Eine weitere bedeutende Maßnahme zur gezielten Förderung des Forschungsdatenmanagement in Deutschland stellte die Ausschreibung **Informationsinfrastrukturen für Forschungsdaten** des Bereiches *Wissenschaftliche Literaturversorgung und Informations-*

*systeme (LIS)*[9] der DFG aus dem April 2010 dar. In diesem Rahmen wurden eine Reihe von Infrastrukturprojekten aus verschiedenen Fachgebieten gefördert[10]. Aufgrund der besonderen Relevanz für die Forschungsdateninfrastruktur in den jeweiligen Disziplinen stellen diese Projekte einen weiteren Schwerpunkt unserer Untersuchung dar. In Tabelle 2 sind die von uns betrachteten Projekte aufgeführt. Während die Eingruppierung in Fachbereiche wiederum der GEPRIS Datenbank entnommen ist, wurde die Verortung in den vier Domänen als Teil dieser Arbeit vorgenommen. Wie zuvor werden auch diese Projekte ausführlicher im Abschnitt *Bestandsaufnahme in den einzelnen Wissenschaftsdisziplinen* besprochen.

| Projekttitel | Fachgebiet | Domäne/Aspekt |
|---|---|---|
| IANUS - Forschungsdatenzentrum Archäologie & Altertumswissenschaften | Geisteswissenschaften | Dauerhafte Domäne<br>Zugangsdomäne |
| OpenInfRA - ein webbasiertes Informationssystem zur Dokumentation und Publikation archäologischer Forschungsprojekte | Geisteswissenschaften<br>Geowissenschaften (einschl. Geographie) | Zugangsdomäne |
| Zentrum für germanistische Forschungsprimärdaten | Geisteswissenschaften | Dauerhafte Domäne<br>Zugangsdomäne |
| Etablierung eines Schwerpunkts "Mehrsprachigkeit und gesprochene Sprache" im Hamburger Zentrum für Sprachkorpora | Geisteswissenschaften | Gruppendomäne<br>Dauerhafte Domäne<br>Zugangsdomäne |
| Wissensspeicher - Daten geisteswissenschaftlicher Grundlagenforschung | Geisteswissenschaften | Gruppendomäne<br>Dauerhafte Domäne<br>Zugangsdomäne |
| LAUDATIO - Entwicklung einer nachhaltigen und nutzerorientierten Speicherung und Bereitstellung von Forschungsdaten für die historische Linguistik | Geisteswissenschaften | Dauerhafte Domäne<br>Zugangsdomäne |
| InFoLiS - Integration von Forschungsdaten und Literatur in den Sozialwissenschaften | Sozial- und Verhaltenswissenschaften | Zugangsdomäne<br>Verlage/Publikationen |
| Einrichtung eines Zentrums für Record-Linkage | Sozial- und Verhaltenswissenschaften | Zugangsdomäne |
| Professionalisierung und Ausbau des Forschungsdatenzentrums Survey of Health, Ageing and Retirement in Europe (SHARE) | Sozial- und Verhaltenswissenschaften | Zugangsdomäne |
| Aufbau einer Serviceeinrichtung für Daten der qualitativen empirischen Sozialforschung. Primärdaten für eScience in den Sozialwissenschaften | Sozial- und Verhaltenswissenschaften | Dauerhafte Domäne<br>Zugangsdomäne |
| da/ra - Aufbau einer Registrierungsagentur für sozialwissenschaftliche Forschungsdaten | Sozial- und Verhaltenswissenschaften | Identifikatoren |
| MISSY 3.0 - Forschungsbasierte Metadaten für amtliche Erhebungen: Ausbau von MISSY | Sozial- und Verhaltenswissenschaften | Gruppendomäne |
| EDaWaX (European Data Watch Extended) - Verbesserte Replizierbarkeit von Forschungsergebnissen in der empirischen Wirtschaftsforschung mit Hilfe eines publikationsbezogenen Datenarchivs | Sozial- und Verhaltenswissenschaften | Zugangsdomäne |
| reBiND - Entwicklung von Workflows und Softwarekomponenten zur Rettung lebenswissenschaftlicher Primärdaten | Biologie | Private Domäne<br>Dauerhafte Domäne |
| Ein generisches Annotationssystem für Biodiversitätsdaten | Biologie | Dauerhafte Domäne<br>Formate<br>Rechtliche Aspekte |
| BExIS- Modularisierung und Skalierung der BEXIS Experimentdaten-plattform | Biologie | Gruppendomäne |
| Development of the Golm Metabolome Database as a central plant metabolomics information resource. | Biologie | Dauerhafte Domäne<br>Zugangsdomäne |
| Extension and modification of Morph D Base producing a system for permanent storage and documentation of volume data of biological objects in high resolution | Biologie | Gruppendomäne |
| SILVA 2.0: Building the next generation databases for ribosomal RNAs | Medizin | Gruppendomäne<br>Zugangsdomäne |
| ARB im Zeitalter der Hochdurchsatzsequenzierung: Anpassung an die Erfordernisse umfassender Umwelt- und Metagenomstudien sowie Pflege entsprechender Datenbanken | Medizin | Gruppendomäne |

[9] Webseite http://www.dfg.de/foerderung/programme/infrastruktur/lis
[10] Vgl. http://www.dfg.de/download/pdf/foerderung/programme/lis/projekte_forschungsdaten.pdf, online.

| | | |
|---|---|---|
| Modular Support System for Planing and Implementation of a central Datenmanagement for Research Projects in Health Sciences | Medizin | Gruppendomäne |
| LaBiMi, Langzeitarchivierung biomedizinischer Forschungsdaten | Medizin | Dauerhafte Domäne |
| Compound Research System (CoRS): Literaturinformationssystem zur Analyse der physiologischen Wirkung von Kleinmolekülen | Medizin | Zugangsdomäne<br>Verlage/Publikationen |
| PubFlow - Publikationsprozesse für Forschungsdaten: Von der Erhebung und Verarbeitung zur Archivierung und Publikation | Geowissenschaften (einschl. Geographie);Elektrotechnik, Informatik und Systemtechnik | Private Domäne<br>Gruppen Domäne<br>Dauerhafte Domäne<br>Verlage/Publikationen |
| KOMFOR - Kompetenzzentrum für Forschungsdaten aus Erde und Umwelt Competence Centre for Geoscientific Research Data | Geowissenschaften (einschl. Geographie) | Private Domäne<br>Gruppen Domäne<br>Dauerhafte Domäne<br>Zugangsdomäne<br>Verlage/Publikationen<br>Institutionen |
| EWIG - Entwicklung von Workflowkomponenten für die Langzeitarchivierung von Forschungsdaten im Bereich Erd- und Umweltwissenschaften | Geowissenschaften (einschl. Geographie);Elektrotechnik, Informatik und Systemtechnik | Dauerhafte Domäne<br>Zugangsdomäne |
| Aufbau des Dateninformationssystems für das GESEP Kern- und Probenlager zur Erfassung und Verwaltung von Bohrkernen und Nachweis der Bestände in einem Internetportal | Geowissenschaften (einschl. Geographie) | Dauerhafte Domäne<br>Zugangsdomäne |
| Aufbau eines Informationssystems für werkstoffwissenschaftliche Forschungsdaten mittels Technologien zur semantischen Wissensverarbeitung | Materialwissenschaft und Werkstofftechnik | Dauerhafte Domäne |
| Implementation of the DNA-Bank-Network as Service for Science in Germany | Geowissenschaften (einschl. Geographie) | Gruppendomäne<br>Dauerhafte Domäne |
| Digitalisierung und Online-Verfügbarmachung des Hohhaus-Herbariums und der Sammlung Goldschmidt | Biologie | Dauerhafte Domäne<br>Zugangsdomäne |
| Entwicklung und Etablierung eines Datenarchivs für die Erfassung, Systematisierung und Bereitstellung von Forschungsprimärdaten humaner Stammzellen für die regenerative Biologie und Medizin | Medizin | Dauerhafte Domäne<br>Zugangsdomäne |

**Tabelle 2:** Übersicht über die Projekte des DFG-Programms *Informationsinfrastrukturen für Forschungsdaten*.

## Verteilung der untersuchten Projekte über die Fachgebiete und Domänen

Nachdem im vorherigen Abschnitt die in dieser Arbeit primär untersuchten Projekte aus dem Programmen *Informationsmanagement und Informationsinfrastruktur in Sonderforschungs-bereichen* und *Infrastrukturen für Forschungsdaten* der DFG vorgestellt worden sind, soll nun die Verteilung dieser Projekte über die Fachgebiete untersucht werden. Zu diesem Zweck

| Wissenschaftsbereich | Fachgebiet |
|---|---|
| **Geistes- und Sozialwissenschaften** | Geisteswissenschaften |
| | Sozial- und Verhaltenswissenschaften |
| **Lebenswissenschaften** | Biologie |
| | Medizin |
| | Agrar-, Forstwissenschaften, Gartenbau und Tiermedizin |
| **Naturwissenschaften** | Chemie |
| | Physik |
| | Mathematik |
| | Geowissenschaften (einschl. Geographie) |
| **Ingenieurswissenschaften** | Maschinenbau und Produktionstechnik |
| | Wärmetechnik/Verfahrenstechnik |
| | Materialwissenschaft und Werkstofftechnik |
| | Elektrotechnik, Informatik und Systemtechnik |
| | Bauwesen und Architektur |

**Tabelle 3:** DFG Systematik der Wissenschaftsbereiche und Fachgebiete.

orientieren wir uns an der Fachsystematik der DFG[11]. Auf der obersten Ebene unterscheidet die DFG die Wissenschaftsbereiche Geistes- und Sozialwissenschaften, Lebenswissenschaften, Naturwissenschaften und Ingenieurswissenschaften. Diese gliedern sich weiter in Fachgebiete auf die in Tabelle 3 dargestellt sind.

Zusätzlich wird für das Programm *Infrastrukturen für Forschungsdaten* die Verteilung der Projekte über die im zweiten Kapitel vorgestellten Domänen des Forschungsdatenmanagements untersucht.

In Abb. 1 ist die Verteilung der INF Projekte über die verschiedenen Fachgebiete dargestellt. Da auch die Anzahl der SFB und TRR stark über die Fachgebiete variiert sind die SFB/TRR mit INF-Projekt denen ohne gegenübergestellt. Es fällt auf, das Instrument der INF-Projekte besonders in den Geistes- und Sozialwissenschaften sowie den Lebenswissenschaften wahrgenommen wird, während es in den Naturwissenschaften (mit Ausnahme der Geowissenschaften) und den Ingenieurswissenschaften deutlich weniger präsent ist.

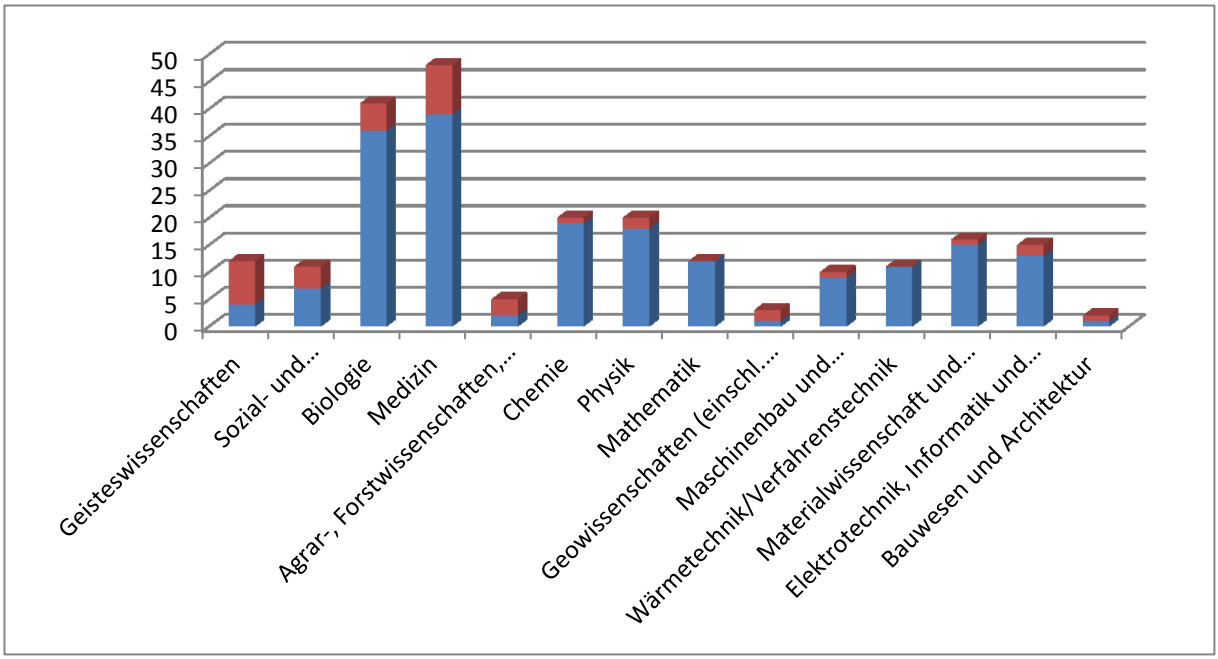


**Abbildung 1:** Verteilung der Sonderforschungsbereiche mit INF-Projekt (*rot*) und ohne (*blau*) über die verschiedene Fachgebiete. Aufgrund der Vergleichbarkeit, werden hierbei nur die SFB bzw. TRR betrachtet deren Förderung im Jahr 2009 oder später begonnen hat.

Bei den Projekten des Programms *Infrastrukturen für Forschungsdaten* ist die Situation ähnlich. In Abb. 2 ist die Verteilung der dieser Projekte über die Fachgebiete in ähnlicher Weise zu Abb. 1 dargestellt. Auch hier ist eine klare Häufung in den Geistes- und Sozialwissenschaften sowie den Lebenswissenschaften erkennbar und den Geowissenschaften erkennbar.

[11] Vgl. Grafik zur Systematik der Fächer und Fachkollegien der DFG für die Amtsperiode 2012-2015 auf der Webseite der DFG, http://www.dfg.de/download/pdf/dfg_im_profil/gremien/fachkollegien/amtsperiode_2012_2015/fachsystematik_2012_2015_de_grafik.pdf, online.

Es stellt sich nun die Frage, welche Gründe diese Inhomogenität in der Verteilung über die Fachgebiete hat. Interessanterweise werden diese Programme gerade von den Wissenschaftsdisziplinen dominiert, welche in der Regel nicht mit großen Datenmengen assoziiert werden. Dies ist zum einen ein Zeichen, dass auch in vermeintlich technikfernen Disziplinen ein moderner Umgang mit Forschungsdaten Einzug hält. Es kann aber auch daran liegen, dass diese Disziplinen traditionell von einer Kultur der Archivierung geprägt sind, die jetzt eine Entsprechung in der Informationstechnik sucht.

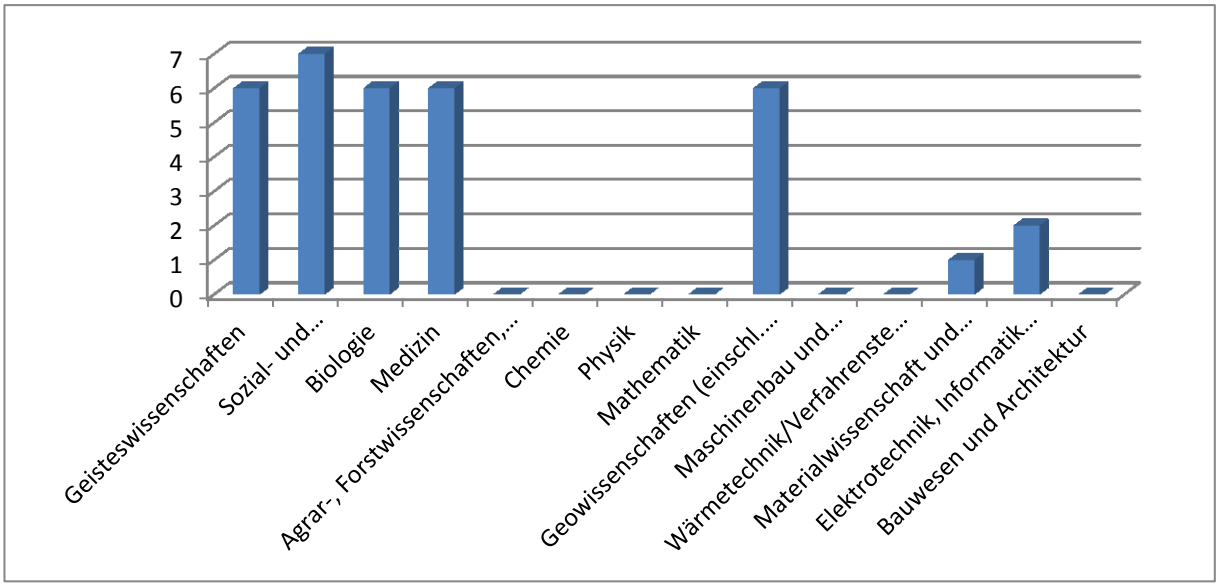


**Abbildung 2:** Verteilung der Projekte des DFG-Calls Infrastrukturen für Forschungsdaten über die verschiedene Fachgebiete. Die Gesamtzahl der betrachteten Projekte beträgt 31.

In diesem Sinne könnte ein Grund für die vielen Projekte in den Geisteswissenschaften die traditionelle Nähe zu Bibliotheken sein. In den Sozial- und Wirtschaftswissenschaften, den Lebenswissenschaften und den Geowissenschaften haben Forschungsdaten schon immer einen dezidiert episodischen Charakter. Es existieren dort daher schon seit längerer Zeit Sammlungen, um diese Daten zukünftigen Forschungsprojekten zugänglich zu machen.

In den übrigen Naturwissenschaften und insbesondere in den Ingenieurwissenschaften ist die technologische Entwicklung in der Vergangenheit ein wichtiger Faktor für neue wissenschaftliche Erkenntnisse gewesen. Dies hatte jedoch zur Folge, dass ältere Forschungsdaten mit dem Aufkommen neuer Experimente schnell obsolet wurden. Als zweiter wichtiger Faktor ist der kommerzielle Wert von Daten zu nennen, der in anwendungsnahen Gebieten dem öffentlichen Zugang zu Archiven entgegenwirkt. Erst mit dem gegenwärtigen Trend zur datenbasierten Wissenschaft wird das Forschungsdatenmanagement zu einem entscheidenden Faktor. In jedem Fall ist jedoch nicht aus den Augen zu verlieren, dass die Anzahl der Wissenschaftlerinnen und Wissenschaftler, der Institutionen und des Förderungsvolumens mit einem Schwerpunkt im Bereich der Forschungsdateninfrastruktur unterschiedlich über die Disziplinen verteilt ist.

Anders als bei den INF-Projekten, die Datenmanagement und Infrastruktur für den jeweiligen SFB bzw. TRR bereitstellen und daher in jedem Fall in der Gruppendomäne zuzuordnen

sind, verteilen sich die Projekt des Programms *Infrastrukturen für Forschungsdaten* über alle vier Domänen. In Abb. 3 ist die entsprechende Verteilung dargestellt. Offensichtlich konzentriert sich die Mehrzahl der Projekte auf die dauerhafte Domäne und die Zugangsdomäne. Projekte, die zum Datenmanagement innerhalb der Gruppendomäne forschen, sind dem gegenüber in der Minderzahl. Die private Domäne, die die Arbeit der Forscherinnen und Forscher in den Fachdisziplinen unmittelbar betrifft, wird nur von wenigen Projekten betrachtet.

Es ist festzuhalten das, zumindest im Programm *Infrastrukturen für Forschungsdaten*, verstärkt Projekte gefördert wurden, welche ihren Schwerpunkt in den Bereich der Langzeitarchivierung und der Nachnutzung von Forschungsdaten gelegt haben. Dies geschieht beispielsweise durch Erarbeitung von Archivkonzepten, Entwicklung entsprechenden Softwarelösungen oder durch die Bereitstellung von Webportalen zur Nachnutzung. Projekte die an Infrastrukturen arbeiten, die die Situation im Datenmanagement von Kollaborationen verbessern sollen, sind dem gegenüber unterrepräsentiert. Die private Domäne wird von Forschung zum Forschungsdatenmanagement bzw. zur Forschungsdateninfrastruktur fast noch nicht beachtet.

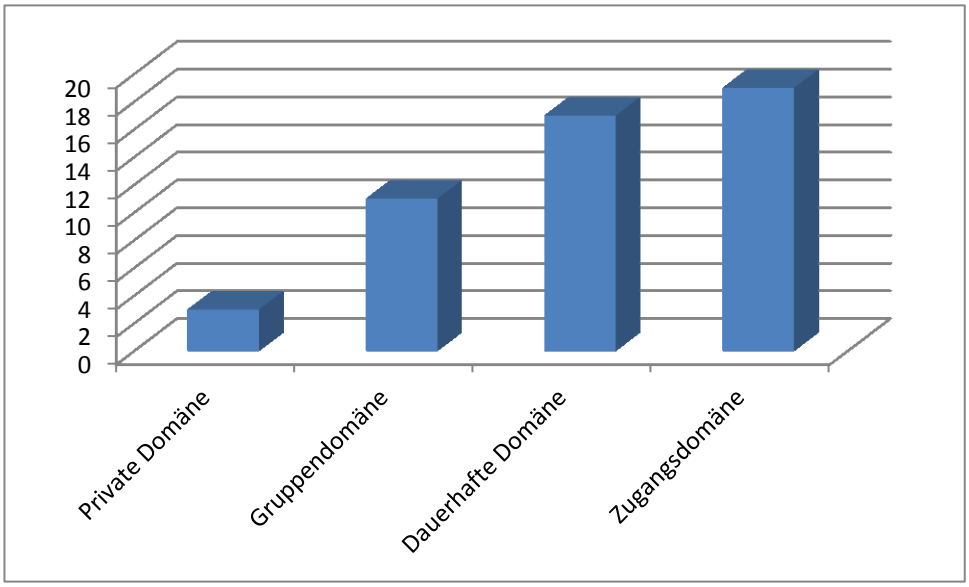


**Abbildung 3:** Verteilung der Projekte des DFG-Calls *Infrastrukturen für Forschungsdaten* über die vier Domänen des Forschungsdatenmanagements . Die Gesamtzahl der betrachteten Projekte beträgt 31. Dadurch, dass verschiedene Projekte in mehr als einer Domäne verortet sind, ist die Addition der Anzahl in den einzelnen Domänen höher als diese Zahl.

## Bestandsaufname in den einzelnen Wissenschaftsdisziplinen

Im Folgenden wird nun ausführlich auf den Umgang mit Forschungsdaten in den einzelnen Forschungsdisziplinen eingegangen. Besondere Bedeutung fällt hierbei den im Forschungsdatenmanagement aktiven Akteuren zu. Wir weichen hierbei von der bisher verwendeten DFG Systematik ab. Dies hat mehrere Gründe:

- Die DFG Systematik fasst teilweise (Sub-)Disziplinen zusammen, die zwar thematisch nahe beieinanderliegen, jedoch mit Blick auf das Forschungsdatenmanagement stark unterschiedlich sind. Das betrifft nicht nur die Art und Menge der Daten, sondern auch die Community, die sich um die Forschungsdateninfrastruktur herausgebildet hat. Beispiele hierfür sind die Astro- und die Hochenergiephysik als Teildisziplinen der Physik, aber auch die Biodiversität als der Biologie nahe, jedoch interdisziplinär angelegte Forschungsdisziplin.
- Die in dieser Arbeit im Fokus stehenden Projekte (*INF* und *Infrastruktur für Forschungsdaten*) sind sehr ungleich über die Fachgebiete verteilt. Insbesondere die Ingenieurswissenschaften sind deutlich unterrepräsentiert.
- Auch die im Radieschen-Projekt durchgeführten Interviews mit ausgewählten Akteuren im Bereich des Forschungsdatenmanagements konnten nicht alle Disziplinen abdecken.

Im Folgenden beschränken wir uns daher auf die Wissenschaftsdisziplinen:

- Geisteswissenschaften und Psycholinguistik
- Altertumswissenschaften
- Sozial- und Wirtschaftswissenschaften
- Biodiversität
- Medizin
- Astrophysik
- Geo-, Meeres- und Klimawissenschaften

Es mag auffallen, dass mit der Hochenergiephysik eine prominente, datenintensive Disziplin fehlt. Wir verweisen hier auf die Arbeiten des Nestor Projektes bzw. die dort zitierten Arbeiten.[12] Weiter sind mit den Geo-, Meeres- und Klimawissenschaften mehrere Disziplinen zusammengefasst die in anderen Publikationen oft getrennt aufgeführt sind. Dies liegt in der starken Überschneidung der Akteure (insbesondere der Zentren des World Data Systems) in diesen Disziplinen begründet.

Projekte die nicht in die oben aufgeführte Systematik passen, jedoch von uns in dieser Arbeit untersucht wurden, werden am Ende diese Kapitels beschrieben.

## Geisteswissenschaften und Psycholinguistik

### Wichtige Akteure

Das Projekt **TextGrid**[13] stellt eine Virtuelle Forschungsumgebung und eine damit verbundene Langzeitarchivierungskomponente für die Geistes- und Kulturwissenschaften zur Verfügung[14].

[12] Vgl. Gülzow, V., Kemp, Y. (2012) in: Neuroth, H. et al. (Hrsg.) (2012) *Langzeitarchivierung von Forschungsdaten – Eine Bestandsaufnahme*, Universitätsverlag Göttingen, S. 257ff
[13] Projektwebseite http://www.textgrid.de
[14] Vgl. Pempe, W. (2012) in: Neuroth, H. et al. (Hrsg.) (2012) *Langzeitarchivierung von Forschungsdaten – Eine Bestandsaufnahme*, Universitätsverlag Göttingen, S. 142ff

Die **Berlin-Brandenburgische Akademie der Wissenschaften (BBAW)** ist im Bereich der Forschungsdaten unter anderem durch das Projekt **TELOTA**[15] aktiv. TELOTA dient den diversen Projekten der BBAW elektronische Arbeitsumgebungen und Beratung im Bereich Forschungsdatenmanagement und Datenpublikation.

Im Bereich der psycholinguistischen Forschungsdaten fällt dem **Max-Plank Institut für Psycholinguistik (MPI-PL)**[16] eine Vorreiterrolle zu. Das seit dem Jahr 2000 durch die Volkswagenstiftung finanzierte und am MPI-PL angesiedelte **DoBes Programm**[17] arbeitet an der Erhaltung bedrohter Sprachen für die Nachwelt. Hierfür werden annotierte Audio- und Videoaufnahmen und wissenschaftliche Beschreibungen der Sprachen in digitaler Form und in offenen Formaten bzw. Standards archiviert und über das Internet verfügbar gemacht. Das Institut ist weiter ein Partner im Europäischen **CLARIN** Projekt und in den deutschen und niederländischen Unterprojekten **CLARIN-D** (gefördert vom BMBF) und **CLARIN-NL**. **CLARIN** arbeitet an der Integration und der Interoperabilität der verschiedenen Ansätze im Forschungsdatenmanagement und der Langzeitarchivierung in der Linguistik und anderen Geisteswissenschaften. Seit September 2010 sind die Aktivitäten zur Archivierung und zum Datenmanagement am MPI-PL in einer neuen Einheit **The Language Archive**[18] gebündelt.

Innerhalb der Max Plank Gesellschaft ist die **Max Plank Digital Library**[19] für die Versorgung der Institute mit Publikationen und Publikationsdatenbanken zuständig.

Das von der **SUB Göttingen** koordinierte Projekt **DARIAH-DE**[20] (gefördert vom BMBF) unterstützt den Aufbau von virtuellen Forschungsumgebungen in den Geisteswissenschaften in Deutschland. DARIAH-DE bietet hierfür Beratung, Aktivitäten der Community-Bildung, Schulungen und technische Infrastruktur[21]. Ziel ist eine nachhaltige dezentrale Infrastruktur für die Digital Humanities in Deutschland. Das Projekt ist Teil des auf europäischer Ebene angesiedelten ESFRI Projekts **DARIAH-EU**[22].

Ebenfalls auf Europäischer Ebene arbeiten im Projekt **DASISH**[23] die Projekte DARIAH und CLARIN, zusammen mit den Sozialwissenschaftlichen Projekten CESSDA, ESS und SHARE an Lösungen für gemeinsame Fragestellungen, primär in den Feldern Qualitätsmanagement, Archivierung, Zugang, sowie rechtlichen bzw. ethischen Fragen.

Der Verband **DHD - Digital Humanities Deutschland**, gegründet im Juli 2012, versteht sich als Forum und Interessenvertretung der Forscher und Forscherinnen in den Digital Humanities im deutschsprachigen Raum und sieht sich auch als Bindeglied zur **Association for Literary and Linguistic Computing (ALLC)** und der **Alliance of Digital Humanities Organizations**

[15] Projektwebseite http://www.bbaw.de/telota
[16] Institutswebseite http://www.mpi.nl
[17] Dokumentation Bedrohter Sprachen, Webseite http://www.mpi.nl/DOBES/dobesprogramme
[18] Webseite http://www.mpi.nl/departments/other-research/research-projects/the-language-archive
[19] Webseite http://www.mpdl.mpg.de
[20] Digital Research Infrastructure for the Arts and Humanities, Website http://de.dariah.eu
[21] Vgl. Selbstdarstellung des Projektes http://de.dariah.eu/projektziele.html, online.
[22] Website http://dariah.eu
[23] Digital Services Infrastructure for Social Sciences and Humanities, Website http://dasish.eu

**(ADHO)**. Auf der Seite Webseite des Verbandes wird unter anderem eine Liste über die verschiedenen Projekte in den Digital Humanities in Deutschland gepflegt.

## Domänenübergreifende Faktoren

### *Art und Menge der Daten*

Bei digitalen Forschungsdaten in den Geisteswissenschaften handelt es sich in der Regel um Digitalisierungen von gedruckten Texten, d.h. Scans oder Fotographien. Hierzu kommen im Laufe der wissenschaftlichen Arbeit getätigte Annotationen, Übersetzungen, sowie eher technische Daten wie Schemata und Workflows. In der Psycholinguistik kommen hierzu noch experimentelle Daten und sog. Observationelle Daten hinzu. Der Umfang der Datensätze welche mit einem Forschungsprojekt assoziiert werden können, reicht von kleinsten Volumina bis zu mehreren zehn Terabyte wie im TextGrid Projekt oder gar bis zu hunderten von TB wie am Archiv des MPI-PL. Diese Zahlen werden sich, nicht zuletzt aufgrund der diversen Digitalisierungsprojekte (z.B. des Göttinger Digitalisierungszentrums[24]), in der nächsten Zeit deutlich erhöhen. Der Umfang der gesamten (analogen) geisteswissenschaftlichen Datenbasis kann auf mehrere Petabyte geschätzt werden[25].

### *Formate*

Als Formate werden meist Standardformate wie TIFF, JPG oder PDF (für Scans und Bilder), bzw. die Microsoft Office Formate (für Textdateien) genutzt. Für annotierte Daten werden meist auf XML aufbauende Formate verwendet. Dies umfasst insbesondere die Formate der Text Encoding Initiative (z.B. das im TextGrid verwendete TEI P5). Die Nutzung von proprietären Formaten schafft Probleme bei der LZA, daher wird beispielsweise im Archiv des MPI-PL zwischen einem Bereich für Standardformate mit strengen formalen Anforderungen und einem Bereich, in dem auch proprietäre Formate verwendet werden können, getrennt.

### *Metadaten*

Aufgrund der traditionellen Nähe zu den Bibliotheken ist in den Geisteswissenschaften die Nutzung von (standardisierten) Metadaten bereits weit verbreitet. Aufgrund der starken Heterogenität der Daten hat sich jedoch kein einheitliches Format durchsetzen können. Eines der Formate, die als etabliert bezeichnet werden können ist Dublin Core[26], bzw. die davon abgeleiteten sog. Application Profiles wie beispielsweise PREMIS. Aber auch Metadatenformate aus dem Bibliotheksumfeld wie MARC oder MODS sind in Verwendung. Eine Übersicht über diese Formate bietet die Webseite der Library of Congress[27]. Im Archiv des MPI-PL werden alle Daten mit IMDI [28]Metadaten beschrieben. CLARIN entwickelt mit CDMI

[24] Webseite http://gdz.sub.uni-goettingen.de/gdz/
[25] Vgl. Pempe, W. (2012) in: Neuroth, H. et al. (Hrsg.) (2012) Langzeitarchivierung von Forschungsdaten – Eine Bestandsaufnahme, Universitätsverlag Göttingen S. 148
[26] Webseite der Dublin Code Metadata Initiative http://dublincore.org
[27] Übersichtsseite Metadatenformat http://www.loc.gov/standards/
[28] ISLE Metatdata Initiative, Webseite http://www.mpi.nl/imdi/

einen neuen Metadatenstandard, um die semantische Interoperabilität der verschiedenen Standards sicherzustellen.

### *Identifikatoren*

Als Identifikatoren werden verschiedene Systeme verwendet. Im Archiv des MPI-PL verfügen beispielsweise bereits alle Datensätze über Identifikatoren im Handle-Format. Auch TextGrid Projekt nutzt das Handle-Format (über EPIC[29]). Daneben werden unter anderem Uniform Resource Names (URN)[30] und DOI im Zusammenhang mit Publikationen verwendet.

### *Rechtliche Aspekte*

Unter den rechtlichen Aspekten sind in den Geisteswissenschaften sind die Urheber- und Verwertungsrechte der untersuchten Texte am wichtigsten. Die Rechte an traditionell publizierten Daten (i.e. Texte) liegen meist bei den Verlagen. Ein Zugang wird meist nur gegen Bezahlung und in einer für die informationstechnische Verarbeitung ungeeigneten Form ermöglicht.

## *Finanzierung*

TextGrid wurde als Teil der D-Grid Initiative vom BMBF gefördert, jedoch erfolgte die Finanzierung aus dem entsprechenden Fachreferat des BMBF, das auch eine zweite und dritte Förderung von TextGrid mitträgt. In der dritten Förderphase ist ein Projekt-Schwerpunkt die Ausarbeitung und Implementierung eines tragfähigen, nachhaltigen Betriebs.[31] Für das „Language Archive“ ist die Unterstützung durch die MPG das wohl entscheidende Kriterium für langfristig-nachhaltige Finanzierung.

## Umgang mit Forschungsdaten in den vier Domänen

### *Private Domäne*

Im Allgemeinen steht der nachhaltige Umgang mit Forschungsdaten in den Geisteswissenschaften noch am Anfang. Die allgemeinen Anforderungen zur Aufbewahrung von Forschungsdaten, wie die Regeln der DFG zur guten wissenschaftlichen Praxis, werden meist nur durch individuelle Lösungen umgesetzt. Insbesondere Daten aus abgeschlossenen Projekten sind oft nicht aufbereitet und nur in proprietären Formaten zugänglich. Teilweise wirken auch etablierte proprietäre Softwarelösungen einer stärker archivierungsfreundlichen Arbeitsweise entgegen. Nichtsdestotrotz entwickelt sich zurzeit eine Community um die sog. *Digital Humanities*, welche sich der Weiterentwicklung der klassischen Geisteswissenschaften durch die Möglichkeiten und Werkzeuge der modernen Informationsgesellschaft verschrieben hat.

---

[29] European Persisten Identifier Consortium, Webseite http://www.pidconsortium.eu
[30] Webseite http://www.persistent-identifier.de
[31] Webseite: http://textgrid.de/ueber-textgrid/materialien/antraege-und-berichte/. Insbesondere die Darlegungen des Antrags für die dritte Förderphase zeigen auf, dass TextGrid durch Konzentration auf eine spezifische Dienstleistung erfolgreich war, die in vielen angrenzenden Fachdisziplinen auch erwünscht war. Das ist jedoch noch nicht ausreichend für eine langfristige Sicherung. (Antrag 3, S. 12ff)

### *Gruppen Domäne*

Auf der Ebene der kollaborativen Zusammenarbeit hat sich in den letzten Jahren eine Vielzahl an Projekten etabliert. Insbesondere ist hier das schon weiter oben beschriebene TextGrid zu nennen. In der Psycholinguistik werden experimentell erhobene Daten, da sie für sehr spezialisierte Rahmenbedingungen erhoben werden, nur selten zwischen Forschergruppen geteilt. Bei den observationellen Daten ist dies anders, hier haben Projekte wie das schon beschriebene *DoBes* ein Bewusstsein für die Erhaltung und den Austausch der Daten geschaffen. Mit CLARIN sollen starke Zentren für die LZA (im Humanities Sektor) in Europa etabliert werden.

## Untersuchte Projekte

- Im Rahmen des INF Projekt des **SFB 600 Fremdheit und Armut** wird das sog. *Forschungsnetzwerk und Datenbanksystem (FuD)* entwickelt[32]. Es handelt sich hierbei um eine virtuelle Forschungsumgebung, die eine Arbeits-, Publikations- und Informationsplattform in sich vereinigt. Sie soll die Wissenschaftler und Wissenschaftlerinnen des SFBs bei allen Schritten ihres Forschungsprojektes unterstützen. Die Aufgaben der VRE umfassen die Datenaufnahme, die Datenanalyse, eine Redaktionsumgebung, eine Publikationsumgebung und ein Langzeitarchiv. Diese Software-Umgebung wird auch schon von externen Projekten eingesetzt. Im Projekt gibt es eine Mitarbeiterin, die speziell die Verbindung zwischen Entwicklern und Wissenschaftlern für Support, Training, aber auch Weiterentwicklung der Plattform koordiniert. Die Software-Umgebung des Projektes wird auch von anderen Instituten eingesetzt, dazu gehören die Akademie der Wissenschaften und der Literatur (Mainz) sowie die Akademie der Wissenschaften und der Künste in NRW. Weiter wird sie genutzt für die digitale Editionen der Werke Arthur Schnitzlers (ein großes Akademievorhaben der nordrhein-westfälischen Akademie), oder die Edition der Korrespondenz von August Wilhelm Schlegel, ein Projekt der sächsischen Landes- und Universitätsbibliothek zusammen mit der Phillips-Universität Marburg.
- Im **SFB 833 Bedeutungskonstitution - Dynamik und Adaptivität sprachlicher Strukturen** ist das INF Projekt zuständig für die nachhaltige Archivierung der im SFB anfallenden Datensätze[33]. Als Begründung werden hierbei die Vorgaben der DFG zur guten wissenschaftlichen Praxis, die Realisierung von erweiterten Publikationen (in denen die eigentliche Veröffentlichung mit den zugehörigen Primärdaten verknüpft ist) und eine etwaige Nachnutzung der Daten innerhalb und außerhalb des SFBs angegeben. Das Projekt kooperiert mit Clarin-D, NaLiDa[34] und dem Informations-, Kommunikations- und Medienzentrum der Universität Tübingen.

[32] Projektwebseite http://www.fud.uni-trier.de
[33] Projektwebseite http://www.sfb833.uni-tuebingen.de/inf-infrastrukturprojekt.html
[34] Projektwebseite http://www.sfs.uni-tuebingen.de/nalida/de/

- Das Projekt X1[35] des **SFB 673 Alignment in Communication** betreut zum einen die vom SFB genutzte (Daten-) Management-Plattform, arbeitet zum andren auch an der Modellierung der im SFB untersuchten Prozesse.
- Das INF Projekt[36] des **SFB 991 Die Struktur von Repräsentationen in Sprache, Kognition und Wissenschaft** bereitet Korpora auf, die verschiedenen Projekten innerhalb des SFB als empirische Ressource dienen sollen. Weiter bietet es den Projekten die Möglichkeit, gewonnene Daten in einheitlicher Weise (nach existierenden Standards) in Datenbanken abzulegen. Ferner wird der Internetauftritt des SFBs betreut.
- Ziel des INF Projektes des **SFB 950 Manuskriptkulturen in Asien, Afrika und Europa** ist die systematische Sicherung und langfristige Nutzung der Projektdaten des SFBs. Hierfür wird eine zentrale MyCoRe-Installation[37] verwendet. Es wird ein Datenrepositorium für digitale Objekte mit beschreibenden Metadaten aufgebaut, welches als virtuelle Forschungsumgebung für die Mitglieder des SFBs dienen soll. Zusätzlich werden die Forscherinnen und Forscher des SFB durch die Bereitstellung eines Wikis und das zum SFB gehörende Graduiertenkolleg durch Beratung zu E-Learning unterstützt.[38]
- Das Projekt D01[39] des **SFB 933 Materielle Textkulturen** arbeitet an dem Aufbau eines Content-Management-Systems für die kollaborative Arbeit im SFB.
- Im **SFB 632 Informationsstruktur: Die sprachlichen Mittel der Gliederung von Äußerung, Satz und Text** wird das Projekt D1[40] bereits der dritten Phase gefördert. Nachdem in den ersten beiden Phasen des SFBs eine linguistische Datenbank ANNIS entworfen wurde, an statistischen Auswertungsmethoden gearbeitet wurde (Mehrebenen-Annotationen, MEA), wird in der dritte Phase weiter an der Erstellung, Verfügbarmachung und Auswertung von Korpusdaten, der Betreuung der Software-Infrastruktur, der Betreuung und Pflege der Daten der empirisch arbeitenden Teilprojekte gearbeitet.

### *Dauerhafte Domäne*

Auf dem Gebiet der Langzeitarchivierung ist die Entwicklung von kooperativen Strukturen in den Geisteswissenschaften noch im Anfangsstadium. In letzten Jahren wurden die technischen Grundlagen unter anderem mit den Projekten **kopal**[41], **Planets**[42] und **SHAMAN**[43] erarbeitet. Digitale Publikationen werden zwar in der Deutschen Nationalbibliothek archiviert, jedoch ist es noch offen, wie mit zugehörigen Forschungsdaten verfahren wird[44]. Sogenannte

[35] Projektwebseite http://www.sfb673.org/projects/X1
[36] Projektwebseite http://www.sfb991.uni-duesseldorf.de/inf/
[37] Open Source Projekt: http://www.mycore.de/index.html
[38] Vgl. Poster Friedrich, M., Thiemann, S. http://www.manuscript-cultures.uni-hamburg.de/Poster/poster_inf_thiemann.pdf, online.
[39] Projektwebseite http://www.materiale-textkulturen.de/teilprojekt.php?tp=INF&up=
[40] Projektwebseite http://www.sfb632.uni-potsdam.de/aprojekte/d1.html
[41] Projektwebseite http://kopal.langzeitarchivierung.de
[42] Projektwebsite http://www.planets-project.eu/
[43] Projektwebseite http://shaman-ip.eu/shaman/
[44] Vgl. Webseite des EU-Projektes DRIVER http://www.driver-community.eu, online.

Digitale Editionen, d.h. dezidierte Internetauftritte mit einer dazugehörigen Datenbank, bieten hier eine Weiterentwicklung. Bei diesen Projekten ist jedoch noch eine stärkere Abstraktion zwischen Präsentation und Datenbestand nötig, um nicht für jede Edition Entwicklungsarbeiten unnötig zu wiederholen[45]. Das schon beschriebene TextGrid bietet bereits eine LZA Komponente. An der BBAW bietet **Telota** unterstützende Services für eine Reihe von relativ unabhängigen Projekten an. Einen herausgehobenen Platz innerhalb der Psycholinguistik hat das MPI-Archiv.

## Untersuchte Projekte

- Das von der DFG geförderte Projekt **Mehrsprachigkeit und gesprochene Sprache**[46] arbeitet am Ausbau der Arbeit des vorangegangenen Projektes Z2 „Computergestützte Erfassungs- und Analysemethoden multilingualer Daten“ (Laufzeit 2005-2011 als Teil des **SFB 538 Mehrsprachigkeit**). Gegenstand der Forschung in diesem Bereich sind Korpora gesprochener Sprache und deren Transkriptionen. Im Projekt wird ein Repositorium mit Portal aufgebaut und verschiedene Instrumente zum Anwendersupport getestet. Das Projekt kooperiert mit dem weiter unten beschriebenen Zentrum für germanistische Forschungsprimärdaten, sowie mit CLARIN-D.
- Das ebenso von der DFG geförderte **Zentrum für germanistische Forschungsprimärdaten**[47] arbeitet an dem Aufbau einer einheitlichen Infrastruktur für die Bereitstellung von Forschungsdaten der germanistischen Linguistik. Hierbei handelt es sich um Korpora, gesprochene Sprache, sowie elektronische Lexika. Neben verschiedenen Standards sollen auch Best-Practice-Richtlinien für Archivierung und Nachnutzung der Daten entwickelt werden. Im Projekt solle ein Portal entstehen, über das auf sich am Projekt beteiligende, bereits bestehende Repositorien zugegriffen werden kann. Die Zugriffskontrolle wird hierbei bei den Anbietern liegen. Eine zentrale Datenablage soll möglich sein. In einer späten Phase des Projektes soll dieses Portal um Korpora externer Einrichtungen erweitert werden. Das Projekt arbeitet, auf Betreiben des Förderers, mit dem Projekt „Mehrsprachigkeit und gesprochene Sprache“ zusammen. Weiter bestehen Kooperationen mit den DFG Projekten InFoLis (im Kapitel Sozial- und Wirtschaftswissenschaften beschrieben) und LAUDATIO.
- Das Projekt **LAUDATIO** arbeitet am Aufbau eines Repositoriums und der damit verbundenen Infrastruktur für Forschungsdaten in der historischen Linguistik (annotierte und komplex strukturierte Textkorpora). LAUDATIO soll umfangreiche Onlinesuche mit persistent referenzierbaren Suchergebnissen, die in einen Repository abgespeichert werden können, ermöglichen. Das Repositorium soll offen sein für Korpora jeglicher Herkunft sein und neben einem Webportal auch Zugriff durch eine API ermöglichen. Das Projekt arbeitet in Zusammenarbeit mit dem Computer- und Medienzentrum der HU Berlin.

[45] Vgl. Pempe, W. (2012) in: Neuroth, H. et al. (Hrsg.) (2012) *Langzeitarchivierung von Forschungsdaten – Eine Bestandsaufnahme, Universitätsverlag Göttingen* S. 140
[46] Projektwebseite http://www.corpora.uni-hamburg.de/lis/index.html
[47] Projektwebseite http://www.ids-mannheim.de/fi/projekte/lis.html

- Im INF Projekt **SFB 600 Fremdheit und Armut**, welches schon weiter oben beschrieben ist, wird vom Universitätsrechenzentrum Trier eine prototypische Langzeitarchivierungsumgebung entwickelt. Innerhalb des SFBs wird von den Wissenschaftlern und Wissenschaftlerinnen individuell entschieden, welche Dokumente in das Archiv übernommen werden sollen, und inwieweit die Daten für die Öffentlichkeit verfügbar sein werden. Gemäß den Vorgaben der DFG werden diese Forschungsdaten für 10 Jahre gespeichert. Das Archiv wird erst zum Abschluss des SFB (Ende 2012) online gehen. Eine Qualitätskontrolle findet während der Entwicklung der Plattform in Zusammenarbeit mit den Wissenschaftlern statt (beim Ingest). Darüber hinaus sind weitere Qualitätsprüfungen zunächst den Forschern überlassen.
- Im Rahmen von TELOTA arbeitet das von der DFG geförderte Projekt **Wissensspeicher – Daten geisteswissenschaftlicher Grundlagenforschung**[48] daran, die digitalen Ressourcen der BBAW erfassen und gebündelt anzubieten. Das Projekt konzentriert sich dabei auf Daten aus geisteswissenschaftlichen Projekten, welche mit Texten aus dem Altertum, dem Mittelalter und der Moderne arbeiten. Geplant sind ein erweiterter Zugriff über Metadaten, und eine linguistische Volltextsuche für Sprachen wie Arabisch, Latein oder Altgriechisch.
- Das Projekt **Yago-Naga**[49] betreibt Web und Text Mining auf Korpora, die einerseits explizit dokumentiert (und evtl. käuflich erworben) sind, z.B. das Archiv der New York Times, andererseits durch die Community erzeugt werden. Teilweise wird auch selbst das Internet systematisch durchsucht und indiziert mit einem Webcrawler. Die Rohdaten werden in jedem Fall konserviert. Die Datenmengen sind hierbei im niedrigen TB Bereich, da z.B. auf Bilder und Videos verzichtet wird. Eine Zusammenarbeit mit dem European Archive, dem europäischen Gegenstück zum Internet Archive, findet statt. Die Metadaten, die mit den Rohdaten verknüpft sind, werden erhalten und durch „added value data“ ergänzt. Alle Arbeitsschritte werden genau dokumentiert. Von einer funktionierenden Langzeitarchivierung kann auf Grund der technischen Entwicklung (Standards, etc.) nicht in jedem Fall ausgegangen werden. Archivierte Daten werden nicht weiter kuratiert, veröffentlichte Daten (Yago) werden jedoch weiter gepflegt. In Yago werden nur freie Datensätze weiterveröffentlicht.

### *Zugangsdomäne*

Im Rahmen der Nachnutzung sind in den Geisteswissenschaften noch keine allgemein akzeptierten Nutzungsbedingungen etabliert. Für die Wissenschaftler sind somit die Konditionen, unter denen auf Datenbestände zugegriffen werden kann, unübersichtlich. Insbesondere wird die Frage, wie die Daten vom potentiellen Nutzer gefunden werden und wie eine Vernetzung mit anderen Repositorien stattfindet, als drängend angesehen. In den Geisteswissenschaften erfolgt eine Nachnutzung von Forschungsdaten auch durch eine fach-

[48] Webportal http://www.bbaw.de/telota/ressourcen/digitaler-wissensspeicher
[49] Projektwebseite http://www.mpi-inf.mpg.de/yago-naga/

fremde Öffentlichkeit (Journalisten, Schulen, etc.) und damit ist auch eine Überlappung des Datenmanagements mit klassischer Öffentlichkeitsarbeit denkbar.

### Zusammenfassung und Fazit

In den Geisteswissenschaften und der Psycholinguistik hat die Einführung moderner informationstechnischer Infrastrukturen nicht nur die Grundlagen eines strukturierten digitalen Forschungsdatenmanagements gelegt, sondern gleich ein neues Forschungsfeld, die e-Humanities, begründet. Etablierte Akteure, wie TextGrid für die Literaturwissenschaften oder CLARIN in der Linguistik, befinden sich bereits in einer Phase der Konsolidierung. Dabei erweitern sie ihr institutionelles Netzwerk, auch im Europäischen Kontext, und versuchen auch angrenzende Subdisziplinen miteinzubeziehen.

Die Benutzung von Metadaten und die Notwendigkeit von gemeinsamen Formaten werden in den Communities verstanden. Die mag auch an der Nähe zu den traditionellen Wissensinstitutionen, Bibliotheken und Archive liegen. Als problematisch wird die Rechtsunsicherheit in Bezug auf die Urheber und Verwertungsrechte der der Forschung zugrundeliegenden Texte und anderen Medien beschrieben. Dies stellt insbesondere einen Hinderungsgrund für die Weitergabe von Forschungsergebnissen und eine damit verbundene Nachnutzung dar.

Die hohe Anzahl der geförderten Projekte zeigt einen hohen Bedarf an Infrastruktureller Unterstützung der sonst als eher weniger IT-affin einzuschätzenden Disziplinen. Es fällt weiter auf, dass die technische Komplexität der Projektvorhaben sehr unterschiedlich ist und bei den INF Projekten beispielsweise von dem Aufsetzen und dem Betrieb eines Content-Management-Systems bis zu komplexen virtuellen Forschungsumgebungen mit Langzeitarchivierungskomponente und Publikationsmöglichkeit reicht.

## Altertumswissenschaften

### Wichtige Akteure

Die Altertumswissenschaften in Deutschland zeichnen sich durch eine breite institutionelle Verankerung aus. Besonders hervorzuheben ist hierbei das **Deutsche Archäologische Institut (DAI)**[50], welches weltweit an Archäologischen Forschungsprojekten beteiligt ist. Das DAI pflegt zusammen mit dem **Archäologischen Institut der Universität zu Köln** die zentrale Objektdatenbank **Arachne**[51], welche als Werkzeug zur Internetrecherche und zum Auffinden von Datensätzen und Objekten für die Altertumswissenschaften konzipiert ist.

Neben dem DAI existieren **Landesdenkmalämter**, deren Aufgabe die Erforschung der Kulturgüter in den jeweiligen Bundesländern ist, sowie die Akademien, die sich in eher langfristigen Projekten engagieren. Weiter ist altertumswissenschaftliche Forschung an **Universitäten** sowie die von der DFG geförderte **Forschungsverbünde** zu finden.

[50] Webseite http://www.dainst.org
[51] Webseite http://arachne.uni-koeln.de

## Domänenübergreifende Faktoren

### *Art und Menge der Daten*

Altertumswissenschaftliche Forschung basiert heute mehr und mehr auf der Arbeit mit digitalen Objekten und Raumdaten anstelle der klassischen Arbeit mit physischen Objekten. Dies umfasst unter anderem GIS Analysen, statistische Untersuchungen, sowie 2D und 3D Rekonstruktionen von archäologischen Stätten. Insbesondere die Dokumentation von Ausgrabungen sind einmalige Primärquellen, da im Verlauf der Ausgrabung der Ausgangszustand der Stätte nach und nach zerstört wird. Außerdem sind für die Altertumswissenschaften auch naturwissenschaftliche Daten, beispielsweise aus der Geologie (Bohrkerne, Fernerkundung, Vermessung) oder der Biologie (Anthropologie, Zoologie) interessant. Wie auch in vielen anderen Disziplinen wachsen die Datenvolumina in den Altertumswissenschaften rapide an. Bei archäologischen Großprojekten fallen inzwischen Daten im Terabyte-Bereich an. Kleinere Projekte bleiben bisweilen aber im einstelligen Gigabyte-Bereich. Besonders datenintensiv sind Retro-Digitalisierungsprojekte wie ARACHNE mit Volumina von mehreren hundert Terabyte.

### *Formate*

Das DAI hat im Jahr 2011 einen *Leitfaden für die Anwendung von IT in der archäologischen Forschung* herausgegeben, der die gebräuchlichen Dateiformate der Disziplin zusammenfasst. Er nennt die Anwendungsfälle, welche durch eine Aufzählung der möglichen Datenformate abgedeckt sind[52]. Das Dokument weist jedoch die Schwäche auf, dass es sich vor allem auf die technischen Aspekte (z. B. Dateiformate) konzentriert, wohin ein für die Disziplin relevantes System von Metadaten nicht erörtert wird. Unter anderem wird auch auf verschiedene internationale Standards verwiesen (u.a. CIDOC CRM[53]). Das Dokument ist ein Resultat einer DFG-Arbeitsgruppe, die von 2008-2011 tätig war[54].

### *Metadaten*

Metadaten sind in den Altertumswissenschaften ausgiebig diskutiert. Es existiert daher auch eine Vielzahl von Vorschlägen zur Standardisierung. Bei den Archäologischen Landesämtern existieren auch schon strikte Vorgaben zum Metadaten Vokabular. Ein bundesweit einheitliches Vokabular ist jedoch noch nicht gefunden und es existiert auch keine zentrale Struktur, um dies zu forcieren. Die praktische Erprobung solcher Metadaten hat keine zufriedenstellenden Resultate bislang ergeben[55].

---

[52] Vgl. Tabelle. S. 17 in *Leitfaden für die Anwendung von IT in der archäologischen Forschung* http://www.dainst.org/de/project/it-leitfaden?ft=all, online.
[53] CIDOC Conceptual Reference Model, Webseite http://www.cidoc-crm.org
[54] Vgl. http://www.ianus-fdz.de/projects/zentr-dig-arch/wiki/Vorarbeiten_DAI, online.
[55] Vgl. Workshopdokumentation, http://www.dainst.org/de/event/report-vernetzte-datenwelten-ein-workshop-zur-umsetzung-von-cidoc-crm?ft=all, online.

#### *Identifikatoren*

Die Nutzung von PIDs (persistenten Identifikatoren) befindet sich noch am Anfang. Dies liegt unter anderem in der noch nicht ausgebildeten Kultur der Publikation der Daten bzw. dem mangelnden Bewusstsein für Nachnutzung begründet. Beim Deutschen Archäologischen Institut bzw. den Landesdenkmalämtern sind PIDs noch nicht in Verwendung. In der schon genannten ARACHNE werden zurzeit Identifikatoren implementiert.

#### *Rechtliche Aspekte*

Ein besonderer rechtlicher Aspekt in den Altertumswissenschaften ist der Schutz vor Raubgräbern, der gegen eine unbeschränkte Veröffentlichung von Forschungsdaten (z.B. Geokoordinaten) spricht. Auch die kulturelle oder religiöse Bedeutung, die untersuchte Objekte in anderen Kulturen haben können, kann gegen eine Datenpublikation sprechen. Auch urheberrechtliche Einschränkungen (Fotographien) oder kommerzielle Interessen (Satellitenbilder) können für altertumswissenschaftliche Daten relevant sein. In der Regel sind die Rechte an den Daten in Deutschland klar geregelt, unterscheiden sich jedoch von Bundesland zu Bundesland.

### *Finanzierung*

Finanziert wird Altertumswissenschaftliche Forschung in Deutschland einerseits durch die beteiligten Institutionen, andererseits durch die DFG. Das DAI ist in die Haushaltspläne des Auswärtigen Amtes eingebunden und Teil der Auswärtigen Kulturpolitik. Die archäologischen Landesämter sind wiederum über die Länder finanziert. Darüber hinaus finden sich bei anderen Akteuren (z.B. Museen) noch andere Geldgeber.

### *Publikationen*

Printmedien haben einen sehr hohen Stellenwert in den Altertumswissenschaften. Hierbei lassen sich interpretatorische Publikationen (Thesen mit Abbildungen und Referenzen) und dokumentarische Publikationen (mit Fotos bzw. Zeichnungen von Objekten) unterscheiden. Darüber hinaus sind die Forschungsobjekte nur über Museen zugänglich. Reine Datenpublikationen sind meist Aufbereitungen bereits vorhandener Daten. Sie werden aus diesem Grund auch von der Community im Allgemeinen nicht als wertvoll (und als nicht zu fördern) angesehen.

## Umgang mit Forschungsdaten in den vier Domänen

### *Private Domäne*

Der allgemeine Umgang mit Forschungsdaten durch die Wissenschaftler wird als äußerst heterogen beschrieben. Die Notwendigkeit von standardisierten Werkzeugen und Formaten wird in der Community zwar akzeptiert, die zu Ihrer Etablierung notwendigen Schritte werden aber nicht unternommen. Von Vertretern der Community wird hier der Anglo-Ameri-

kanische Bereich um Jahrzehnte fortschrittlicher wahrgenommen. Dies wird unter anderem mit Defiziten in der Ausbildung betreffend den Umgang mit Forschungsdaten begründet.[56]

### *Gruppendomäne*

Kollaboratives Arbeiten mit Forschungsdaten ist in der Regel über Webinterfaces zu Datenbanken, verbunden mit dem entsprechenden Zugriff auf Rohdaten per Fileserver, realisiert. Beispiele hierfür sind das schon angesprochene ARACHNE Projekt und das **iDAI.field**[57] Projekt des DAI. Es existieren auch Ansätze der Nutzung von Semantic-Web-Technologien, beispielsweise im Rahmen von **CLAROS**[58] sowie von Web-GIS.

#### Untersuchte Projekte

- Im DFG Projekt **IANUS - Forschungsdatenzentrum Archäologie & Altertumswissenschaften**[59] soll eine Bestands- und Bedarfsanalyse der bereits vorhandenen Ansätze zum Forschungsdatenmanagement in den Altertumswissenschaften vorgenommen werden. Zusätzlich werden Anwendungsszenarien, sog. Testbeds, definiert, welche in Form von kleineren Softwareprodukten der Community zur Verfügung gestellt werden. Dies soll unter anderem die Akzeptanz des Projektes seitens der Community steigern. In einer zukünftigen, noch zu beantragenden Projektphase soll dann ein Geschäftsmodell für ein Kompetenzzentrum und ein Stufenplan zu dessen Umsetzung entwickelt werden.

### *Dauerhafte Domäne*

Auch das Bewusstsein zur Langzeitarchivierung ist höchst unterschiedlich ausgeprägt. In den Akademien existieren, nicht zuletzt aufgrund der längeren Projekt- und damit auch Finanzierungszeiträume schon Ansätze zur Archivierung. In den Landesdenkmalämtern ist kein Workflow zur Archivierung bekannt. Das DAI arbeitet noch mit einzelnen, pro Projekt getrennten Datensammlungen. Im Vereinigten Königreich ist hier die Entwicklung schon weiter und es sind bereits dezidierte Forschungsdaten-Repositorien in Betrieb.

### *Nachnutzung*

#### Untersuchte Projekte

- Das von der DFG geförderte Projekt **OpenInfRA**[60] arbeitet am Aufbau eines webbasierten Informationssystems welches für archäologische Forschung und Bauforschung kostenfrei zur Verfügung gestellt wird. Das System soll Recherchefunktionen bereitstellen und die zeitnahe digitale Publikation von Forschungsdaten ermöglichen.

---

[56] Vgl. Dally, O., Fless, F., Förtsch, R. (2012) in: Neuroth, H. et al. (Hrsg.) (2012) *Langzeitarchivierung von Forschungsdaten – Eine Bestandsaufnahme*, Universitätsverlag Göttingen S. 167ff
[57] Projektwebseite http://www.dainst.org/de/project/idaifield?ft=all
[58] Projektwebseite http://www.clarosnet.org/XDB/ASP/clarosHome/
[59] Projektwebseite http://www.ianus-fdz.de
[60] Projektwebseite http://www.tu-cottbus.de/projekte/de/openinfra/

### Zusammenfassung und Fazit

Innerhalb der Altertumswissenschaften besitzt das DAI durch seinen Status als Bundesanstalt eine herausgehobene Rolle. Das Projekt IANUS, welches konzeptionell am Forschungsdatenmanagement in dieser Disziplin arbeitet, ist vom DAI initiiert. In diesem Projekt sind besonders die Zentrierung auf die Nutzerinteressen und die Entwicklung von Anwendungsszenarien zur Steigerung der Akzeptanz der Community zu begrüßen.

Dadurch, dass neben der üblichen Forschungsfinanzierung über DFG, BMBF bzw. den entsprechenden Landesministerien auch das Auswärtige Amt (als Träger des DAI), das Umweltministerium (verantwortlich für den Denkmalschutz) und diverse kleinere Strukturen eine Rolle spielen, sind die Förderungsstrukturen deutlich komplexer als in anderen Disziplinen. Potentielle kooperative Strukturen müssen sich mit diesem Problem auseinandersetzen.

Eine einzigartige Einschränkung der Weitergabe von altertumswissenschaftlichen Forschungsdaten besteht im Schutz der Ausgrabungsstätten vor Raubgräbern. Für die Altertumswissenschaften lassen sich im Bereich der Forschungsdaten Überschneidungen zu den Geowissenschaften bei Archivierung von physischen Objekten und allgemein durch Verwendung von Geodaten feststellen. Abschließend ist es positiv zu bewerten, dass in den Altertumswissenschaften begonnen wird, die vorhandene Kultur von Archiven und Sammlungen in die digitale Welt zu übertragen.

## Sozial- und Wirtschaftswissenschaften

### Wichtige Akteure

Der Rat für **Sozial- und Wirtschaftsdaten (RatSWD)**[61] ist ein unabhängiges, vom BMBF berufenes Beratungsgremium, welches aus Forschern und Forscherinnen der empirischen Sozialwissenschaften, sowie Vertretern und Vertreterinnen von Datenproduzenten gebildet wird und die Aufgabe hat, durch Beratung und Konzeption die weitere Entwicklung der Forschungsdateninfrastruktur in den empirischen Sozialwissenschaften zu fördern. Insbesondere die Einrichtung und kontinuierliche Evaluation von Datenservicezentren wird hierbei forciert.[62] Die in der qualitativen Sozialforschung tätigen Forscherinnen und Forscher sind vor allem in der **Deutschen Gesellschaft für Soziologie (DGS)**[63] organisiert.

Das aus der Gesellschaft Sozialwissenschaftlicher Infrastruktureinrichtungen hervorgegangene **Leibniz-Institut für Sozialwissenschaften (GESIS)**[64] sieht sich als bedeutendste sozialwissenschaftliche Infrastruktureinrichtung im deutschsprachigen Raum. Neben diversen Angeboten, die den ganzen Lebenszyklus von Forschungsdaten abdecken sollen, forscht GESIS auch im Bereich des Umgangs mit Forschungsdaten.

[61] Webseite http://www.ratswd.de
[62] Vgl. Beschreibung auf der Webseite des RatSWD http://www.ratswd.de/rat/aufgaben.php, online.
[63] Vgl, Beschreibung auf der Webseite der DGS http://www.soziologie.de/de/die-dgs/ueber-die-dgs.html, online.
[64] Webseite http://www.gesis.org

Die Datensammlungen des **statistischen Bundesamts** und der **statistischen Landesämter, sowie des IAB (Bundesanstalt für Arbeit) und des DRV-Bund** sind insbesondere im Bereich der Sozial- und Wirtschaftswissenschaften relevant. Besondere Bedeutung kommt hierbei den Daten der Volkszählungen (**Makrozensus**) und des **Mikrozensus**, einer ergänzenden Erhebung, die jedes Jahr in 1% der deutschen Haushalte durchgeführt wird, zu.

Im Kontext des Datenmanagements in den Sozialwissenschaften besonders interessant sind sog. Surveys. Hierbei handelt es sich um methodisch aufwändige Erhebungskampagnen, die dezidiert auf die Sekundärnutzung durch andere Forscherinnen und Forscher ausgelegt sind. Beim **Sozioökonomischen Panel (SOEP)**[65] handelt es um eine repräsentative Wiederholungsbefragung, die unter anderem Daten über Einkommen, Erwerbstätigkeit, Bildung oder Gesundheit erhebt. Der **Survey of Health, Ageing and Retirement in Europe (SHARE)**[66] ist ein europaweites Projekt, welches über Erhebungen bei Menschen über 50 Jahren die Prozesse des Älterwerdens untersucht. Seit März 2011 hat SHARE als erstes Projekt den Status eines European Research Infrastructure Consortium (ERIC). Das Beziehungs- und Familienpanel **pairfam**[67] untersucht die Partner- und Lebensverhältnisse in Deutschland.

Die wirtschaftswissenschaftlichen Datensammlungen und -provider sind teilweise die gleichen wie für die Sozialwissenschaften (**Statistische Ämter, Institut für Arbeitsmarkt- und Berufsforschung**[68]). Darüber hinaus besitzen auch **Weltbank**, diverse UN-Organisationen, Banken und Wirtschaftsverbände Archive und es existiert eine Vielzahl kommerzieller Datensammlungen.

Die Deutsche **Zentralbibliothek für Wirtschaftswissenschaften – Leibniz Informationszentrum Wirtschaft (ZBW)** bietet traditionelle Bibliotheksdienste für Wirtschaftswissenschaftler an. Es ist geplant, diese Services um Forschungsprimärdaten zu erweitern. Die ZBW arbeitet bei der Langzeitarchivierung mit der GESIS und dem RatSWD zusammen.

Abgesehen von diesen Projekten ist sind die Sozial- und Wirtschaftswissenschaften an den **Universitäten** als mittelgroßes Fach vertreten. Die Forschungsstrukturen sind jedoch eher kleinteilig.[69]

## Domänenübergreifende Faktoren

### *Art und Menge der Daten*

In den Sozialwissenschaften ist zunächst zwischen den Daten der empirischen Sozialforschung, wie z.B. Umfragen, und der qualitativen Sozialforschung, wie Ton- und Videoaufnahmen und Interviews zu unterscheiden. Verglichen mit den Daten anderer Wissenschaftsdisziplinen haben sozialwissenschaftliche Forschungsdaten ein eher kleines Volumen. Dies reicht von wenigen MB für Kleinstatistiken bis in den TB Bereich für die Archive der GESIS

[65] Webseite http://www.diw.de/soep
[66] Webseite http://www.share-project.org
[67] Panel Analysis of Intimate Relationships and Family Dynamics, Webseite http://www.pairfam.de
[68] Webseite http://www.iab.de
[69] Vgl. Quand, M., Mauer, R. (2012) in: Neuroth, H. et al. (Hrsg.) (2012) *Langzeitarchivierung von Forschungsdaten – Eine Bestandsaufnahme*, Universitätsverlag Göttingen S. 62

oder die Zensus-Daten beim Statistischen Bundesamt. Gerade bei den Daten der qualitativen Forschung steigt der benötigte Speicherplatz aber rapide. In den Sozialwissenschaften sind Forschungsdaten in der Regel nicht reproduzierbar und haben daher auch immer eine historisch beschreibende Funktion.[70]

### *Formate*

Sozial- und Wirtschaftswissenschaftliche Forschungsdaten liegen entweder in Dokumenten oder in den nativen Formaten z.B. der verwendeten Softwarepakete vor. In der empirischen Sozialforschung sind diese in der Regel proprietär, beispielsweise die Formate der Softwarepakete SPSS, Stata, oder SAS. In letzter Zeit wird verstärkt freie Software wie R[71] verwendet. Hinzukommen die kommerziellen Datensammlungen, welche Datenbank-Abfragen in verschiedenen Formaten anbieten usw.

Bei Forschungsdaten der qualitativen Sozialforschung handelt es sich in der Regel um textuelle Quellen, bzw. Audio- und Videodateien. Informationen über Standardformate hierfür sind derzeit nicht erhältlich bzw. nicht öffentlich dokumentiert.[72]

### *Metadaten*

Die systematische Nutzung von Metadaten ist in den Sozialwissenschaften bereits etabliert. Es werden sowohl nicht-standardisierte *Code-Books*, als auch Standards wie DDI[73], SDMX[74] und DataCite[75] verwendet. Daten und Metadaten werden, zumindest beim Statistischen Bundesamt, physisch zusammen gespeichert. Für die Wirtschaftswissenschaften sind keine Metadatenstandards bekannt, die über die sozialwissenschaftlichen Ansätze hinausgehen. Gleiches gilt für die qualitative Sozialforschung.

### Untersuchte Projekte

- Das von der DFG geförderte **da|ra**[76] Projekt, arbeitet an einem Nachweis- und Registrierungssystems für die Sozialwissenschaften in Deutschland. Mit dem Abschluss des Projektes soll dieses System für die Community geöffnet werden.

### *Identifikatoren*

Die Weiterentwicklung des Forschungsdatenmanagements in den Sozialwissenschaften resultiert auch in der systematischen Verwendung von persistenten Identifikatoren in den diversen Archiven. Beim statistischen Bundesamt befindet sich die Nutzung von persistenten Identifikatoren befindet sich zurzeit im Aufbau. Bei GESIS wird im Rahmen von da|ra auch die Vergabe von DOI für die registrierten Daten organisiert.

---

[70] Vgl. Quand, M., Mauer, R. (2012) in: Neuroth, H. et al. (Hrsg.) (2012) *Langzeitarchivierung von Forschungsdaten – Eine Bestandsaufnahme*, Universitätsverlag Göttingen S. 62
[71] Webseite http://www.r-project.org
[72] Die DGS weist für 2011 (ca. 175) Projekte der Empirischen Sozialforschung aus, welche aus DFG Mitteln gefördert werden, keine der qualitativen Sozialforschung. (http://www.soziologie.de/index.php?id=310, (03-2013)
[73] Data Documentation Iniative, Webseite http://www.ddialliance.org
[74] Statistical Data and Metadata eXchange, Webseite http://sdmx.org
[75] Projektwebseite http://datacite.org
[76] Projektwebseite http://www.da-ra.de

### *Rechtliche Aspekte*

In den Sozialwissenschaften spielen sowohl urheber- als auch datenschutzrechtliche Aspekte eine herausgehobene Rolle. Oft muss eine Datennutzung durch den Forscher erst beim Datenzentrum beantragt werden. Auch kann eine Drittnutzung untersagt sein. Außerdem können offene Datenzentren keine Kopien von anderen Datenarchiven bereithalten, da Datensätze auf Grund datenschutzrechtlicher Bestimmungen in der Regel nicht weitergegeben oder miteinander verknüpft werden können und in der erhebenden Institution gespeichert werden. Für besonders sensible Daten ist die Dateneinsicht nur in den Räumlichkeiten des Datenhalters möglich. In Zukunft soll dies auch über geschützte Verbindungen über das Internet möglich sein. Es erweist sich in diesem Kontext als problematisch, dass, auch aufgrund der föderalen Struktur der deutschen Wissenschaftslandschaft, in Deutschland keine zentrale Autorisierungsstelle existiert.

### *Finanzierung*

Neben der traditionellen Finanzierung über das BMBF und die DFG wird der Aufbau von Archiven in den Sozial- und Wirtschaftswissenschaften in erheblichem Umfang durch Mittel aus Behörden, Ämter und andere staatliche Stellen finanziert. Neben dem schon genannten Statistischen Bundesamt und entsprechenden Landesämtern sind hier besonders die **Deutsche Rentenversicherung** und die **Bundesagentur für Arbeit** aktiv. Es existieren auch eine Reihe von privatwirtschaftlichen Forschungsinstituten.

### *Publikationen*

Wie auch in den Geisteswissenschaften, sind klassischen Publikationen in den den Sozial- und Wirtschaftswissenschaften eigenständige Forschungsobjekte und daher sind auch mehrere Projekte zum Forschungsdatenmanagement im bibliotheksnahen Bereich tätig.

## Untersuchte Projekte

- Im Projekt **InFoLis**[77] wird an der Verknüpfung des Datenbestandskatalogs von GESIS mit dem Literaturbestand im Recherchesystem der UB Mannheim gearbeitet. Das Projekt untersucht verschiedene technische Ansätze, setzt sie technisch um und spricht anschließend Empfehlungen aus.
- An der ZBW Projekt angesiedelt ist das Projekt **EDaWaX**[78], welches ein Datenarchiv für Fachzeitschriften in den Wirtschaftswissenschaften, im Rahmen eines ganzheitlichen Ansatzes, aufbaut. Insbesondere fehlende Anreizstrukturen im Bereich der Publikation von Forschungsdaten sollen adressiert werden. In Bezug auf die Datenschutzproblematik arbeitet das EDaWaX-Projekt an einem Rechtsgutachten. Die Arbeit umfasst neben der Softwareentwicklung und der Entwicklung von Metadaten auch die Untersuchung der Data-Policies der verschiedenen Journale und für das Hosting in Sozialwissenschaftlichen Datenzentren. Ein Pilot-Archiv wird von Au-

[77] Projektwebseite http://www.gesis.org/forschung/drittmittelprojekte/projektuebersicht-drittmittel/infolis/
[78] European Data Watch Extended , Projektwebseite http://www.edawax.de

gust 2011 an für zwei Jahre für das Journal of Applied Social Sciences in Betrieb sein. EDaWaX ist eine Kollaboration zwischen dem RatSWD, dem an der LMU München angesiedelten INNO-tec Institut, und der ZBW.

## Umgang mit Forschungsdaten in den vier Domänen

### *Private Domäne*

Die Projektstrukturen in den Sozialwissenschaften sind dominiert von kleinen Projekten mit nur wenigen Mitarbeitern. Dies spiegelt sich auch in den Publikationen wieder, die oft von Einzelautoren oder auch von kleinen Forschergruppen verfasst werden. Oft werden Forschungsdaten nicht direkt für ein spezielles Forschungsvorhaben erhoben, sondern als Mehrthemenumfragen oder Surveys, und dann weiterverwertet. Zum Teil wird die Erhebung der Daten auch an kommerzielle Dienstleister ausgelagert. Eine besondere Stellung in den Sozialwissenschaften haben amtlich erhobene Daten, welche beispielsweise von Ministerien oder dem Statistischen Bundesamt erhobenen werden, sowie die Daten der Sozialämter, des Arbeitsamtes etc. Datenzentren wie GESIS spielen für die alltägliche Forschungsarbeit bereits eine hervorgehobene Rolle.

#### Untersuchte Projekte

- Das von der DFG geförderte Projekt **MISSY 3.0**[79] arbeitet an der Weiterentwicklung des Mikrodateninformationssystems MISSY[80]. MISSY wurde bei GESIS, mit teilweiser Finanzierung des BMBF, für die Forschungsarbeiten um den Mikrozensus entwickelt. Die Erweiterung soll das Anwendungsgebiet auf weitere Mikrodaten erweitern. Die zentralen Dienste von MISSY sollen als Open-Source-Software zugänglich gemacht werden.
- Am ZBW werden zusammen mit dem Institut für Informatik der Christian-Albrechts Universität Kiel im Projekt **MaWiFo**[81] die technischen Gestaltungsmöglichkeiten für Infrastrukturlösungen im Bereich der wirtschaftswissenschaftlichen Forschungsdaten, insbesondere am Standort Kiel, untersucht.

### *Gruppen Domäne*

Wie auch in anderen Wissenschaftsdisziplinen werden in den Sozialwissenschaften größere Kollaborationen immer wichtiger. Hierbei handelt es sich beispielsweise um die schon erwähnten Surveys, wie SOEP, SHARE und PAIRFAM, welche in jüngster Zeit immer aufwendiger werden und daher neue, angepasste Organisationsstrukturen erfordern.

[79] Forschungsbasierte Metadaten für amtliche Erhebungen: Ausbau von MISSY, Projektwebseite http://www.gesis.org/forschung/drittmittelprojekte/projektuebersicht-drittmittel/missy-30/
[80] Webportal http://www.gesis.org/missy
[81] Management wirtschaftswissenschaftlicher Forschungsdaten, Projektwebseite http://mawifo.zbw.eu

### Untersuchte Projekte

- Im **SFB 884 Politische Ökonomie von Reformen** wird im Projekt Z1[82] das zentrale Datenzentrum des SFBs realisiert. Neben der Infrastruktur erstellt das Projekt eine Internet-Panel-Umfrage zur Erhebung individueller Mikrodaten. Diese Datensätze bilden eine der Grundlagen für die Forschung des gesamten SFB.
- Das INF Projekt[83] im **SFB 882 Von Heterogenitäten zu Ungleichheiten** arbeitet am Aufbau einer Virtuellen Forschungsumgebung für den SFB. Diese besteht aus einer konventionellen Arbeitsplattform, die die Forscher bei der täglichen Arbeit unterstützt, einer Forschungsdatenplattform, die Dienste zur Archivierung und zur Nachnutzung der Datensätze zur Verfügung stellt und einem Schnittstellenmodul, welches eine Anbindung an externe Ressourcen (SOEP, FDZ[84], etc.) herstellt. Weiter entwickelt es Standards für die Dokumentation der anfallenden Forschungsdaten, erarbeitet Anonymisierungskonzepte und übernimmt Beratungsfunktionen gegenüber den Forscherinnen und Forschern des SFB.
- Im **SFB 649 Economic Risk** wird im Projekt INF[85] das Research Data Center (RDC) aufgebaut. Es soll die Grundlage für die empirische und theoretische (d.h. numerische) Forschung im SFB bilden. Es dient als Austauschplattform für Software und numerische Algorithmen und verwaltet den Zugang zu Diskussionspapieren bzw. Veröffentlichungen. Das RDC steht auch anderen DFG-finanzierten Projekten offen. Ziel ist die Schaffung eine Kommunikationsplattform für empirische Risikoforschung und Wirtschaftswissenschaften für räumlich getrennte Arbeitsgruppen.

## *Dauerhafte Domäne*

Datenzentren, deren Aufgraben die Langzeitarchivierung, aber auch die Bereitstellung von Forschungsdaten für die nationale und z. T. auch internationale Community umfassen, sind in der empirischen Sozialforschung schon seit den 1960er Jahren etabliert. Dies umfasst nationale Einrichtungen wie GESIS und das Statistische Bundesamt, aber auch kleinere Datensammlungen, welche meist bei den Auftraggebern der jeweiligen Forschungsprojekte angesiedelt sind. Hierbei handelt es sich meist um Ämter und Behörden, Ministerien, Stiftungen, aber auch kommerzielle Dienstleister. Für die qualitative Sozialforschung existieren zurzeit noch keine Datenzentren.

### Untersuchte Projekte

- Das von der DFG geförderte Projekt **Serviceeinrichtung qualitative Sozialforschung** hat das Ziel die Speicherung, Referenzierung und Verfügbarkeit von Forschungsprimärdaten aus der qualitativen Sozialforschung zu initiieren und konzeptuell zu gestalten. Hierzu wird auf internationale Standards gesetzt (Metadaten nach DDI,

[82] Projektwebseite http://reforms.uni-mannheim.de/projekte/datenzentrum/index.html
[83] Projektwebseite http://www.sfb882.uni-bielefeld.de/de/projects/inf
[84] Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder, Webportal http://www.forschungsdatenzentrum.de
[85] Projektwebseite http://sfb649.wiwi.hu-berlin.de/fedc/index.php

ESDS Qualidata[86]). Dienste werden in Form von WSDLV-Soap-Webservices im Rahmen des Referenzmodels der Workflow Management Coalition bereitgestellt. Die geschieht in Zusammenarbeit mit GESIS.

- Im Bereich der Langzeitarchivierung wird an der ZBW im Projekt **Digitale Reichsstatistik**[87] an der Digitalisierung historischer Statistiken (1873 bis 1883) als PDF oder Tabellenkalkulationsdaten gearbeitet.

### *Zugangsdomäne*

Das Angebot von Daten zur Nachnutzung ist von der Art der Daten abhängig. Oft bestimmen die oben beschriebenen rechtlichen Rahmenbedingungen, insbesondere der Datenschutz, ob eine Veröffentlichung in Frage kommt. Bei den Daten des Bundesamts für Statistik ist es von den Policies der jeweiligen Quellen abhängig, ob die Daten freigegeben werden, die Tendenz geht hier zu Open Data. Auch in den Wirtschaftswissenschaften ist in der Regel die Offenheit der Daten unterschiedlich. Daten von privatwirtschaftlichen Firmen (z.B. Thomson Reuters) sind nicht offen und der Zugang ist teilweise sehr teuer (bis zu 10000 Euro/Jahr).

#### Untersuchte Projekte

- Die Verknüpfung verschiedener sozialwissenschaftlicher Datensätze (sog. Record-Linkage) wird in Deutschland dadurch erschwert, das Datensätze in der Regel nicht mit einem eindeutigen Identifikator (z.B. Sozialversicherungsnummer) versehen sind. Das von der DFG geförderte **Zentrum für Record-Linkage**[88] soll als überregionale und interdisziplinäre Institution durch Serviceleistung und Forschungsarbeiten die Zahl und Qualität dieser Verknüpfungen fördern.
- Von der DFG wird ferner im Projekt **Professionalisierung und Ausbau des Forschungsdatenzentrums Survey of Health, Ageing and Retirement in Europe (SHARE)** die Nutzbarkeit der Daten des SHARE Surveys durch verstärkte Nutzerbetreuung, Schulungen, Austauschprogramme und die Erstellung eines vereinfachten Datensatzes für die Ausbildung zu verbessern. Ferner werden SHARE Zusatzdateien, wie zusätzliche Biomarker und Daten der Deutschen Rentenversicherung, zur Verfügung gestellt.

#### Zusammenfassung und Fazit

Obwohl die Forschungslandschaft in den Sozial- und Wirtschaftswissenschaften eher kleinteilig organisiert ist, sind im Bereich der Forschungsdaten, insbesondere in der Dauerhaften Domäne, große Akteure wie beispielsweise GESIS etabliert. Auch nicht primär akademische Institutionen wie die Statistischen Ämter von Bund und Ländern sind hier von Bedeutung. Zusätzlich machen es Surveys wie SHARE oder SOEP notwendig, kollaborative Dateninfrastrukturen in der Gruppendomäne zu etablieren. Dementsprechend sind viele der geför-

[86] Projektwebseite http://www.esds.ac.uk/qualidata/
[87] Projektbeschreibung http://www.zbw.eu/ueber_uns/projekte/reichsstatistik.htm
[88] Projektbeschreibung http://www.uni-due.de/gesellschaftswissenschaften/profilschwerpunkt/record_linkages.shtml

derten Projekte auch in diesem Bereich angesiedelt. Im Bereich der Wirtschaftswissenschaften versucht die ZBW sich im Forschungsdatenmanagement zu etablieren. Die Forschungsdateninfrastruktur in den qualitativen Sozialwissenschaften befindet sich erst in einem frühen Entwicklungsstadium.

In den Sozial- und Wirtschaftswissenschaften ist die Benutzung von Metadatensystemen bereits etabliert. Ein Problem für Archivierung und Nachnutzung besteht eher in der Dominanz von proprietären Softwareprodukten und den damit verbundenen geschlossenen Formaten. Als ein weiteres Problem werden immer wieder die diversen datenschutzrechtlichen Probleme geschildert. Hier ist noch Forschungsbedarf gegeben. In diesem Kontext ist das Projekt EDAWAX zu erwähnen, welches versucht auch rechtliche Fragen zu beantworten.

Abschließend ist zu bemerken, dass das Forschungsdatenmanagement in den der Sozial- und Wirtschaftswissenschaften nicht nur das klassische akademische Umfeld betrifft, sondern auch, insbesondere durch die Open-Data Bewegung, für die breite Öffentlichkeit relevant wird. Dies gilt verstärkt für amtlich erhobenen Daten.

## Biodiversität

### Wichtige Akteure

An der Biodiversitätsforschung sind die traditionellen Biologie-Fachbereiche der Universitäten sowie Forschungsinstitute und naturkundliche Sammlungen an Museen, Zoos und Botanischen Gärten beteiligt. Ähnlich wie in der Medizin hat sich mit der Bio(diversitäts)-Informatik ein eigener Forschungszweig gebildet, der unter anderem disziplinspezifische Fragen des Forschungsdatenmanagements adressiert.

In diesem Kontext existiert seit 2001 die internationale **Global Biodiversity Information Facility (GBIF)**[89] welche den kostenfreien und offenen Zugang zu Biodiversitätsdaten fördern soll. Der deutsche **GBIF-D**[90], wird seit 2010 als Verbundprojekt durch das BMBF gefördert.

Seit 2006 werden von der DFG **Biodiversitäts-Exploratorien (BE)**[91] als Schwerpunktprogramm SPP 1374 gefördert. Hierbei handelt es sich um Programme von verschiedenen interdisziplinären Forschungsvorhaben, welche auf den gleichen Forschungsflächen stattfinden. Die Projektmitglieder verpflichten sich dabei zur frei zugänglichen Veröffentlichung der Daten und die werden fünf Jahre nach Erhebung veröffentlicht.

Mit der Diversity-Workbench der **Staatl. Nat. Wiss. Sammlungen Bayern** (SNSB)[92], BioCASE am **Botanischer Garten und Botanisches Museum Berlin-Dahlem** (BGBM)[93] und der weiter unten beschriebenen Plattform BExIS der BE existieren zum Teil sich ergänzende bzw. konkurrierende Services, die in verschiedenen Forschungsdaten-Repositorien Anwendung finden.

[89] Webseite http://www.gbif.org
[90] Global Biodiversity Information Facility, Webseite http://www.gbif.de
[91] Webseite des Schwerpunktprogramms http://www.biodiversity-exploratories.de
[92] Webseite http://diversityworkbench.net
[93] Webseite http://www.bgbm.org

## Domänenübergreifende Faktoren

### *Art und Menge der Daten*

Aufgrund der interdisziplinären Natur der Biodiversität findet sich dort auch eine Vielzahl von verschiedenen Datensätzen. Hierzu gehören Datensätze aus Felderhebungen und Feldversuchen, Laborauswertungen, Fotos, GIS Daten, Karten, Modelle, Annotationen von Naturhistorischen Sammlungsobjekten, mikrobiologische Daten, sowie Spektral- und Audiodateien.

### *Formate*

Felddaten werden oft mit Textverarbeitungsprogrammen wie Word oder Excel erfasst oder liegen in üblichen Image-Formaten vor, welche in jüngerer Zeit durch GPS Informationen ergänzt sind. In der Analytik werden vor allem die meist proprietären Formate der verwendeten Sequenzierer bzw. der genutzten Statistik-Programme verwendet. Auf Seiten der Repositorien und Datenzentren existiert es eine Vielzahl von genutzten Formaten und Standards. Als Austauschformate werden unter anderem ASCII-Tabellen, XML und teilweise proprietäre Formate für MS Access, Excel, Oracle DB unterstützt.

### *Metadaten*

Auch bei den Metadaten spiegelt sich die Vielfalt der verschiedenen Datentypen wieder. Es werden unter anderem die Metadatenstandards ABCD[94], Darwin Core, EML[95], GML[96], und PMML[97] verwendet. Hinzu kommen extensive Thesauri und Keyword-Sammlungen für spezifische Bereiche der Biodiversitätsforschung.

### *Identifikatoren*

Eine breit geführte Diskussion um die Verwendung von Identifikatoren wie z.B. von DOI wird zurzeit noch nicht disziplinweit geführt, wohl aber für einzelne Datenbanken. In der Taxonomie und vor allem in der Biodiversitätserfassung wird die Methode des DNA-Barcoding eingeführt (z.B. GBOL[98] , unterstützt vom BMBF). Hierbei wird eine Probe (d.h. ein Datensatz) mit einer DNA-Sequenz verknüpft. Unterschieden wird dabei nach verschiedenen DNA Markern, z.B. COI (für Tiere), ITS (Pilze), rbcL und matK (Pflanzen). Diese sind über BOLD[99] (nicht Open Access) oder GenBank[100] (Open Access) registriert.

### *Rechtliche Aspekte*

Auf Grund der möglichen Ortung von bedrohten Arten kann es notwendig sein, Daten durch eine Software zu verschleiern. Die Arten sind dann nicht mehr verortbar, für Analysezwecke

[94] Webseite http://wiki.tdwg.org/twiki/bin/view/ABCD/WebHome
[95] Ecological Metadata Language, Webseite http://knb.ecoinformatics.org/software/eml/
[96] Geography Markup Language, Webseite http://www.opengeospatial.org/standards/gml
[97] Predictive Model Markup Language, Webseite http://www.dmg.org/
[98] Webseite http://www.bolgermany.de/
[99] Barcode of Life Database, Webseite http://www.boldsystems.org
[100] Webseite http://www.ncbi.nlm.nih.gov/genbank/

sind die Daten jedoch weiterhin nutzbar. Eine Technik zu diesem Zweck wurde vom BGBM entwickelt wird, aber nicht von anderen Projekten genutzt. Als Grund hierfür wird der dafür benötigte Mehraufwand angegeben.

### *Finanzierung*

Die Publikation von digitalisierten Sammlungsdaten wird von vielen Institutionen als Teil ihrer Aufgabe gesehen und auch aus Gründen des Naturschutzes aus Bundes- und Landesmitteln mitgetragen. Darüber hinaus werden sowohl Digitalisierungsprojekte as auch Projekte der Bioinformatik von der DFG flankierend gefördert. Das GBIF-D[101] wird seit Ende 2010 als Verbundprojekt vom BMBF in der 2. Phase gefördert. BEs und BExIS werde für 3 Jahre von DFG durch ein Schwerpunktprogramm finanziert.

### *Publikationen*

Wie auch in vielen andern Disziplinen ist auch in der Biodiversitätsforschung die zitierbare Publikation von Daten nicht weit verbreitet. Einige Journale erlauben sog. *supporting online material*. Mit **Ecological Archives**[102] betreibt die ökologische Gesellschaft der USA ein eigenes Datenarchiv für ihre Zeitschriften. Das Archiv wird aber seitens der Community nur wenig genutzt.[103]

## Umgang mit Forschungsdaten in den vier Domänen

### *Private Domäne*

Eine von der DFG finanzierte Studie[104], hat die derzeitige Wahrnehmung und die Hinderungsgründe für die Nutzung der Angebote in den involvierten Fachgemeinschaften untersucht. Die Fragestellung war wesentlich auf die Bereitschaft zum Data-Sharing ausgerichtet, zeigt jedoch auch die derzeitigen Datenerfassungs- und Speicherungs-Methoden auf. Der wissenschaftliche Workflow zur Erhebung von Daten ist wesentlich durch die textliche Erfassung von Daten und Nutzung von Tabellenprogrammen wie Excel geprägt.

#### Untersuchte Projekte

- Das von der DFG geförderte Projekt **ReBIND**[105] arbeitet an einem Workflow zum effizienten Retten digitaler Daten aus kleinen Projekten wie Doktor- oder Diplomarbeiten. Die Daten werden hierzu in einer nativen XML-Datenbank in Verbindung mit den institutionellen Datenbanken gespeichert. Auf Grund des Bedarfs an Unterstützung und Schulungen im Umgang mit Forschungsdateninfrastruktur in der Biologischen Community, organisiert das Projekt ferner Summer Schools, Seminare, und Übungsgruppen in Video-Konferenzen.

---

[101] Global Biodiversity Information Facility, Webseite http://www.gbif.de/
[102] Webseite http://www.esapubs.org/archive/
[103] Vgl. Nieschulze, N., König-Ries, B. (2012) in: Neuroth, H. et al. (Hrsg.) (2012) *Langzeitarchivierung von Forschungsdaten – Eine Bestandsaufnahme*, Universitätsverlag Göttingen S. 218
[104] Vgl. Enke, N., et al. (2012) The user's view on biodiversity data sharing, Ecological Informatics, doi:10.1016/j.ecoinf.2012.03.004
[105] Projektwebseite http://rebind.bgbm.org

### *Gruppen Domäne*

In der Biodiversitätsforschung lassen sich mehrere Ansätze des kollaborativen Arbeitens mit Forschungsdaten unterscheiden. Zum einen finden sich Webportale, in denen die Daten von verschiedenen Institutionen oder Projekten transparent zusammengeführt, jedoch nicht zentral gespeichert werden. Beispiele hierfür sind **LifeWATCH**[106] oder **GEO BON**[107]. Weiter existieren Zusammenfassungen von Mess- und Monitoring-Flächen über Landes- und Institutsgrenzen hinweg wie **NEON**[108], **ILTER**[109] und **LTER-D**[110]. Auch hier liegen das Datenmanagement und die Archivierung in Verantwortung der einzelnen Partner. In verschiedenen Museen und Sammlungen werden (z.T. konkurrierende) Dienste angeboten. Ein Beispiel hierfür ist die schon erwähnte BioDiversity-Workbench. Auch in den bereits beschriebenen **Biodiversitäts-Exploratorien** sind als Plattformen für die Zusammenarbeit von verschiedenen, teils aus verschiedenen Disziplinen stammenden Arbeitsgruppen konzipiert. Dies umfasst auch einen regen Datenaustausch.[111]

#### Untersuchte Projekte

- Ein neuer Ansatz ist der Austausch von Multimedia-Dateien anstatt digitaler Objekte. Das von der DFG geförderte Projekt **Annosys** beschäftigt sich mit der Annotation dieser digitalen Objekte. FilteredPush[112] ist ein verwandtes Projekt aus den USA.

### *Dauerhafte Domäne*

Zwischen 2001 und 2008 existierte mit dem **National Biological Information Infrastructure** ein World Data Center für Ökologie und Biodiversität. Seine Aufgabe bestand hauptsächlich in der Referenzierung von verteilt vorliegenden, nicht standardisierte Daten. Aufgrund von Sparmaßnahmen der US-Amerikanischen Bundesregierung wurden die Services des NBII am 15. Januar 2012 eingestellt.

Sonstige zentrale LZA Dienste existieren nur für spezielle Daten aus der Mikrobiologie (**GenBank**[113], **TreeBase**[114]). Daneben existieren nur noch vereinzelt andere fragmentierte Ansätze.

Für die weiter oben beschriebenen Biodiversitäts-Exploratorien wird mit dem **Biodiversity Exploratory Information System (BExIS)** ein Zentrales Repository für die BE (ebenfalls im Rahmen des DFG SPP 1374) aufgebaut. Angesiedelt ist BExIS am MPI für Biogeochemie und an der Friedrich-Schiller-Universität Jena. Im Allgemeinen sind in BExIS nur die Metadaten

---

[106] Webseite http://www.lifewatch.com
[107] Group on Earth Observations, Webseite http://www.earthobservations.org/geobon.shtml
[108] National Ecological Observatory Network, Webseite http://www.neoninc.org
[109] International Long Term Ecological Research, Webseite http://www.ilternet.edu
[110] Long Term Ecological Reserach, Webseite http://www.ufz.de/lter-d/
[111] Vgl. Nieschulze, N., König-Ries, B. (2012) in: Neuroth, H. et al. (Hrsg.) (2012) *Langzeitarchivierung von Forschungsdaten – Eine Bestandsaufnahme*, Universitätsverlag Göttingen S. 216
[112] Webseite http://etaxonomy.org/mw/FilteredPush
[113] Webportal http://www.ncbi.nlm.nih.gov/genbank/
[114] Webportal http://treebase.org

frei zugänglich, während für die Primärdaten eine Karenzzeit von 5 Jahren vorgesehen ist. Die Daten können aber auch vom Primärforscher manuell freigestellt werden.

### Untersuchte Projekte

- Im von der DFG geförderten Projekt **BExIS++ - Modularisierung und Skalierung der BExIS Experimentdatenhaltungsplattform für die Umweltsystemforschung** wird die in BExIS verwendete Software um die Unterstützung von unterschiedlichen Metadatenstandards, die direkte Anbindung verschiedener Datenproduzenten, sowie erweiterte Analysewerkzeuge erweitert. Weiter soll BExIS in ein modulares System, welches für neue Projekte einfach adaptiert werden kann umgewandelt werden. Gleichberechtigt mit der Arbeit an der Software steht die Einführung einer Lehrveranstaltung zur Experimentdatenhaltung.

## *Zugangsdomäne*

Die Nachnutzung von Forschungsdaten ist in der Biodiversität vor allem für die Analyse von Zeitreihen interessant. Der Bedarf wird sich hier mit der Zeit erhöhen. Als problematisch wird jedoch die mangelnde Sichtbarkeit der Datenarchive genannt. Auch liegen viele interessante Forschungsobjekte nicht in digitaler Form vor.[115] Die wissenschaftlich genutzten Datendienste wie GBIF, obwohl für jeden im Internet zugänglich, sind eher der Dauerhaften Domäne zuzuordnen. Daneben gibt es die Public Sites wie von z.B. Birdlife International, in der NGOs v.a. hinsichtlich Nachhaltigkeit und Artenschutz Outreach und Lobby-Arbeit machen, die jedoch für die wissenschaftliche Arbeit nur von marginalem Interesse ist.

### Zusammenfassung und Fazit

Im Zusammenhang mit den Biodiversitäts-Exploratorien fördert die DFG schon seit 2006 geeignete Infrastrukturen, die die kollaborative Zusammenarbeit in dieser jungen interdisziplinären Forschungsdisziplin unterstützen.

Eine spezielle rechtliche Einschränkung betrifft den Schutz bedrohter Arten, der es unmöglich machen kann, Forschungsdaten weiterzugeben oder zu veröffentlichen.

Die Biodiversität ist intrinsisch interdisziplinär angelegt und zeigt daher auch Überschneidungen, beispielsweise mit den restlichen Lebenswissenschaften und den Geowissenschaften. Äußerst interessant ist das Projekt reBind am BGBM, welches sich zur Aufgabe gemacht hat anderweitig verlorene Daten direkt bei der Forscherin oder dem Forscher zu retten. Es ist damit eines der wenigen Projekte in unserer Untersuchung, welches sich in der privaten Domäne verorten lässt. Ein warnendes Beispiel für die Nachhaltigkeitsdiskussion im Bereich der Forschungsdateninfrastruktur ist das NBII, dessen Angebote durch Einstellung der Finanzierung durch die US-amerikanische Bundesregierung vollständig verschwunden sind.

[115] Vgl. Nieschulze, N., König-Ries, B. (2012) in: Neuroth, H. et al. (Hrsg.) (2012) *Langzeitarchivierung von Forschungsdaten – Eine Bestandsaufnahme*, Universitätsverlag Göttingen S. 222f

In Verbindung sowohl mit der politisch brisanten Frage des Umweltschutzes, als auch dem massiven ökonomischen Interesses an pharmakologischer und anderweitiger Nutzung, sind internationale wissenschaftliche Datensammlungen zur Biodiversität mit GBIF weit fortgeschritten. Flankierend sind vielfältige Arbeiten zur Standardisierung erfolgt bzw. in der Implementierungsphase. Wie in anderen Fachgebieten besteht eine erkennbare Lücke zwischen den von Institutionen getragenen Datensammlungen und Portalen, wo Workflows und Standards weitgehend klare Konturen haben, und dem Workflow der Daten in der privaten und der Gruppendomäne, in denen noch starker Bedarf an weiterer Entwicklung besteht.

## Medizin

### Wichtige Akteure

Medizinische Forschung findet in Deutschland zum Großteil in Einrichtungen wie **Universitätskrankenhäusern** oder ihnen angeschlossenen Forschungsinstituten statt, die immer auch der medizinischen Versorgung der Bevölkerung dienen. Es besteht eine starke Interdependenz zwischen dieser Versorgungstätigkeit und in den Einrichtungen durchgeführten Forschungstätigkeit. Neben den öffentlich finanzierten Einrichtungen existiert auch ein Bereich der privatwirtschaftlichen Forschung, insbesondere im Bereich der Pharmaindustrie.

Zentraler Dachverband der medizinischen Forschung in Deutschland ist die **Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften (AWMF)**[116] mit 163 Fachgesellschaften als Mitgliedern. Weitere wichtige nationale Institutionen mit wesentlichem Einfluss sind das **Bundesinstitut für Arzneimittel und Medizinprodukte (BfArM)**[117], verantwortlich für die Zulassung und Registrierung von Medikamenten, sowie das **Paul-Ehrlich-Institut (PEI)**[118]. Speziell im Bereich des Datenmanagements aktiv ist die Deutsche Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie. Auf internationaler Ebene haben Organisationen wie die **World Health Organisation (WHO)**[119] als Unterorganisation der Vereinten Nationen, sowie die US amerikanische **Food and Drug Administration (FDA)**[120] bzw. in **Europa die European Medicines Agency (EMEA)**[121] entscheidenden Einfluss auf die Regeln, nach dem medizinische Forschung und damit auch das Forschungsdatenmanagement ablaufen. Für letzteres ist auch die **International Medical Informatics Association (IMIA)**[122] als Internationale Kooperationsplattform für die Medizininformatik von Bedeutung.[123] Innerhalb Deutschlands dient die als Verein konstituierte **TMF – Technologie- und**

[116] Webseite http://www.awmf.org
[117] Webseite www.bfarm.de
[118] Webseite http://www.pei.de
[119] Webseite http://www.who.int
[120] Webseite http://www.fda.gov
[121] Webseite http://www.ema.europa.eu
[122] Webseite http://www.imia-medinfo.org/new2/
[123] Vgl. Dickmann, F., Rienhoff, O. (2012) in: Neuroth, H. et al. (Hrsg.) (2012) *Langzeitarchivierung von Forschungsdaten – Eine Bestandsaufnahme*, Universitätsverlag Göttingen S. 237f

**Methodenplattform für die vernetzte medizinische Forschung e.V**[124] der Koordination der medizinischen Verbundforschung.

Im Rahmen der D-Grid Initiative haben Projekte wie **MediGrid**, **MedInfoGRID** und **Pneumo-GRID** haben insbesondere auch die Möglichkeiten der verteilten Datenspeicherungen unter der speziellen Anforderungen der Medizin erforscht.[125]

### Domänenübergreifende Faktoren

#### *Art und Menge der Daten*

Digitale Daten in der Medizin umfassen Bild- und Sensordaten wie zum Beispiel MRT oder EKG Aufnahmen, Biomaterialdaten wie Blutproben, aber auch Statistiken oder Klassifikationen. Dazu kommen Befunddaten und Daten der Patentenverwaltung, welche über sog. Krankenhausinformationssysteme verwaltet werden. In den letzten Jahren haben besonders Genomdaten an Bedeutung gewonnen. Medizinische Forschungsdaten zeichnen sich durch ein starkes Wachstum bis in den Petabyte-Bereich aus. Dies betrifft besonders Daten aus der medizinischen Bildgebung sowie Genomdaten aus der DNS-Sequenzierung.[126]

#### *Formate*

Für medizinische Bilddaten ist *DICOM*[127] das führende Format. Für die Verwaltung der Daten wird *Picture Archive and Communication Systems (PACS)* genutzt. Bei den Sensordaten haben sich in der Community *EDF*[128] und *aECG*[129] durchgesetzt. Ein weiteres Format für Biosignale ist *GDF*[130]. Biomaterialdaten werden je nach Verfahren bzw. Laborgerät in verschiedenen Formaten gespeichert. Es kommen Tabellenformate wie CSV oder XLS zum Einsatz, aber auch proprietäre Formate. Daten aus der Mikroskopie werden analog zu den oben beschriebenen Bilddaten gespeichert. Für molekulare Strukturen wird unter anderem *FASTA*[131] verwendet. Für Befunddaten aus Krankenhausinformationssystemen haben sich die Formate des *HL7*[132] durchgesetzt. Statistikdaten werden in den Formaten der entsprechenden Softwarelösungen, wie R (frei), *SPSS* oder *SAS* (beide proprietär), aber auch in *CSV* gespeichert. Die Formate des *CDISC*[133] versuchen, die Datenformate aus den verschie-denen Bereichen zu einem gemeinsamen XML Standard zusammenzuführen.

---

[124] Webseite http://www.tmf-ev.de/
[125] Vgl. Dickmann, F., Rienhoff, O. (2012) in: Neuroth, H. et al. (Hrsg.) (2012) *Langzeitarchivierung von Forschungsdaten – Eine Bestandsaufnahme*, Universitätsverlag Göttingen S. 229
[126] Vgl. Dickmann, F., Rienhoff, O. (2012) in: Neuroth, H. et al. (Hrsg.) (2012) *Langzeitarchivierung von Forschungsdaten – Eine Bestandsaufnahme*, Universitätsverlag Göttingen S. 242
[127] Digital Imaging and Communication in Medicine, Webseite der National Electrical Manufacturers Association http://medical.nema.org, NEMA ist Copyright Holder von DICOM.
[128] European Data Format, Webseite http://www.edfplus.info
[129] HL7 v 3.0 annotated ECG , Webseite http://www.hl7.org/implement/standards/product_brief.cfm?product_id=102
[130] General Data Format for biomedical signals, vgl. Schlögl, A. (2011) GDF - A general dataformat for biosignals, arXiv:cs/0608052
[131] Formatbeschreibung http://blast.ncbi.nlm.nih.gov/blastcgihelp.shtml
[132] Health Level 7, Webseite http://www.hl7.de
[133] Clinical Data Interchange Standards Consortium, Webseite http://www.cdisc.org

### *Metadaten*

Das DICOM Format besitzt eine extensive Liste von Keywords zur Speicherung von Metadaten. Es nutzt das Uniform Type Identifier (UTI) System, um weitere Metadaten zu speichern.

### *Identifikatoren*

Im Rahmen von **Geoportis** bietet die ZB MED über das Projekt **DataCite**[134] einen Service für Digital Object Identifier (DOI) an.

### *Rechtliche Aspekte*

Humanmedizinische Forschungsdaten haben in der Regel einen direkten Personenbezug und unterliegen daher starken Datenschutzbestimmungen. Die Erhebung von Daten für ein Forschungsprojekt bzw. die Nutzung von Daten aus der Krankenversorgung erfordert daher eine explizite Einwilligung des Patienten. Des Weiteren liegt der Erhebung ein sog. Studienprotokoll zu Grunde, welches mit den zuständigen Ethikkommissionen abgestimmt werden muss. Da personenbezogene Daten nur sehr eingeschränkt verarbeitet oder weitergegeben werden dürfen, werden Forschungsdaten durch verschiedene Maßnahmen anonymisiert, indem die medizinischen Datenbestandteile von den identifizierenden Teilen getrennt werden. In bestimmten Situationen kann es aber auch erforderlich sein, diese Anonymisierung wieder rückgängig zu machen. Um diesen Ansprüchen gerecht zu werden, empfiehlt die TMF ein spezielles zweistufiges Verfahren.[135] Auch vollständig anonymisierte Daten können unter Umständen datenschutzrechtlich problematisch sein, da auch rein medizinische Daten (z.B. das vollständig sequenzierte Genom eines Menschen) Rückschlüsse auf die Identität des Patienten zulassen. Diese beschriebene Datenschutzproblematik gilt selbstverständlich nicht für Daten, die von Forschungstätigkeiten an Tieren (z.B. Mäusen) stammen.

### *Publikationen*

Mit **Medical Literature Analysis and Retrieval System Online (MEDLINE)**, bzw. **PubMed**[136] als Meta-Datenbank, existiert in der Medizin eine etablierte, öffentlich zugängliche Literaturdatenbank. Die **Deutsche Zentralbibliothek für Medizin (ZB MED)**[137] bietet im Rahmen von **Geoportis** einen Service zur Langzeitarchivierung von medizinischen Forschungsdaten an.[138]

## Umgang mit Forschungsdaten in den vier Domänen

### *Private Domäne*

Wie schon beschrieben sind die beiden Bereiche Versorgung und Forschung in der Humanmedizin eng verknüpft. Versorgungsdaten sind jedoch einerseits aus rechtlichen Gründen äußerst sensibel, aber auch von Struktur und Formaten stark heterogen und werden daher

[134] Projektwebseite http://datacite.org
[135] Vgl. Dickmann, F., Rienhoff, O. (2012) in: Neuroth, H. et al. (Hrsg.) (2012) *Langzeitarchivierung von Forschungsdaten – Eine Bestandsaufnahme*, Universitätsverlag Göttingen S. 249
[136] Webportal http://www.ncbi.nlm.nih.gov/pubmed/
[137] Webportal http://www.zbmed.de
[138] Vgl. Selbstbeschreibung der ZB Med, http://www.zbmed.de/ueber-uns/kernkompetenzen/langzeitarchivierung.html, online.

als Forschungsdaten kaum genutzt. Bei den Daten aus der Grundlagenforschung existieren diese Probleme in abgeschwächter Form. Die IT-Affinität der Forscher wird auch als eher gering beschrieben. Oft sind Daten und Dokumentationen noch papierbasiert. Die Bereitschaft der Forscherinnen und Forscher ihre Daten mit Metadaten zu versehen wird als gering beschrieben. Als Begründung wird genannt, dass dem damit verbundenen Aufwand dem kein entsprechender Nutzen gegenüberstehe.

### *Gruppen Domäne*

Kooperative Strukturen entstehen in der Medizin zwischen verschiedenen Instituten im Rahmen gemeinschaftlicher Projekte. Im Allgemeinen gibt es jedoch nur ein schwaches Interesse an der Bereitstellung von Daten für andere Forscher. Im Vergleich zu anderen Disziplinen ist die medizinische Forschung strukturell konservativer und stärker hierarchisch organisiert.

### Untersuchte Projekte

- Das Projekt Z03[139] des **TRR 54 Growth and Survival, Plasticity and cellular Interactivity of lymphatic Malignancies** arbeitet am Aufbau einer zentralen Datenbank zur Speicherung der im Forschungsverbund erhobenen Daten. Darauf aufbauend wird eine einheitliche Datenaufbereitung und Qualitätssicherung realisiert. Zum System gehört ein Web-basiertes Analysewerkzeug, genannt **LymphomExplorer**, mit Hilfe dessen die Mitglieder des TRR 54 ihre Daten analysieren und mit den Daten anderer Mitglieder vergleichen können. Um ein strenges Qualitätsmanagement zu gewährleisten, werden beim Ingest die Daten aufwändig getestet, normalisiert, gereinigt und gefiltert. Aus diesem Grund können die Daten nicht von den Forschern und Forscherinnen eigenständig hochgeladen werden. Über Seminare und Schulungen werden die Mitglieder des TRR 54 auf die Nutzung des LymphomExplorers vorbereitet. Das Projekt verwendet nur wenige Metadaten und keine Identifikatoren und bietet keine eigene Lösung zur Langzeitzeitarchivierung. Es wird vielmehr auf vorhandene Services in der Community verwiesen.
- Im **SFB 850 Control of Cell Motility in Morphogenesis, Cancer Invasion and Metastasis** wird im Projekt Z01[140] die zentrale Validationsplattform bereitgestellt. Weiter führt das Projekt von anderen Projekten des SFB angefragte Analysen durch.
- Die Integration der verschiedenen, heterogenen Daten aus den 22 Teilprojekten des **TRR77 Leberkrebs** ist die Aufgabe des Projekts Z2[141]. Ziel ist unter anderem die Ermöglichung projektübergreifender Auswertungen. Hierfür werden die Daten mit beschreibenden Metadaten ausgestattet und den Forscherinnen und Forscher des SFB nach einem dezidierten Zugriffskonzept zugänglich gemacht.
- Das Projekt Z1[142] des **SFB 1074 Experimentelle Modelle und klinische Translation bei Leukämien** arbeitet an neuen, parallelen Algorithmen für die Datenanalyse von

[139] Projektwebseite http://www.trr54.de/index.php?id=13&project=22&L=1
[140] Projektbeschreibung http://www.sfb850dev.uni-freiburg.de/projects-de/z/veelken_passlick_zurhausen
[141] Projektbeschreibung http://www.klinikum.uni-heidelberg.de/Biomedical-Research-Network.116298.0.html
[142] Projektwebseite http://www.uni-ulm.de/en/einrichtungen/sfb-1074/projects/z1.html

Sequenzdaten von Genomen. Diese Algorithmen sollen dann in die Daten-Analyse-Pipelines des SFB integriert werden und so die Forscherinnen und Forscher der anderen Projekte bei ihrer Arbeit unterstützen.

### *Dauerhafte Domäne*

Prinzipiell besitzt die Langzeitarchivierung von Forschungsdaten in der Medizin eine hohe Relevanz, da nur über eine funktionierende Archivierung aktuelle Daten mit denen aus zurückliegenden Behandlungen verglichen werden können. Die Dokumentation von Krankheitsverläufen ist schon immer wichtiger Bestandteil des medizinischen Alltags gewesen. In der heutigen Zeit ist dies durch die starke Personalfluktuation noch wichtiger geworden. Darüber hinaus existieren es auch gesetzliche Verpflichtungen zur Archivierung. Im Bereich der Forschungsdaten gibt es jedoch, obwohl Gegenstand intensiver Forschung, noch keine einheitliche Institution zur LZA auf nationaler Ebene. Neben der schon beschriebenen ZB MED existieren noch weitere publikationsorientierte Datenbanken, beispielsweise die **Cochrane Collaboration**[143]. International existiert mit **INSD**[144] eine Sammlung von DNA Sequenzen, welche vor der Publikation hochgeladen und dann referenziert werden können. Mit **wwPDB**[145] existiert ein ähnliches Angebot für Proteine. Weitere Internationale Datenbanken existieren für die Molekularbiologie. Von den Forschern wird besonders das Vertrauen in die Datenbestände als enorm wichtig angesehen.

#### Untersuchte Projekte

- Das von der DFG geförderte Projekt **LaBiMi**[146] beschäftigt sich mit dem Aufbau eines Zentralen Datenarchivs von Daten sowohl der Bildverarbeitung, als auch der Genomforschung. Es konzentriert sich absichtlich auf diese Gebiete, um für ein überschaubares Gebiet ein funktionierendes System zu etablieren, welches dann erweitert werden kann. Zur Projektlaufzeit werden Daten aus den beteiligten Partnern Magdeburg und Kiel verwendet. Anschließend soll das entstehende Archiv auch anderen Radiologen und Genomforschern zur Verfügung stehen.

### *Zugangsdomäne*

Aufgrund der starken Einschränkungen für die Weitergabe von Forschungsdaten ist die Zugangsdomäne in der Medizin kaum entwickelt.

### Zusammenfassung und Fazit

Medizinische Forschungsdaten an Universitätskrankenhäusern existieren immer im Spannungsfeld zwischen der medizinischen Versorgungstätigkeit und wissenschaftlicher Forschung. In Folge dessen gehen auch die Anforderungen an Datenschutz und Anonymisierung über andere Disziplinen hinaus. Eine Weitergabe von Forschungsdaten und ein etwaiger

[143] Webseite http://www.cochrane.de/de/arbeitsgebiet-cc
[144] International Nucleotide Sequence Database Collaboration, Webseite http://www.insdc.org
[145] Worldwide Protein Data Bank, Webseite http://www.wwpdb.org
[146] Projektwebseite http://www.labimi-f.med.uni-goettingen.de

Zugang für Forscherinnen und Forscher an anderen Institutionen sind nur unter engen Voraussetzungen möglich. Die für Forschungsdaten verwendeten Formate sind in der Regel über die verwendeten Geräte und Softwaresysteme vorgegeben und in der Regel proprietär.

Mit der TMF existiert bereits ein Akteur mit dem Potential, im Bereich des Forschungsdatenmanagements Koordinationsaufgaben wahrzunehmen. Unter den untersuchten Projekten ist unter anderem LaBiMi interessant, welches sich im Projekt stark auf wenige Partner fokussiert, aber dennoch weitergehende Konzepte entwickelt. Mit Pubmed existiert schon eine etablierte offene Datenbank für Publikationen.

## Astrophysik

### Wichtige Akteure

Die wichtigste Vereinigung im deutschsprachigen Raum ist die **Astronomischen Gesellschaft (AG)**[147] welche die Interessen der in der Astronomie arbeitenden Wissenschaftlerinnen und Wissenschaftler vertritt und die Astronomie im Allgemeinen durch Veranstaltungen, Publikationen und Öffentlichkeitsarbeit fördert. Der **Rat Deutscher Sternwarten (RDS)** vertritt die Interessen der Astronomischen Forschungsinstitute in Deutschland. Der RDS war nach dem zweiten Weltkrieg als nationale Einrichtung innerhalb der Bundesrepublik entstanden, da sich die AG damals (und auch heute) als internationaler Verein versteht und daher auch viele Mitglieder aus Österreich und der Schweiz hat. Seit 2012 ist der RDS ein Organ der AG. Der RDS vertritt die deutsche Astronomie auch in der wichtigsten internationalen Vereinigung, der **International Astronomical Union (IAU)**. Innerhalb der IAU beschäftigt sich die **IAU Working Group on Astronomical Data**[148] mit allen Aspekten des Datenmanagements.

Die wichtigste Organisation im Bereich der Forschungsdaten ist derzeit die 2002 gegründete **International Virtual Observatory Alliance (IVOA)**[149]. Als Zusammenschluss von 20 nationalen Programmen zur Etablierung von **virtuellen Observatorien (VO)**, d.h. Einrichtungen, durch die der Zugang zu Archivdaten ermöglicht wird, gegründet, arbeitet die IVOA an der Entwicklung von gemeinsamen Standards, Protokolle und Datendefinitionen. Die für Deutschland zuständige Unterstruktur ist das seit 2003 vom BMBF geförderte **German Astrophysical Virtual Observatory (GAVO)**[150]**.** Im selben Kontext wird auf europäischer Ebene das FP-7 Projekt **EURO-VO**[151] gefördert. Eine der erfolgreichsten Einrichtungen im Umfeld des europäischen VO ist das **Centre de Données astronomiques de Strasbourg (CDS)**[152]. Das CDS stellt über verschiedene Services Datensätze zu bestimmten Himmelsobjekten oder -bereichen in verschiedensten Archiven zur Verfügung, die sich auffinden und zur Nachnutzung herunterladen lassen.

[147] Webseite http://www.astronomische-gesellschaft.org
[148] Webseite http://www.atnf.csiro.au/people/rnorris/WGAD/
[149] Webseite http://www.ivoa.net
[150] Projektwebseite http://www.g-vo.org
[151] Projektwebseite http://www.euro-vo.org
[152] Webportal http://cdsweb.u-strasbg.fr

Besondere Bedeutung in der Astronomie haben die Träger der diversen weltweit verteilten Beobachtungseinrichtungen. Für die Europäische Astronomie sind das besonders die **Europäische Südsternwarte (ESO)**[153] als Betreiber diverser Observatorien hauptsächlich in Chile, und die **Europäische Weltraumorganisation (ESA)**[154], welche für diverse astronomische Satelliten-Missionen verantwortlich ist. Sowohl ESO als auch ESA haben schon seit längerer Zeit Archive aufgebaut, in denen die Daten aus den jeweiligen Observatorien/Missionen nach einer Karenzzeit (z.B. ein Jahr im **ESO-Archiv**[155]) der Öffentlichkeit zur Verfügung gestellt werden.

Im Rahmen der D-Grid-Initiative wurden von 2005-2010 im **AstroGrid-D**[156] IT Infrastrukturen der Astronomischen Community in Deutschland zur Verfügung gestellt. Einzelne Dienste und Webservices zu Datenarchiven werden durch die an der Entwicklung beteiligten Institute weitergeführt.

### Domänenübergreifende Faktoren

#### *Art und Menge der Daten*

Astrophysikalische Forschungsdaten zeichnen sich dadurch aus, das sie nicht durch Experimente im Labor erzeugt werden können, sondern nur durch Beobachtung mittels Observatorien oder ähnlichen Einrichtungen erstellt werden können. Zusätzlich produzieren Computersimulationen astrophysikalischer Modelle Datensätze, welche dann mit den Beobachtungen verglichen werden können. Die Größe der einzelnen Datensätze hängt stark vom jeweiligen Projekt ab: Einzelne Beobachtungsdaten sind eher klein (bis einige GB), während umfangreiche Archive von Beobachtungkampagnen (sog. Surveys) oder Simulations-Sets hunderte von Terabyte umfassen können. Insgesamt liegt der Zuwachs an Daten in der Astronomie im dreistelligen Terabyte Bereich pro Jahr.[157]

#### *Formate*

Als Standard für astronomische Beobachtungsdaten hat sich seit den 1980er Jahren das *FITS*[158] Format durchgesetzt. In neuerer Zeit, besonders im Zusammenhang mit den schon erwähnten Surveys, werden SQL Datenbanken genutzt. Im Zuge des VO wurden mit *VOTables*[159] standardisierte Formate eingeführt. Bei den Simulationsdaten sind in der Regel die Formate der Datenprodukte im Prozess der Entwicklung des jeweiligen Codes entstanden und sind nicht standardisiert. Einige populäre Codes haben so inoffizielle de-facto Standards geschaffen.

---

[153] Webseite http://www.eso.org
[154] Webseite http://www.esa.int
[155] Webseite http://archive.eso.org/eso/eso_archive_main.html
[156] Webseite http://www.astrogrid-d.org
[157] Vgl. Enke, H., Wambsganß, J. (2012) in: Neuroth, H. et al. (Hrsg.) (2012) *Langzeitarchivierung von Forschungsdaten – Eine Bestandsaufnahme*, Universitätsverlag Göttingen S. 276
[158] Support Webseite der NASA http://fits.gsfc.nasa.gov
[159] Webseite http://www.ivoa.net/Documents/VOTable/

### *Metadaten*

Standardisierte Metadaten werden in der Astronomie kaum verwendet. Es gibt einen Standard für die Einträge im FITS-Header, der jedoch sehr wenig festes Vokabular (nur ca. 80 Keywords) enthält. Im VO wurden *Unified Content Deskriptoren (UCD)* und *UTypes* für komplexere Systeme entwickelt, die zum Ziel haben alle Elemente eines Datenmodells in Key/Value-Listen zu beschreiben. Die IVOA hat für die verschiedenen Beobachtungsarten Datenmodelle entworfen (Photometrie, Spektrometrie, etc.), deren Nutzung jedoch noch auf wenige Werkzeuge beschränkt ist. Auch für die Modellierung von Simulationsdaten sind Metadaten (UTypes) entwickelt worden[160].

### *Identifikatoren*

In der Regel werden astronomische Objekte über ihre Position am Himmel Identifiziert. Hierfür existieren Richtlinien seitens der IAU. In Archiven der verschiedenen Organisationen (ESO, HST), werden zusätzliche Merkmale in der Form von Buchstabenkombinationen hinzugefügt. In der Regel haben größere Archive ein eigenes System von Identifiern, welches aber nur innerhalb des jeweiligen Archivs eine systematische Bedeutung hat[161]. Für ganze Kataloge/Datensätze hat die IVOA ein auf URN basiertes System von Identifiern erarbeitet. Eine Klärung, wie dieses System in Kombination mit den anderswo genutzten Identifiern wie DOI gemeinsam genutzt werden kann, steht aus. Anwendung findet das URN-System derzeit nur in Werkzeugen, die innerhalb des VO arbeiten.

### *Rechtliche Aspekte*

Für Daten aus der Astronomie existieren keine rechtlichen Einschränkungen.

### *Finanzierung*

Die Finanzierung von Archiven obliegt in der Regel den Instituten, die die entsprechende Beobachtungskampagne leiten bzw. Datenmanagement hierfür übernehmen. Die Archive der supranationalen Observatorien (z.B. ESO) werden aus deren Haushalt finanziert. GAVO wird seit 2003 kontinuierlich vom BMBF Fachreferat gefördert.

### *Publikationen*

Die Suche nach Literatur findet in der Regel über das **SAO-NASA Astrophysics Data System**[162], welches normalerweise als **ADS** abgekürzt wird, statt. In der Astrophysik ist es weit verbreitet und von den wichtigen Zeitschriften akzeptiert, zur Publikation eingereichte Fachar-

---

[160] Vgl, *Utype: A data model field name convention*, IVOA Note May 24, 2009, und *UCD: The UCD1+ controlled vocabulary*, IVOA Recommendation 02 April 2007

[161] Beispielsweise hat der TWOMASS Katalog einen aus 2 8-stelligen Ziffernfolgen bestehenden Objektbezeichner. Andere Kataloge nutzen eine Kombination aus Buchstaben und Koordinaten oder fortlaufender Nummer (BD +20 111 = Bonner Durchmusterung, +20 Grad RA, 111 Objekt, usw.).

[162] Webportal http://adswww.harvard.edu

tikel auf den Open Access Server **arXiv**[163] **(astroph)** als Preprint zu speichern und so der Allgemeinheit schneller Forschungsergebnisse zur Verfügung zu stellen.

## Umgang mit Forschungsdaten in den vier Domänen

### *Private Domäne*

In ihrer Arbeit mit Astronomischen Daten benutzen Astronomen größtenteils kommerzielle oder freie Programmpakete (IDL, IRAF, MIDAS, MATLAB). In der theoretischen Astrophysik wird normalerweise eigener Code entwickelt. Die IT-Abteilungen der Institute kümmern sich um den Betrieb der Hardware und Software, stehen aber nicht für höheren Support zur Verfügung. Für kommerzielle Software gibt es Trainingskurse, die von den entsprechenden Firmen veranstaltet werden. Die Wartung und Weiterentwicklung von selbstgeschriebener Software wird durch den jeweiligen Autor geleistet. Einige Codes werden auch als Open Source Software veröffentlicht und sind teilweise, wenn die Zitate des dazugehörigen Papers betrachtet werden, außerordentlich erfolgreich (z.B. DAOPHOT[164] auf der Beobachtungsseite oder GADGET2[165] bei den Simulationen). Die Archivierung der anfallenden Daten obliegt dem einzelnen Forscher. Meistens existieren keine institutsweiten Regelungen über die DFG-Richtlinien zur guten wissenschaftlichen Praxis hinaus.

### *Gruppen Domäne*

Da die Aufstellung von wissenschaftlich konkurrenzfähigen Teleskopen nur an wenigen Standorten weltweit möglich ist, ist die astrophysikalische Forschung schon lange stark international ausgerichtet und durch kooperative Strukturen gekennzeichnet. Die Observatorien werden durch Institutionen wie die ESO betrieben, die in der Regel Archive der Rohdaten betreiben. Diese Archive sind jedoch, was Formate und Metadaten betrifft, äußerst heterogen. Für größere Kollaborationen ist Datenmanagement inzwischen etabliert. In vielen Fällen geschieht dies zumindest teilweise unter Nutzung der Standardisierungsarbeiten des VO. Insgesamt ist die Dateninfrastruktur jedoch noch nicht genügend ausgebaut. Die zunehmend bedeutender werdenden Surveys sind Projekte von grossen Kollaborationen, die auch einen Teil des Datenmanagements einschliesslich Kuratierung und Archivierung übernehmen.

#### Untersuchte Projekte

- Das INF Projekt im **SFB 881 The Milky Way System** Teil des SFB ist ein Unterprojekt namens „Information Infrastructure“ (in Kooperation mit GAVO) dessen Ziel die Publikation der vom SFB erzeugten Daten ist. Hierzu sollen die Standards des VO genutzt werden. Dies geschieht kurzfristig auch mit Mitteln aus dem SFB. Langfristig soll sich die Universität ebenfalls beteiligen.

[163] Webportal http://de.arxiv.org/archive/astro-ph
[164] ADS Eintrag http://adsabs.harvard.edu/abs/1987PASP...99..191S
[165] ADS Eintrag http://adsabs.harvard.edu/abs/2005MNRAS.364.1105S, Webseite http://www.mpa-garching.mpg.de/gadget/

- Im **SFB 963 Astrophysikalische Strömungsinstabilität und Turbulenz** stellt das Projekt INF[166] die zentrale Forschungsdateninfrastruktur des SFB bereit. Neben dem Austausch und der Analyse der im SFB anfallenden Beobachtungsdaten wird auch an der Kalibrierung von Simulationsmodellen gearbeitet. Das Projekt orientiert sich in seiner Arbeit an Standards des VO und arbeitet mit internationalen Netzwerken wie EURO-VO zusammen.

### *Dauerhafte Domäne*

In der Astronomie sind schon seit einiger Zeit Datenzentren in den bedeutenden Observatorien (**ESO**, **ESA**, **HST**[167]) etabliert, welche sämtliche aufgenommenen (Roh-)Daten archivieren, zusammen mit der Software, die instrumentbezogene Charakteristiken "bereinigt".

Bei den Computersimulationen befindet sich diese Entwicklung noch in den Anfängen. Im Prinzip ist der Aufwand, Simulationen nach einiger Zeit neu zu berechnen geringer als der, die Daten zu speichern. Eine Archivierung ist daher hauptsächlich im Hinblick auf eine Veröffentlichung, insbesondere für Astronomen außerhalb der eigentlichen Simulationscommunity, interessant. Projekte wie die **Millennium Database**[168] und **MultiDark**[169] bieten kosmologische Simulationsdaten über Webportale an, und ermöglichen einer großen Anzahl interessierter Forscher die Arbeit mit solchen Daten. Mit dem **Astrophysical Software Laboratory** ist eine europäische Einrichtung für die Unterstützungen der Entwicklung von Simulationscodes geplant.

### *Zugangsdomäne*

Im letzten Jahrzehnt hat in der Astrophysik ein Umdenken hin zu einer Kultur der Nachnutzung von Forschungsdaten begonnen. Dies ist eng an den Erfolg von Projekten wie dem **Sloan Digital Sky Survey**[170] oder der Millennium Database gekoppelt. Diese Projekte haben gezeigt, dass das zur Verfügung stellen von Daten sich auch in klassischen wissenschaftlichen Erfolgsparametern wie Publikationen und Zitationen auszahlt.

Die verschiedenen Dienste und Protokolle des VO wie z.B. das **Table Access Protocol (TAP)** wurden nicht zuletzt für Nachnutzung von Forschungsdaten konzipiert. Innerhalb des VO dient die VO-Registry und auch das **GAVO Datencenter**[171] dem Auffinden von Datensätzen zur Nachnutzung. Diese Dienste werden nach und nach durch die Community angenommen, wobei dies stark an einen Effizienzgewinn für den Forschen bzw. die Forscherin gebunden ist.

[166] Projektübersicht http://www.uni-goettingen.de/de/215363.html
[167] Hubble Space Telecope, Webportal zum Mikulski Archive for Space Telescopes http://archive.stsci.edu
[168] Webportal zur Datenbank http://gavo.mpa-garching.mpg.de/Millennium/
[169] Webportal zur Datenbank http://www.multidark.org/MultiDark/
[170] Webseite des Projektes http://www.sdss.org
[171] Webportal des GAVO Datenzentrums http://dc.zah.uni-heidelberg.de

### Zusammenfassung und Fazit

Die Astronomie ist eine stark international aufgestellte Forschungsdisziplin. Insbesondere auf Kollaborationsebene, in der Gruppendomäne, besteht die Notwendigkeit Forschungsdaten von signifikantem Volumen Forscherinnen und Forschern über Instituts- und Ländergrenzen hinweg zugänglich zu machen.

Durch das Aufkommen von Leuchtturmprojekten wie dem SDSS Survey ist die Akzeptanz von Maßnahmen zu einem professionelleren Datenmanagement deutlich gestiegen. Archive werden eher im Hinblick auf eine Datenveröffentlichung als auf die bloße dauerhafte Aufbewahrung akzeptiert.

In der beobachtenden Astronomie besteht mit FITS ein etabliertes Format, jedoch sind Metadatenstandards wenig akzeptiert und unterentwickelt. Bei den Simulationen herrscht auch bei den Formaten keine Übereinkunft. Durch die VO Bewegung, in Deutschland durch GAVO, sind zwar Systeme für Formate und Metadatenstandards erarbeitet, setzen sich jedoch in der Community nur sehr langsam durch.

## Geo-, Meeres- und Klimawissenschaften

### Wichtige Akteure

Zu den wichtigste Akteuren im Bereich der Forschungsdaten in den Geo-, Meeres- und Klimawissenschaften sind die Zentren des **World Data Systems**[172] des **Internationalen Wissenschaftsrats**[173]. Es handelt sich hierbei um die Weiterentwicklung der **World Data Center (WDC)**, einem System aus über 50 Datenzentren in 12 Ländern, welches auf eine über 50 Jährige Geschichte zurückblickt und in Folge des Internationalen Geophysikalischen Jahrs 1957 – 1958 geschaffen wurde. In Deutschland existieren drei WDC. Das auf die Bereiche Globaler Wandel und Erdsystemforschung ausgerichtete **World Data Center for Marine Environmental Sciences (WDC-MARE)**[174] wird gemeinschaftlich vom **Alfred-Wegener-Institut für Polar- und Meeresforschung (AWI)**[175] und dem **Zentrum für Marine Umweltwissenschaften (MARUM)**[176] der Universität Bremen betrieben. Vom WDC-Mare wird insbesondere das Datenportal **PANGAEA**[177] betrieben. Das **World Data Center for Climate (WDCC)**[178] sammelt Daten aus der Klimaforschung und wird vom **Deutschen Klimarechenzentrum (DKRZ)**[179] betrieben. Auf Satellitendaten spezialisiert ist das **World Data Center for Remote Sensing of the Atmosphere (WDC-RSAT)**[180], welches vom **Deutsches Fernerkun-**

[172] Webseite http://www.icsu-wds.org
[173] International Council for Science (ICSU), Webseite http://www.icsu.org
[174] Webseite http://www.wdc-mare.org
[175] Webseite http://www.awi.de
[176] Webseite http://www.marum.de
[177] Webseite http://www.pangaea.de
[178] Webseite http://www.dkrz.de/daten/wdcc/
[179] Webseite http://www.dkrz.de
[180] Webseite http://wdc.dlr.de

**dungsdatenzentrum (DFD)** des **Deutschen Zentrums für Luft- und Raumfahrt (DLR)** in Oberpfaffenhofen betrieben wird[181].

Das **Helmholtz-Zentrum Potsdam - Deutsches GeoForschungsZentrum (GFZ)** ist im Bereich der Forschungsdaten zum einen beratend im Bereich Daten- und Informationsmanagement durch das **Zentrum für Geoinformationstechnologie (CeGIT)** tätig und stellt zum anderen diverse Dienste zur Publikation von Forschungsdaten zur Verfügung[182]. Zusammen mit den drei deutschen WDC bildet das GFZ den deutschen **WDC Cluster für Erdsystemforschung**. Auch das **GEOMAR Helmholtz-Zentrum für Ozeanforschung Kiel** stellt diverse Datensätze über ein Webportal[183] zur Verfügung.

Im Zuge der D-GRID Initiative wurde im **C3GRID**[184] eine Infrastruktur entwickelt und aufgebaut, welche einen die einen einheitlichen Zugriff auf die Daten der Projektpartner innerhalb der klimawissenschaftlichen Community ermöglicht. C3 GRID strebt hierbei einen einheitlichen, für den Nutzer transparenten Zugang zu den verschiedenen Datensätzen an. Mit **C3Grid - INAD: Towards an Infrastructure for General Access to Climate Data** wird die Infrastruktur weiterentwickelt und noch stärker auf den Zugang zu den Forschungsdaten ausgerichtet.

### Domänenübergreifende Faktoren

#### *Art und Menge der Daten*

Generell lassen sich in den Geo- und Meereswissenschaften drei Arten von Daten unterscheiden. Dies sind zunächst Daten aus Messgeräten, beispielsweise von Satelliten oder Großgeräten. Diese Daten werden in der Regel automatisiert prozessiert und sind gut maschinell zu verarbeiten. Sie sind außerdem meist episodischer Natur, d.h. sie bilden immer einen Zustand zu einem bestimmten Zeitpunkt ab und lassen sich daher nicht durch nochmalige Messung wiederherstellen. Die zweite Gruppe umfasst Daten aus Computermodellen bzw. numerischer Modellierung. Sowohl bei den Messdaten als auch bei den Simulationsdaten gehen die Datenmengen heute bis in den Petabyte Bereich. Auch diese Daten liegen in verarbeitbarer Form vor. Die letzte Gruppe schließlich umfasst individuell erstellte Datensätze aus Labormessungen, Felderhebungen, Literaturrecherche, etc. Diese Datensätze sind hoch heterogen und oft schlecht dokumentiert, aber in der Erstellung pro Datenmenge sehr teuer.

#### *Formate*

In den Geowissenschaften werden für Daten von Sensoren, Großgeräte und aus der Modellierung bzw. von Simulationsdaten werden je nach Subdisziplin verschiedene Formate verwendet. So wird beispielsweise *SEED*[185] für die Seismologie, *netCDF*[186] für die Fern-

[181] Vgl. http://www.dlr.de/eoc/de/desktopdefault.aspx/tabid-5278/8856_read-15911/, online.
[182] Vgl. http://www.gfz-potsdam.de/portal/gfz/Services/Forschungsdaten, online.
[183] Webseite https://portal.geomar.de
[184] Projektwebseite http://www.c3grid.de
[185] Standard for the Exchange of Earthquake Data, Offizielles Manual http://www.fdsn.org/seed_manual/SEEDManual_V2.4.pdf

erkundung und die Erdsystemforschung, *GeoTIFF*[187] für die Fernerkundung und *Shapefile*[188] für geographische Informationssysteme verwendet.

In der Klimaforschung haben sich bei den Computersimulationen das schon erwähnte *netCDF* und *GRIB*[189] als Formate durchgesetzt. Insbesondere akzeptiert das WDCC nur diese beiden Formate neben ASCII für die Archivierung. Die in den Simulationsdaten verwendeten Variablen und Maßeinheiten werden durch die *Climate Forecast Convention*[190] vorgegeben. Für Beobachtungsdaten in den Klimawissenschaften wird meist ASCII verwendet. Es kommen aber auch diverse Bild-, Video-, Audioformate zum Einsatz.

### *Metadaten*

Die Verwendung von Metadaten ist in den Geo- und Klimawissenschaften bereits weit verbreitet. Genau wie bei den Formaten existiert auch bei den Metadaten eine Vielzahl von Standards. Hervorzuheben sind hierbei, wegen ihrer Verbreitung, aber auch aufgrund ihrer Nutzung durch staatliche Stellen, die Formate der ISO 19115 Familie. Weitere, im Rahmen des **GeoSpatial Consortium (OGC)**[191] entstandene Standards sind *Sensor Web Enablement (SWE)*, *Observation&Measurements Modell (O&M)*, und *GeoSciML*. Auf Europäischer Ebene wird mit *INSPIRE*[192] versucht, auf Basis von ISO 19115 ein einheitliches System von Metadaten zu schaffen. Weiter ist hier das *Directory Interchange Format (DIF)*, das durch das *Global Change Master Directory (GCMD)*[193], ein von der NASA betriebenes Datenportal zur Klimaforschung, starke Verbreitung gefunden hat, zu erwähnen. In den einzelnen Subdisziplinen sind auch weitere, nicht XML-basierte Standards wie *QuakeML*[194], Darwin Core oder *ThermoML*[195] in Verwendung.

### *Identifikatoren*

Bei den Identifikatoren haben seit der Jahrtausendwende zunächst das Projekt **CODATA**[196] und später das DFG-Projekt **Publikation und Zitierbarkeit wissenschaftlicher Primärdaten (STD-DOI)** ein Konzept und eine Infrastruktur für PID auf Basis der Digital Object Identifiers (DOI) erarbeitet. Aus STD-DOI ist 2009 das Projekt **DataCite**[197] hervorgegangen, welches unter anderem auch die Zitierbarkeit von Forschungsdaten über Publikationsportale wie beispielsweise ScienceDirect[198] weiterentwickelt.

---

[186] Network Common Data Form, Webseite http://www.unidata.ucar.edu/software/netcdf/
[187] Webseite http://trac.osgeo.org/geotiff/
[188] Technische Beschreibung http://www.esri.com/library/whitepapers/pdfs/shapefile.pdf
[189] Webseite http://www.weatheroffice.gc.ca/grib/index_e.html
[190] Webseite http://cf-pcmdi.llnl.gov
[191] Webseite http://www.opengeospatial.org
[192] Projektwebseite http://inspire.jrc.ec.europa.eu
[193] Webportal http://gcmd.nasa.gov
[194] Webseite https://quake.ethz.ch/quakeml/
[195] Webseite http://www.trc.nist.gov/ThermoML.html
[196] Projektwebseite http://www.codata.org
[197] Projektwebseite http://datacite.org
[198] Webportal http://www.sciencedirect.com

### *Rechtliche Aspekte*

Datenschutz oder andere rechtliche Vorgaben spielen in den Geo- und Klimawissenschaften kaum eine Rolle. Eine Beschränkung der Veröffentlichung von Forschungsdaten ist aber insbesondere bei der Zusammenarbeit mit Industriepartnern möglich. Weitere Nutzungsbeschränkungen existieren durch mögliche kommerzielle Interessen bei Luft- und Satellitenbildern sowie Karten, und auch bei Umweltmessungen.

### *Finanzierung*

Zentren des World Data Systems (vormals WDC) werden über den Grundhaushalt der jeweiligen Einrichtungen, die das Datenzentrum betreiben, finanziert. Das C3Grid wurde von 2005 bis 2009 als Teil der D-Grid Initiative vom BMBF finanziert. Das Nachfolgeprojekt C3-INAD wird vom BMBF aus dem Fachreferat „Globaler Wandel“ finanziert.[199]

## Umgang mit Forschungsdaten in den vier Domänen

### *Private Domäne*

In den Geo- und Meereswissenschaften lässt sich die Verantwortung für den angemessenen Umgang mit Forschungsdaten klar bei den einzelnen Forschern verorten. In der heutigen Zeit können die Forscher den Lebenszyklus der Forschungsdaten häufig noch komplett selbst abdecken. Hierbei werden sie von auch von Softwarepaketen wie *PanMetaDocs*[200] unterstützt.

### *Gruppen Domäne*

Größere Forschungsvorhaben in den Geowissenschaften werden schon seit längerer Zeit in Forschungsverbünden durchgeführt. Die umfasst beispielsweise diverse ozeanische und kontinentale Bohrprojekte. Über die Zeit haben diese Projekte eine gemeinsame Forschungsdateninfrastruktur aufgebaut (siehe SEDIS[201]). Auch in anderen Bereichen bestehen langjährige Zusammenarbeiten zwischen verschiedenen Institutionen. Diese Projekte können dann auch über spezialisierte Datenmanager verfügen, deren Aufgaben vom Sammeln und der Beschreibung der Datensätze bis zu Entwicklungsaufgaben reicht. Nicht alle dieser Strukturen werden jedoch durch eine adäquate Forschungsdateninfrastruktur abgedeckt.

Auch in den Klimawissenschaften sind große Instituts- bzw. Länderübergreifende Kollaborationen eher die Regel als die Ausnahme. Hierbei werden Großgeräte, wie z.B. der Eisbrecher Polarstern, in der Regel von einzelnen Instituten (in diesem Fall AWI) betrieben, sind aber von der ganzen Community auf Bewerbung nutzbar.

[199] Vgl. Präsentation http://www.d-grid-gmbh.de/fileadmin/downloads/Ergebniskonferenz_2012/WissGrid-Klima.pdf
[200] Webseite http://sourceforge.net/projects/panmetadocs/
[201] Projektwebseite http://sedis.iodp.org

### Untersuchte Projekte

- Das Projekt Z1[202] im **TR32 Patterns in Soil-Vegetation-Atmosphere-Systems** arbeitet an der Speicherung und dem Austausch der Daten innerhalb des TR über eine Datenbank. Dies stellt besondere Anforderungen, da diese Daten auf Grund der fachübergreifenden Natur des TR äußerst heterogen sind. Neben geographischen Daten werden auch Metadaten, Literatur bzw. Publikationsdaten sowie Videos und Bilder verwaltet.

## *Dauerhafte Domäne*

Die dauerhafte Archivierung von Forschungsdaten hat in den Geowissenschaften, nicht zuletzt durch ihre schon angesprochene episodischer Natur, eine lange Tradition. Wie weiter oben beschrieben, wurden schon in den 1950er Jahren die Grundlagen für das heutige System der World Data Centers (WDC) geschaffen. Zusammenarbeit mit den Bibliotheken existiert, wie schon beschrieben, über das DataCite Projekt. Insgesamt ist ein kultureller Wandel bei der LZA zu beobachten.

### Untersuchte Projekte

- Im von der DFG geförderten Projekt **EWIG**[203] soll ein Konzept und Workflowkomponenten für ein zentrales Langzeitarchiv im Bereich Erd- und Umweltwissenschaften erarbeitet werden. Ein besonderer Schwerpunkt ist hierbei die Optimierung des Ingestprozesses an der Schnittstelle zwischen disziplinspezifischen Datenprodukt und generischem Langzeitarchiv. Das Projekt wird begleitet von öffentlicher Diskussion und Dokumentation des Entwicklungsprozesses und iterativen Test durch Wissenschaftler und Studierende.
- Ebenfalls von der DFG wird das vom **GESEP e.V.**[204] initiierte **Deutsche Bohrkernlager** gefördert. In einem ersten Schritt wird eine Probenverwaltungssoftware mit der Unterstützung für persistente Identifikatoren entwickelt. Weiter wird ein digitales Portal aufgebaut. Am Ende soll eine projektunabhängige, zentrale Datenbank stehen, welche Primärdaten, Metadaten, Proben und Publikationen zusammenführt. Zur persistenten Identifikation der Probenstücke in Sammlungen, Daten und Literatur wird die **International Geo Sample Number (IGSN)**[205] verwendet.

## *Zugangsdomäne*

Generell besteht, wieder aus dem Grund der Einmaligkeit der untersuchten Naturvorgänge, ein großes Interesse an der Nachnutzung von Forschungsdaten. Dies gilt aber nicht uneingeschränkt für alle Daten, da verbesserte Forschungsmethoden auch in den Geo- und Meereswissenschaften kontinuierliche Erneuerung erforderlich machen. Beispiele für Daten Portale sind Projekte wie das schon erwähnte SEDIS, sowie die oben beschriebenen WDC. Eine

[202] Webseite des TR32 http://tr32.uni-koeln.de
[203] Projektwebseite http://ewig.gfz-potsdam.de
[204] Deutsches Forschungsbohrkonsortium, Webseite http://www.gesep.de
[205] Webseite der SESAR – System for Earth Sample Registration http://www.geosamples.org

Sperrfrist für die Nachnutzung von zwei Jahren nach Projektende ist üblich. Ein starker Hinderungsgrund für eine effiziente Nachnutzung liegt jedoch in der mangelhaften Aufbereitung der Daten für die Archivierung und die Auffindbarkeit (Metadaten, Identifikatoren). Es besteht daher ein Bedarf an Ansprechpartnern und Werkzeugen, um den einzelnen Wissenschaftler bei dem datenkuratorischen Prozess zu unterstützen.

### Untersuchte Projekte

- Das Projekt **KOMFOR**[206] ist als Bindeglied zwischen Forschungseinrichtungen, Verlagen, Bibliotheken und dem bestehenden Archivnetzwerk aus Erd- und Umweltforschung geplant. Im Projekt sollen nachhaltige und verlässliche Wege zur Publikation wissenschaftlicher Daten geschaffen werden. Hierbei ist ein ganzheitlicher Ansatz angedacht, der den Forscher bei allen Fragen des Datenmanagements begleitet. Kernstücke sind die organisatorische und technische Zusammenarbeit der einzelnen Akteure, sowie eine webbasierte personalisierte Arbeitsplattform. Die Möglichkeit, Kontextinformationen in einer Projektsituation zur Erhebung von Metadaten zu nutzen, soll für den Datenproduzenten die Dokumentation seiner Daten weitgehend automatisieren. Der Forscher als Konsument von Forschungsdaten wird Daten über ein übergreifendes Datenportal recherchieren können.
- Das von der DFG geförderte Projekt **Re3Data**[207] erstellt ein Register der bestehenden Forschungsdaten Repositorien in Deutschland in den Geowissenschaften, um einerseits Antragstellern einen Überblick über die vorhandenen Einrichtungen zu ermöglichen, und andererseits Qualitätskriterien für Repositorien weiterzuentwickeln.

### Zusammenfassung und Fazit

In den Geo-, Meeres- und Klimawissenschaften besteht schon seit den langem das System der WDC bzw. der WDS. In Deutschland bilden die drei Zentren WDC-MARE, WDCC, und WDC-RSAT zusammen mit dem GFZ den WDC Cluster für Erdsystemforschung. Im Projekt KOMFOR arbeiten diese Akteure gemeinschaftlich an übergreifenden Lösungen zu allen Bereichen des Forschungsdatenmanagements.

Durch die lange Tradition von Datenzentren ist auch die Verwendung von Metadatensysteme etabliert. Aufgrund der hohen Anzahl der verschiedenen Datenarten und Formate gibt es auch eine Vielzahl an verwendeten Metadatensystemen.

Neben großvolumigen, eher homogenen Datensätzen finden sich in den Geo-, Meeres- und Klimawissenschaften auch viele kleine Datensätze die, gemessen an der Datenmenge, äußerst teuer zu reproduzieren wären und dementsprechend erhaltenswert sind. Auch physikalische Objekte wie Bohrkerne werden archiviert und werden mit denselben Techniken des Forschungsdatenmanagements behandelt.

[206] Projektwebseite http://www.komfor.net
[207] Projektwebseite http://www.re3data.org

Ähnlich wie in den Sozialwissenschaften sind auch Geodaten im Open-Data Kontext interessant. Insbesondere im Bereich der Bodenschätze können geowissenschaftliche Daten kommerziell interessant sein.

### Weitere untersuchte Projekte aus anderen Disziplinen

Die folgenden Projekte lassen sich nicht klar einem der obigen Bereiche zuordnen und werden hier in einem eignen Abschnitt behandelt.

#### *Domänenübergreifend*

- Im von der DFG geförderten Projekt **PubFlow**[208] sollen wissenschaftliche Workflows im Umgang mit Forschungsdaten entwickelt werden. Dies soll alle Arbeitsschritte von der Erhebung der Daten bis zu ihrer Archivierung und Publikation umfassen. In der ersten Projektphase konzentriert sich das Projekt auf Meeres- und Geowissenschaftliche Publikationsprozesse (ins Zusammenarbeit mit dem IFM-Geomar), um später perspektivisch andere Wissenschaftsbereiche mit einzubeziehen. Ziel ist ein semi-automatischer Workflow bei dem Aufgaben, die nicht maschinell abgearbeitet werden können, an vorgegebene Personen delegiert werden. Der PubFlow Workflow beinhaltet auch Transport und Qualitätskontrolle (Metadatenvalidierung, Plausibilitätskontrollen) der Daten.
- Das vom BMBF geförderte Projekt **eScience Interfaces**[209] beschäftigt sich vom soziologischen Standpunkt mit der Entwicklung von E-Infrastrukturprojekten in Deutschland. Insbesondere werden die Projekte TextGrid aus den Geisteswissenschaften und C3-Grid aus der Klimaforschung begleitet. Das Projektteam nimmt an den Arbeitstreffen der Projekte teil, ist an den WiKis beteiligt und führt Einzel- und Gruppeninterviews durch. Betrachtet werden sowohl die Software, als auch die soziale Konstellation der Mitarbeiter in dem jeweiligen Projektrahmen. Zentrale Fragen sind hierbei die Arbeitsteilung in den Projekten, die Akzeptanz durch die jeweiligen Communities, sowie Qualität bzw. der Qualitätsbegriff.

#### *Gruppen Domäne*

- Im Bereich der Mikrobiologie wird durch die DFG im Projekt **SILVA 2.0** die Weiterentwicklung der SILVA[210] Datenbank gefördert. SILVA stellt eine automatische Software-Pipeline für die Sequenzierung von ribosomaler Ribonukleinsäure (rRNA) bereit. Die Software ist hierbei kompatibel zum in der Mikrobiologischen Community verbreiteten ARB Softwarepaket. Projektpartner sind das MPI für marine Meeresbiologie, die TU München und die MPI Ausgründung Ribocon. Inzwischen enthält SILVA rund 2,8 Millionen Einträge. Es wird hierbei zwischen Referenzdatensätzen, welche höchste Qualitätskriterien erfüllen müssen und auch als internationale Benchmarks dienen,

[208] Projektwebseite http://www.pubflow.uni-kiel.de
[209] Projektwebseite http://escience-interfaces.net
[210] Projektwebseite http://www.arb-silva.de

und Parkdatensätzen unterschieden, an die geringere Qualitätsanforderungen gestellt werden. SILVA bündelt Daten der drei Institutionen INSDC (International Nucleotide Sequence Database Collaboration), EMBL-EBI/ENA (European Molecular Biology Laboratory, European Bioinformatics Institute/European Nucleotide Archive) und DNA Data Bank of Japan zusammen mit Zusatzinformationen. Die Speicherung der Daten geschieht im Hause. Das zurzeit laufende Projekt soll die bestehende SILVA Website anhand des Feedbacks von Workshops und der Mailingliste weiterentwickeln. Die umfasst primär eine bessere Suchfunktion, weitere Tools und die Möglichkeit der personalisierten bzw. standardisierten Datenabfrage. Außerdem soll die Zusammenarbeit mit öffentlichen Repositorien verbessert werden. Auch soll eine neue Pipeline zur Datenextraktion aus den Repositorien geschaffen werden.

- Im verwandten, ebenfalls von der DFG geförderten Projekt **ARB in the age of high throughput sequenzing: adaption to the requirements of large scale environmental and metagenomic studies and maintenance of the respective databases**[211] wird, auch aufgrund der Erfahrungen mit SILVA, an der Verbesserung des ARB Softwarepaketes gearbeitet. Insbesondere ist an der seit 1992 kontinuierlich erweiterten Software eine Bereinigung des Quellcodes erforderlich.
- Im **TR62 Companion Technology** wird im Projekt Z3[212] eine Architektur für die im TR untersuchten Companion-Systeme (kognitive technische Systeme, die ihre Funktionalität vollkommen individuell auf den jeweiligen Nutzer abstimmen) entwickelt. Darauf aufbauend wird eine Experimentierplattform zum Aufbau solcher Systeme aufgebaut. Weitere Aufgaben sind das zentrale Datenmanagement für den TR.
- Im stark interdisziplinären **SFB 806 Our Way to Europe: Culture-Environment Interaction and Human Mobility in the Late Quaternary** wird in einem INF Projekt durch die Bereitstellung einer Spatial-Database die Datenspeicherung und der Austausch von Daten zwischen den Mitgliedern aus Geologie, Geographie und Archäologie gewährleistet. Spezielles Augenmerk liegt auf der Nachnutzung der Daten auch nach Ende des SFBs.[213]
- Das von der DFG geförderte Projekt **Extension and modification of Morph D Base producing a system for permanent storage and documentation of volume data of biological objects in high resolution** hat die Aufgabe **Morph D Base**[214], die zentrale Datenbank des DFG SPP 1174 „Phylogeny of Animals - Deep Metazoan Phylogeny“[215], für die Verwaltung großer Mengen dreidimensionaler Daten zu optimieren. Dies wird begleitet von intensiver Kommunikation mit den Nutzern. Nach Projektende soll die Datenbank von Museum König Bonn als Kommunikationsplattform und Speichersystem weiterbetrieben werden.

[211] Webseite des ARB Projektes http://www.arb-home.de
[212] Projektwebseite http://www.uni-ulm.de/in/sfb-transregio-62/teilprojekte/z3.html
[213] Vgl. http://www.sfb806.uni-koeln.de/images/sfb806/projects/poster/806_Z2_Poster.jpg, online.
[214] Webportal https://www.morphdbase.de
[215] Webseite des SFB http://www.deep-phylogeny.org

- Im **TRR 51 Roseobacter** wird durch ein INF Projekt[216] eine zentrale Analyse-Pipeline von Software-Tools und Datenbanken bereit gestellt. Die Abfrage und Integration der Forschungsdaten, sowie deren Analyse wird in einer standardisierten Form sichergestellt. Zusätzlich werden den Mittgliedern des TRR Support und Schulungen angeboten.
- Im Projekt INF[217] des **SFB 990 Ökologische und sozioökonomische Funktionen tropischer Tieflandregenwald-Transformationssysteme** wird am Web-GIS basierten Datenmanagementsystem **EFForTS-IS** gearbeitet. Hierbei wird auf das System BExIS aus den Biodiversitäts-Exploratorien (siehe Abschnitt *Biodiversität*) aufgebaut. Die spezifischen Anforderungen werden in einer Serie von Workshops zusammen mit den am SFB beteiligten Forscherinnen und Forschern erarbeitet. Teil des Systems ist auch eine Komponente zur Langzeitarchivierung und zur zitierbaren Nachnutzung. Weiter werden diverse Services zur wissenschaftlichen Kommunikation innerhalb der Kollaboration bereitgestellt.
- Im Bereich der Tiermedizin wird im **SFB 852 Ernährung und intestinale Mikrobiota - Wirtsinteraktionen beim Schwein**[218] in einem INF Projekt den Mitgliedern des SFBs verschiedene, an das Schwein adaptierte Untersuchungstechniken angeboten. Dies umfasst unter anderem Software für Qualitätskontrolle, Datenanalyse und Dateninterpretation.

### *Dauerhafte Domäne*

- Im Bereich der Biologie arbeitet das von der DFG geförderte Projekt **Development of the Golm Metabolome Database as a central plant metabolomics information resource** an der Weiterentwicklung der Golm Metabolome Database[219], einer Datenbank zum Metabolom (Stoffwechsel) von Pflanzen. Teil der Arbeit wird die Integration von sog. Metabolite Profile Datasets und einer darauf angepassten Präsentation sein.

### *Nachnutzung*

- In der Mikrobiologie wird im von der DFG geförderten **CoRS** Projekt ein automatisiertes Molekülrecherchesystem entwickelt das Informationen aus aktueller Literatur extrahiert und präsentiert. Das Projekt ist als Nachfolge-Projekt zu *Compounds In Literature* geplant, welches alle Abstracts aus PubMed zusammen mit weiteren Schlüsselinformationen gespeichert hat. CoRS durchsucht nun Abstracts nach Schlüsselwörtern, eine Volltextsuche ist geplant. Eine Zusammenarbeit mit Bibliotheken ist vorgesehen.
- Mit dem Projekt **Informationssystem Werkstoffwissenschaften** wird von der DFG der Aufbau einer dezentralen Dateninfrastruktur, in welche die diversen Forschungs-

[216] Projektübersicht http://www.roseobacter.tu-bs.de/ProjektbereichZ
[217] Projektübersicht http://www.uni-goettingen.de/de/412103.html
[218] Projektübersicht http://www.sfb852.de/Teilprojekte/Projektbereich_C/index.html
[219] Projektwebseite http://gmd.mpimp-golm.mpg.de

daten in den Werkstoffwissenschaften überführt werden können. Hierbei sollen die Daten jeweils in den Instituten gespeichert werden die sie auch erzeugen. Gemeinsam mit den Fachwissenschaftlern soll ein systematischer Workflow abgeleitet werden. Darauf aufbauend wird eine Open-Source Softwarelösung entwickelt, welche Standards unterstützt und Schnittstellen zu anderen Datenbanken bzw. Suchmaschinen bietet. Die Daten werden, solange sie nicht aus Industrieprojekten stammen, offen zugänglich sein. Für die Qualität der Daten ist jeweils der Wissenschaftler verantwortlich, der die Daten in das System einstellt. Partner des Projektes sind das TZI[220], das IWT[221] und das BIBA[222].

## Schlussfolgerungen

Im Folgenden werden die in der Bestandsaufnahme gefundenen Entwicklungen in verschiedenen Bereichen der Forschungsdateninfrastruktur diskutiert und daraus abgeleitete Handlungsempfehlungen vorgestellt.

### Forschungsdatenmanagement in den einzelnen Disziplinen

Die Akzeptanz der Ergebnisse in den Projekten zum Datenmanagement in den einzelnen Communities, wie beispielsweise Formate, Standards und Workflows, aber auch bereitgestellte Infrastruktur, ist immer noch verbesserungsbedürftig. Daher sollte bei der Bewilligung von zukünftigen Projekten mehr noch auf die Nutzung von existierenden, in der Community etablierten Standards und Formaten geachtet werden. Als eine Maßnahme, um die Verankerung innerhalb der Fachdisziplin zu verbessern, sollten neue Projekte dazu angehalten werden, die Kompetenz der diversen Gremien in dieser Fachdisziplin bzw. in den relevanten Wissenschaftsorganisationen stärker zu nutzen. Dies sollte auch den betreffenden Gremien und Organisationen klarer kommuniziert werden.

In den meisten Fach-Communities fehlen klare Analysen der grundlegenden wissenschaftlichen Workflows. Nur durch Konzentration auf solche existierenden Workflows ist es möglich, Werkzeuge zu entwickeln oder aus dem bestehenden Angebot auszuwählen, die einerseits einen Effizienzgewinn liefern und andererseits durch die Wissenschaftlerinnen und Wissenschaftler in der jeweiligen Community akzeptiert werden. Ein Beispiel dafür ist der in Abb. 4 abgebildete Workflow für den Bereich der beobachtenden Astronomie.

[220] Technologiezentrum Informatik, Webseite http://www.tzi.de.
[221] Stiftung Institut für Werkstofftechnik, Webseite http://www.iwt-bremen.de.
[222] Bremer Institut für Produktion und Logistik GmbH, Webseite http://www.biba.uni-bremen.de.

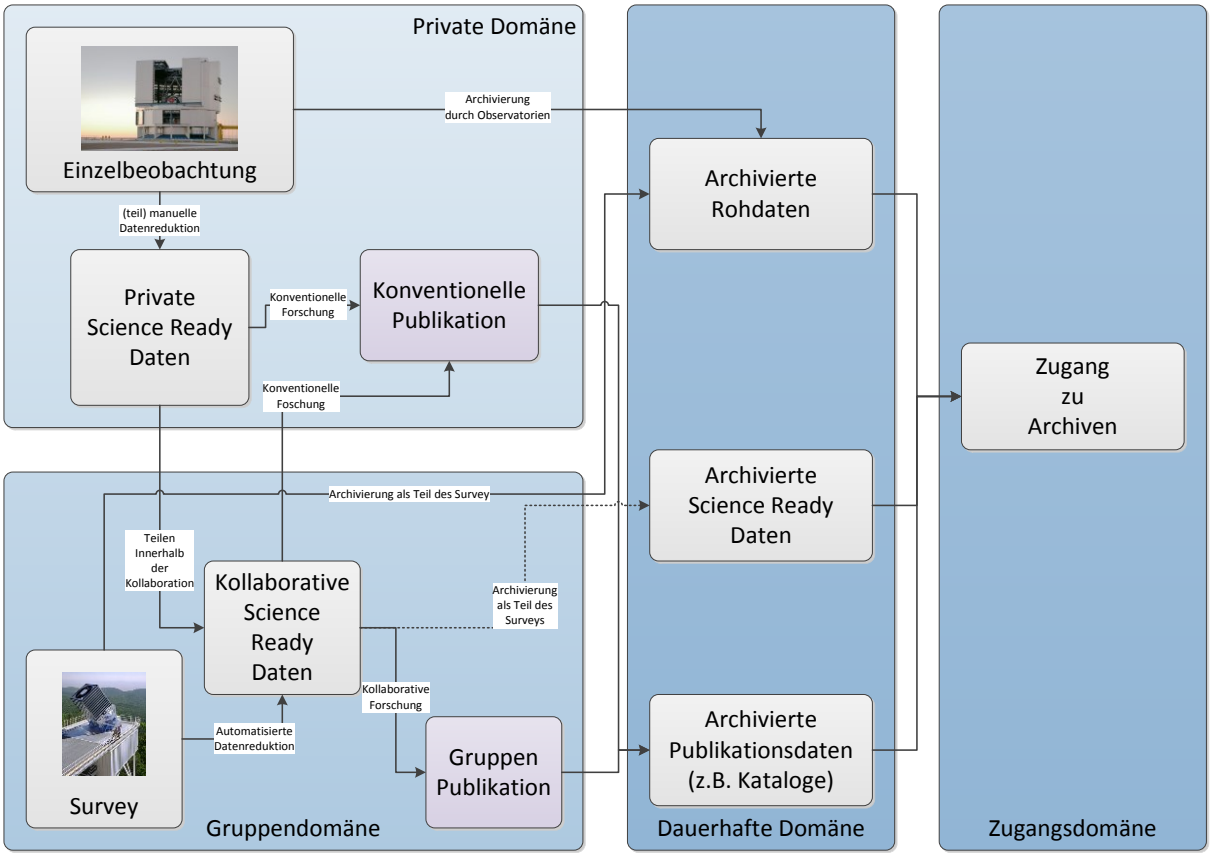


**Abbildung 4:** Prototypischer Daten-Workflow für die beobachtende Astrophysik. Die Daten werden entweder in der privaten Domäne oder, beispielsweise bei Surveys, bereits in der Gruppendomäne erzeugt. Durch Weiterverarbeitung entstehen die sog. Science-Ready Daten, die dann weiter als Grundlage für Publikationen dienen, archiviert werden (dauerhafte Domäne) und gegebenenfalls der Öffentlichkeit zugänglich gemacht werden (Zugangsdomäne).

In den Fachdisziplinen sollten gezielt Maßnahmen gefördert werden, welche die Sichtung, Klärung und Herausbildung von geeigneten disziplinbezogenen Metadaten-Systemen fördern, wobei der Schwerpunkt von den fachlichen Aspekten bestimmt wird. Bei der (Weiter-) Entwicklung der Metadatensystemen und Implementierung sollte ein klarer Fokus auf Auffindbarkeit und Nachvollziehbarkeit (oder *Provenance*) gelegt werden. Komplexe, darüber hinausgehende Systeme verlieren mit ihrem zu spezifischen Vokabular zu schnell Wiederverwendbarkeit und Aktualität. Metadaten sollten auch weitgehend durch die Werkzeuge selbst bereitgestellt werden.

## Projekte zur Forschungsdateninfrastruktur

Als weiteres Problem ist zu nennen, dass oft die auch vom Förderer gewünschte Betonung der interdisziplinären Komponenten von Projekten eine mangelhafte Verankerung in der Fach-Community selbst evoziert bzw. die hierzu notwendigen Anstrengungen in den Hintergrund drängt.

Es fällt außerdem auf, dass innerhalb der von uns untersuchten Projekte die dauerhafte Domäne und die Zugangsdomäne zahlenmäßig überrepräsentiert sind. Es scheint daher nötig, die Förderung künftig stärker auf Konzepte zu fokussieren, welche deutlicher die

Bereiche miteinbeziehen, in denen aktiv mit Daten gearbeitet wird (private und Gruppendomäne).

Es wird aber auch sichtbar, dass die großen, insbesondere auch im europäischen Kontext geförderten Projekte, die vor allem einen fachübergreifenden Kontext adressieren, noch zu wenig durch die Disziplinen selbst gestützt sind, die sich erst in den Anfangsstadien des Forschungsdatenmanagements befinden und denen disziplinspezifische Strukturen für das Datenmanagement fehlen. Dies gilt nicht in gleichem Maße für Klimawissenschaften und die Biodiversität, die stark politische und ökonomische relevante Themen als Hintergrund haben, und für stark organisierte Disziplinen mit wenigen Datenquellen wie die Hochenergiephysik.

Im Allgemeinen wird die Frage, inwieweit Projektergebnisse nachhaltig verfügbar gehalten werden können, stärker durch die äußeren infrastrukturellen Bedingungen bestimmt als durch die Ergebnisse der Projektarbeit selbst. Die von der DFG geförderten INF-Projekte sind hierfür ein gutes Beispiel: Primäres Ziel dieser Projekte ist das Management der Datensammlungen eines SFB über dessen Laufzeit. Danach erwartet die DFG, dass die Hochschule der am SFB beteiligten Institute diese Aufgabe übernimmt. Aus der Förderperspektive ist dies zwar sinnvoll, geht jedoch an der Wirklichkeit vorbei, insofern als die Hochschulen derzeit erst in wenigen Fällen dafür gerüstet sind, eine solche Aufgabe zu bewältigen. Auch ist die Frage zu stellen, inwieweit disziplinweite Infrastrukturen nicht effizienter diese Aufgabe übernehmen können. Dies ist beispielsweise bei den Daten der Klimaforschung schon der Fall, oder auch in den Sozialwissenschaften. Die Entwicklung derartiger Rahmenbedingungen geht über die Reichweite und das Vermögen der INF-Projekte weit hinaus.

## Sonderforschungsbereiche und Transregio

Die von der DFG geförderten INF-Projekte setzen an der Stelle an, wo derzeit, wie schon weiter oben genannt, die größte Lücke im Forschungsdatenmanagement festgestellt werden kann: der Gruppendomäne bzw. dem Übergang in diese. In Zukunft sollte jeder SFB/TRR eine solche Infrastrukturkomponente aufweisen. Nur so ist es möglich, die in den Projekten erarbeiteten Forschungsdaten nachhaltig zukünftigen Forschungen zugänglich zu halten. Eine starke Beteiligung der im SFB/TRR arbeitenden Fachwissenschaftlerinnen und Fachwissenschaftler an der Konzeption der INF Projekte ist anzustreben.

Generell sollten für die in den INF-Projekten verwendeten Komponenten nicht die Neuartigkeit oder Einzigartigkeit entscheiden sein, sondern die Effizienz der nötigen Prozesse und Einbindung in den infrastrukturellen Kontext. Die INF-Komponenten sollten auch nicht den Versuch beinhalten, die Infrastrukturen des gesamten Fachgebiets gleich mit zu erschaffen oder sich zu stark auf die Entwicklung von generischen Komponenten zu konzentrieren. Die Wiederverwendbarkeit von entstehenden Tools ist jedoch selbstverständlich zu begrüßen.

Der Weiterbetrieb von INF-Projekten über das Projektende hinaus ist in der Regel nicht gesichert. Nachhaltigkeit wird von den Projekten meist nur als einer unter vielen weiteren Aspekten behandelt. Die am Projekt arbeitenden Wissenschaftlerinnen und Wissenschaftler

besitzen außerdem primär Kompetenz im Aufbau von Infrastrukturen und nicht notwendigerweise im Bereich Nachhaltigkeit. Es sollte daher ein Kompetenznetzwerk zum Thema Nachhaltigkeit mit Expertinnen und Experten aus verschiedenen Disziplinen aufgebaut werden. Dieser Personenkreis sollte als Projektpartner für dezidierte Nachhaltigkeits-Projekte aus den Communities zur Verfügung stehen. Die Initiative muss hierbei aber direkt von den Communities ausgehen, um die individuellen fachspezifischen Gegebenheiten nicht aus dem Blick zu verlieren.

Auch zu anderen Themenkomplexen sollte die Möglichkeit bestehen, sich durch spezialisierte Kompetenznetzwerke beraten zu lassen. Insbesondere zu den Themen Standards und Software-Tools sollte im Vorhinein stärker kommuniziert werden, um unnötige Doppelentwicklungen zu vermeiden.

## IT-Infrastruktur

Ein großes und im Moment gänzlich ungelöstes Problem besteht darin, dass die Forschungs-IT-Infrastruktur und deren Provider sich zu stark auf das Scientific Computing konzentrieren und sich daher dem Problem des Datenmanagements bislang nur unzureichend angenommen haben. In Hinsicht auf eine bessere Unterstützung des Forschungsdatenmanagements ist jedoch eine Flexibilisierung der gegenwärtigen Strukturen notwendig. Einige Ansätze zur Flexibilisierung gibt es bereits, jedoch sind es noch Insel-Lösungen.

Mit dem D-Grid wurden einige wichtige Grundvoraussetzungen der Flexibilisierung adressiert und der Versuch gestartet, eine effizientere IT-Infrastruktur hierfür zu schaffen. Es hat sich jedoch gezeigt, dass noch erhebliche Probleme existieren, die Bereitstellung von IT-Ressourcen zu flexibilisieren.

Obschon es Anstöße in die richtige Richtung gegeben hat, konnte das D-Grid eine wichtige Hürde nicht überwinden: die traditionelle Organisation der IT-Infrastruktur in der Wissenschaft ist weitgehend im Versuch erstarrt, für ihren jeweiligen Bereich alle IT-Services aus einer Hand anzubieten. An den Universitäten umfasst dies die Bereitstellung der Arbeitsumgebung für die Wissenschaftler und an den Rechenzentren -komplementär- nur das Computing. Die Überschneidungen zwischen diesen Bereichen sind minimal. Natürlich existieren vereinzelt Flexibilisierungsansätze, jedoch gilt im Wesentlichen, dass das Forschungsdatenmanagement flexiblere Strukturen erfordert, als die derzeitige Landschaft der IT-Provider im wissenschaftlichen Bereich bietet.

Das D-Grid hat auch offengelegt, dass es noch erhebliche Probleme gibt, die Bereitstellung von IT-Ressourcen adäquat abzurechnen. Stärker als technische Gründe spielten hierbei rechtliche (Dedizierung der IT-Ressourcen in den Institutionen, Landes- und Bundesrecht), soziale (wissenschaftliche Kollaborationen basieren auf und leben vom Austausch) und kulturelle (infrastrukturelle Dienste und Leistungen sind nur in wenigen Disziplinen "verkaufbar") Fragen. Dies ist insbesondere im Zusammenhang mit den vielzähligen Nachhaltigkeitsbemühungen von Infrastrukturprojekten zu bedenken: Es existiert ein Zielkonflikt zwischen

dem wissenschaftlichen Streben nach Gewinnen von Erkenntnis und der betriebswirtschaftlich gewerteten Nachhaltigkeit im Sinne einer sich selbst tragenden Unternehmung.

## Disziplinübergreifende Strukturen

Obwohl die drängendsten Probleme im Bereich des Forschungsdatenmanagements immer noch innerhalb der jeweiligen Disziplin zu verorten sind, sind auch im organisatorischen Bereich disziplinübergreifende Lösungsansätze erstrebenswert. Insbesondere sollte, um die Ergebnisse der vielfältigen abgeschlossenen Projekte optimal zu nutzen, die Kommunikation zum Thema Forschungsdatenmanagement innerhalb der Communities und übergreifend gefördert werden. Während der Arbeit des Radieschen-Projektes wurde seitens Wissenschaftlerinnen und Wissenschaftler immer wieder artikuliert, dass ein großes Informationsbedürfnis über vorhandene Tools und Standards besteht, das zurzeit jedoch nur unzureichend befriedigt werden kann.

In diesem Kontext sollte ein Kompetenznetzwerk oder Expertengremium erwogen werden, welches durch Wissenschaftlerinnen und Wissenschaftler, die Erfahrung im Forschungsdatenmanagement besitzen und in vorangegangenen Projekten beteiligt waren, gebildet wird. Neue Projekte können dann auf diese Netzwerke zurückgreifen und sich beratend unterstützen lassen. Hierbei ist es entscheidend, dass die Initiative von diesen neuen Projekten (bzw. mittelbar durch die jeweilige Förderinstitution) ausgeht und es nicht die Aufgabe des Netzwerkes ist, neue „Kunden“ zu akquirieren.

Auch im Bereich der Nachhaltigkeit könnte sich ein solches Expertengremium als nützlich erweisen. In der Regel wird von Projekten zum Forschungsdatenmanagement ein Konzept zur Nachhaltigkeit verlangt. Da die am Projekt beteiligten Wissenschaftlerinnen und Wissenschaftler jedoch ihre primären Kompetenzen in der Fachdisziplin oder im Bereich der IT-Infrastruktur haben, verfügen sie nicht notwendigerweise über die erforderlichen Kenntnisse zum Thema Nachhaltigkeit. Als Querschnittthema lässt sich dieses Thema auch besonders gut disziplinübergreifend vermitteln.