

Forschungsdatenmanagement (FDM) in den Natur- und Umweltwissenschaften

Klaus-Dieter Warzecha

2017-06-15

Forschungsdaten

Definitionen

Deutsche Forschungsgemeinschaft (DFG), 2015

„[...] u. a. Messdaten, Laborwerte, audiovisuelle Informationen, Texte, Surveydaten, [...], die in der wissenschaftlichen Arbeit entstehen, entwickelt oder ausgewertet werden. Methodische Testverfahren, wie Fragebögen, Software und Simulationen [...].“

Allianz der deutschen Wissenschaftsorganisationen, 2010

„[...] Daten, die im Zuge wissenschaftlicher Vorhaben z.B. durch Digitalisierung, Quellenforschungen, Experimente, Messungen, Erhebungen oder Befragungen entstehen.“

OECD, 2007

“[...] factual records (numerical scores, textual records, images and sounds) used as primary sources for scientific research [...].”

Forschungsdaten

Lebenszyklus

- ▶ Erzeugung (Laborversuch, Feldversuch, Simulation, Interview)
- ▶ Verarbeitung

in der Vergangenheit

- ▶ in geringen Teilen: unmittelbare Nutzung in Publikationen
- ▶ Daten größtenteils unzusammenhängend verstreut über
 - ▶ Laborjournale
 - ▶ Aktenordner
 - ▶ Dateien auf Mess- und Desktop-Computern
- ▶ keine langfristige Nutzung

Forschungsdaten

Lebenszyklus

- ▶ Erzeugung (Laborversuch, Feldversuch, Simulation, Interview)
- ▶ Verarbeitung

in Zukunft

- ▶ obligatorisch zur Unterstützung der Hypothesen in Publikationen
- ▶ interne Nachnutzung durch nachrückende Mitarbeiter
- ▶ gemeinschaftliche (Nach)-Nutzung in Kooperationen
- ▶ data sharing in öffentlich zugänglichen Repositorien

Forschungsdatenmanagement

Richtlinien

DFG

im Projektantrag:

- ▶ Datenmanagementplan
- ▶ Auswahl adäquater Datenformate
- ▶ Forschungsdaten so zeitnah wie möglich verfügbar machen
- ▶ Forschungsdaten für mindestens 10 Jahre archivieren

BMBF

Forschungsdatenmanagement nach den **FAIR**-Prinzipien

- ▶ **F**indable — auffindbar
- ▶ **A**ccessible — zugänglich
- ▶ **I**nteroperable — kompatibel
- ▶ **R**eusable — wiederverwendbar

Lizenzmodelle für Forschungsdaten

Creative Commons

- ↗ Attribution 4.0 International (CC BY 4.0)
- ↗ Attribution-ShareAlike 4.0 International (CC BY-SA 4.0)

Open Knowledge International

- ↗ Open Data Commons Attribution License (ODC-By)
- ↗ Open Data Commons Open Database License (ODbL)

Persistente Identifier für Wissenschaftler

Herausforderung

Jede Ressource ist mit Personen (in unterschiedlichen Rollen) verbunden:

Namen als Identifier?

Personenverzeichnis Uni Vechta

Nachname	Häufigkeit
Kaiser	4
Meyer	4
Müller	4

Beliebteste deutsche Vornamen

Zeitraum	weiblich	männlich
1980–1989	Julia	Christian
1990–1999	Anna	Jan
2000–2009	Anna	Lu[ck]as

Namen sind keine eindeutigen Identifier!

Persistente Identifier für Wissenschaftler

Anforderungen

Eindeutige ID für jede beteiligte Person

- ▶ unabhängig von Schreibweise und Wechsel des Namens
- ▶ unabhängig von Hochschulwechsel
- ▶ in jedem Stadium des FDM verwendbar
- ▶ bereits international etabliert (**keine In-House-Lösung**)
- ▶ **in XML Schema Definition für Metadaten verwendbar**

Persistente Identifier für Wissenschaftler

ORCID

ORCID (Open Researcher and Contributor ID)

- ▶ Beginn: Oktober 2012; aktuell ca. 3.5 Millionen IDs
- ▶ kostenlos
- ▶ Open Source (MIT-Lizenz)

bei Publikationen unterstützt oder **zwingend vorgeschrieben** von:

- ▶ American Chemical Society (ACS)
- ▶ European Molecular Biology Organization (EMBO)
- ▶ Public Library of Science (PLOS)
- ▶ Royal Society of Chemistry (RSC)
- ▶ Scopus (Elsevier, auch nachträglich)

orcid.org/0000-0002-1247-4508 Gunther Schmidt (Uni Vechta)

orcid.org/0000-0001-6912-3234 Klaus-Dieter Warzecha

Dateiformate

Kriterien:

- ▶ nicht-proprietär, vollständig dokumentiert
- ▶ maschinell (Python, Perl, R, Java, etc.) verarbeitbar

Beispiele

- ▶ CSV, HDF5, NetCDF (tabellarische Daten)
- ▶ CIF, PDBx/mmCIF, NMR-STAR (Chemie, Kristallographie)
- ▶ CML, JCAMP-DX (Chemie)
- ▶ FASTA, FASTQ (Bioinformatik)
- ▶ TIFF, PNG (Bilder)
- ▶ OGG/Opus (Audio)
- ▶ WebM/VP9/Vorbis (Video+Audio)

Keine Daten ohne Metadaten

Funktion der Metadaten

- ▶ beschreiben den Datensatz
- ▶ sind indexierbar
- ▶ sind die Grundlage für die Zuweisung eines persistenten **DOI** (Digital Object Identifier)

Realisierung

- ▶ XML-Datei mit definierten Elementen
- ▶ Elemente und erlaubte Werte in XSD definiert

de-facto-Standard: **DataCite Metadata Schema** [!\[\]\(a03a7eb2f4046e1d3c76772003e549ea_img.jpg\)](#)

Deutsche DOI-Vergabestellen

Domänenspezifische Mitglieder von DataCite

- ▶ TIB – Technische Informationsbibliothek
Technik, Architektur, Chemie, Informatik, Mathematik und Physik
- ▶ ZB MED – Leibniz-Informationszentrum Lebenswissenschaften
Medizin, Gesundheitswesen, Ernährungs-, Umwelt- und Agrarwissenschaften
- ▶ SUB – Niedersächsische Staats- und Universitätsbibliothek
Geisteswissenschaften
- ▶ da|ra – Registrierungsagentur für Sozial- und Wirtschaftsdaten
Sozial- und Wirtschaftswissenschaften

Alle DOI-Vergabestellen nutzen das DataCite Metadata Schema!

Repositorien für Forschungsdaten

Existierende Repositorien

- [✚](#) Zenodo (allgemeines Repository am CERN)
- [✚](#) PUBLISSO
- [✚](#) re3data (Registry of Research Data Repositories)
- [✚](#) OAD (Open Access Directory)

Eigenes institutionelles Repository

- [✚](#) DSPACE
- [✚](#) Samvera (Hydra)
- [✚](#) Isladora

Erste Schritte

1. Bedarfsermittlung bei beteiligten Fachbereichen
 - ▶ verwendete Datenformate und mögliche Alternativen
 - ▶ mögliche Workflows für Konvertierung/Transcodierung von Daten
2. ORCID-Kampagne für Wissenschaftler und Studierende
3. Evaluierung verschiedener Plattformen für Digital Asset Management (institutionelle Repositorien)
4. Online-Informationsportal
5. FDM in bestehendes OTRS einbinden
6. Schulungen