Mit maßgeschneiderten Metadatenprofilen zu validierten und nachhaltigen Forschungsdaten

Matthias Grönewald [1], Nils Preuß [2]

[1]Universitäts- und Landesbibliothek, Technische Universität Darmstadt;
[2]Institut für Fluidsystemtechnik, Technische Universität Darmstadt

Zeitgemäßes Forschungsdatenmanagement (FDM) beinhaltet zunehmend auch die Integration reichhaltiger, maschinennutzbarer Metadaten, allgemein zur Sicherstellung wissenschaftlicher Qualität insbesondere aber im Kontext von Reproduzierbarkeit und Nachnutzung. Bestehende Standards umfassen meist nur deskriptive Metadaten und die in umfassenderen Metadatenschemata enthaltenen Informationen sind in der Regel weder standardisiert noch maschinennutzbar. Fachspezifische Metadaten sind jedoch notwendig, um Forschung präzise und reichhaltig zu dokumentieren. Die Abläufe und Werkzeuge dafür sind jedoch nicht umfassend verfügbar. Ein vielversprechender Ansatz ist die Anwendung von Metadatenprofilen, die es ermöglichen hochspezifische Terminologien, aufbauend auf bestehenden Community-Standards, in flexible und interoperable Metadatenbeschreibungen zu überführen. Basierend auf etablierten Technologien ermöglichen Metadatenprofile eine Lösung zum Gestalten und Verarbeiten von komplexen, maschinennutzbaren und letztlich FAIRen Metadaten.

Anhand eines Beispiels aus den Ingenieurswissenschaften, wird die Datenvalidierung mittels Metadatenprofilen basierend auf kontrolliertem Vokabular gezeigt. Dieser Prozess kann dann zu fast jedem Zeitpunkt im Lebenszyklus von Forschungsdaten genutzt werden. Ein Anwendungsbeispiel demonstriert außerdem die sich daraus ergebenden Möglichkeiten im Bereich der Datenanalyse bzw. der Archivierung.

Erst die Kombination aus praktischer Integration in die Forschungslandschaft in Verbindung mit der Umsetzung in verschiedenen FDM-Projekten, Initiativen und Werkzeugen ermöglicht die für eine Standardisierung notwendigen Synergien. Damit wird die Forschung durch höhere Datenqualität gefördert, sowie für die nachhaltige Bewahrung von Forschungsinhalten durch spezifischere Dokumentation ein Mehrwert gebildet.

Publiziert in: Vincent Heuveline, Nina Bisheh und Philipp Kling (Hg.): E-Science-Tage 2023. Empower Your Research – Preserve Your Data. Heidelberg: heiBOOKS, 2023. DOI: https://doi.org/10.11588/heibooks.1288.c18080

1 Einleitung, Ziele und Bedarf

Metadaten, allgemein oft als „Daten über Daten“ oder „Daten die Daten beschreiben“ (Furner 2019) charakterisiert, sind in der Forschung von entscheidender Bedeutung. Sie beschreiben Zusammenhänge, wie zeitliche Bezüge, administrative Zuordnung zu Projekten oder auch Urheberschaft. Neben solchen weit verbreiteten deskriptiven, administrativen oder bibliographischen Metadaten, die sich in der Forschungslandschaft weitgehend standardisiert etabliert haben, sind unterschiedliche Informationen über eingesetzte Methoden, Software, Materialien und Abläufe (Abbildung 1) notwendig um Forschung nachvollziehbar und reproduzierbar zu machen (Deutsche Forschungsgemeinschaft e.V. 2019). Damit sind sie zwar von mindestens genauso entscheidender Bedeutung, werden jedoch in der Praxis allerdings weit weniger wahrgenommen.

Um erstgenannte Metadaten zu erfassen und beschreiben liegen standardisierte Terminologien und Metadatenschemata vor[1], auch im Bezug auf Forschungsdaten (Grönewald u. a. 2023a; Albertoni u. a. 2023; DataCite Metadata Working Group 2021), für nachfolgend genannte fachspezifische Metadaten jedoch nicht. Auch ist nicht klar ob die moderne stark heterogene Forschung, selbst innerhalb einer Disziplin eine solche Standardisierung zulässt. Gerade das Forschungsdatenmanagement von großen Datenmengen stellt allerdings an Dokumentation und Datenbeschreibung umfassende Herausforderungen. Forschende, die damit konfrontiert sind, gestalten dann eigene Lösungen, teilweise mit hoch individuellen Metadatenschemata, die in Ausgestaltung und Betrieb aufwendig und in der Regel nicht auf andere Forschende übertragbar sind. Letzteres ist besonders mit Blick auf Nachnutzung, aber auch Archivierung, von Forschungsdaten nachteilig. Im Rahmen des DFG-Projekts AIMS (Grönewald u. a. 2023b) wurde eine Softwareplattform entwickelt, die es erlaubt stattdessen interoperable und trotzdem hochspezifische Metadatenprofile zu gestalten und zu teilen.

Metadatenprofile, technisch als Applikationsprofile umgesetzt, erlauben es durch die Beschreibung von Anforderungen Eigenschaften von Metadatensätzen festzulegen bzw. gemäß diesen zu validieren (Coyle 2017). Neben dem Bedarf bei der Handhabung, Analyse und Dokumentation von Daten in der aktiven Forschung, sind fachspezifischen Metadaten insbesondere auch notwendig zur Erstellung und Veröffentlichung von Datensätzen gemäß den FAIR Prinzipien (Wilkinson u. a. 2016). Am deutlichsten wird das an der Nachnutzbarkeit von Forschungsdaten, die durch detaillierte Metainformationen insbesondere verbessert wird, aber durch die Bereitstellung von Metadatenprofilen gemäß etablierter Standards (RDF, SHACL, etc.) werden auch Auffindbarkeit, Zugänglichkeit und Interoperabilität gestärkt. Letztlich umfasst der Anwendungsbereich damit den gesamten Forschungsdatenlebenszyklus. Anhand von Beispielen aus den Ingenieurwissenschaften wird gezeigt, in welcher Weise maßgeschneiderte Metadatenprofile es erlauben Forschungsmetadaten gemäß den FAIR-Prinzipen zu beschreiben und diese in im aktiven Forschungsdatenmanagement zur Datenvalidierung zu nutzen.

1 Vgl. (a) `http://dublincore.org/documents/dcmi-terms/`; (b) `http://www.rdaregistry.info/Elements/u/#`, (c) `https://github.com/tibonto/DFG-Fachsystematik-Ontology`, uvm., *zuletzt aufgerufen am 12. Mai 2023.*

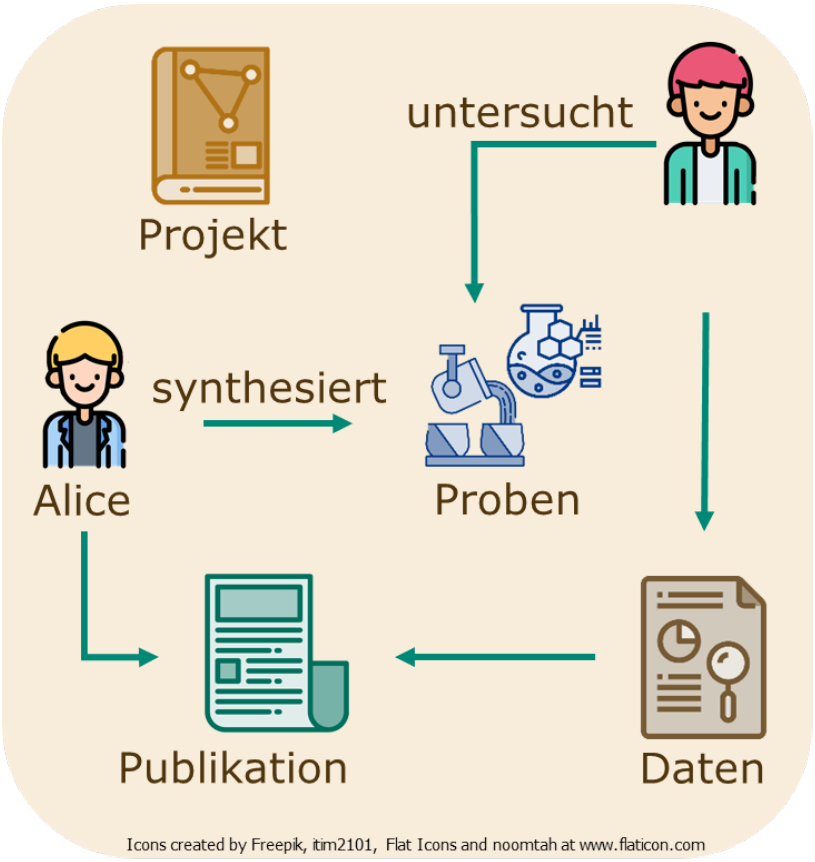

Abbildung 1: Forschungsmetadaten - Neben deskriptiven und administrativen Angaben zu Projektbezug, Forschenden oder Forschungsdisziplin, stellen die Bezügen zwischen einzelnen Forschungsdaten, sowie Metadaten über angewandte Parameter, eingesetzte Geräte oder Analyseverfahren, wichtige Informationen zum Verständnis der eigentlichen Inhalte dar. In aller Regel unterliegen sie hochspezifischen disziplinabhängigen Bedingungen. Es handelt sich um fachspezifische Forschungsmetadaten.

2 Reichaltige fachspezifische Metadaten erlauben FAIRe Inhalte

Für die (Nach-) Nutzung von Forschungsdaten, das Vorhandensein von beschreibenden Forschungsmetadaten zentral. Sie werden benötigt um eine Bewertung der Beschaffenheit der Daten und ihrer Eignung für das jeweilige Nutzungsszenario zu erlauben, und nehmen daher auch im Kontext der FAIR-Prinzipien eine besondere Stellung ein.

Es ist also zweckdienlich, zunächst die Erfüllung formaler Kriterien sicherzustellen, d.h. zu validieren. Dies wird bereits am einfachen Beispiel eines Messwertes ersichtlich: Unabhängig vom Nutzungsszenario ist es für ingenieurwissenschaftliche Versuche unerlässlich, dass für jede Messgröße die physikalische Dimension (Länge, Masse, Zeit, etc.), und für jeden Messwert die physikalische Einheit (Meter, Kilogramm, Sekunde, etc.) unmissverständlich angegeben ist. Gleiches gilt etwa für den Zeitpunkt und die genaue Messstelle, außerdem die Herkunft des Messwertes, etwa Messinstrument und Messverfahren, bzw. Messunsicherheit. Ist die Erfüllung solcher Kriterien nicht sichergestellt, ist eine Beurteilung der Eignung der Daten für ein bestimmtes Nutzungsszenario nicht möglich oder zumindest deutlich erschwert.

```
soil:MeasurementShape
    a sh:NodeShape ;
    sh:name "measurement"@en ;
    sh:property [
       sh:path qudt:hasQuantityKind ;
       sh:name "quantity"@en ;
       sh:minCount 1 ;
       sh:maxCount 1 ;
       sh:class qudt:QuantityKind ;
    ] ;
    sh:property [
       sh:path qudt:applicableUnit ;
       sh:name "unit"@en ;
       sh:minCount 1 ;
       sh:maxCount 1 ;
       sh:class qudt:Unit ;
    ] .
```

```
ex:TemperatureReading
    qudt:hasQuantityKind quantitykind:Temperature ;
    qudt:applicableUnit unit:DEG_C ;
    schema:value "20.54513"^^xsd:float .
```

Abbildung 2: Links das Metadatenprofil einer Messung, rechts ein valider Messwert eines Temperatursensors. Das Profil bedingt dabei die zwei Eigenschaften, das Vorhandensein einer Größe samt Einheit.

Wenn gleich der manuelle Aufwand und die benötigte fachliche Expertise zur letztlichen Bewertung von Daten nur schwer reduzierbar ist, kann die Überprüfung von formalen Kriterien, beispielsweise das bloße Vorhandensein bestimmter bewertungsrelevanter Informationen, aber auch anderer Aspekte der FAIR-Prinzipien maschinell stattfinden. Für eine solche automatisierte Validierung werden formalisierte Vorgaben benötigt, z.B. in Form von maschinennutzbaren Metadatenprofilen. Im Rahmen des vorgestellten Projekts werden RDF (Schreiber und Raimond 2023) und SHACL (Knublauch und Kontokostas 2023) hierfür als Basis-Technologien eingesetzt. Hierdurch werden bereits einzelne FAIR Prinzipien (Wilkinson u. a. 2016) adressiert (im folgenden referenziert durch ihre Nummerierung gemäß der Veröffentlichung, z.B. F1 für das erste FAIR Prinzip).

Der Einsatz einer formalen, zugänglichen, gemeinsamen und breit anwendbaren Sprache zur Wissensdarstellung ist sichergestellt (I1). Zugleich sind die Beschränkung auf Vokabu-

lare und Terminologien, die den FAIR-Prinzipien genügen (I2), und Verweise auf andere (Meta-) Daten (I3) einfach realisierbar.

Darüber hinaus kann analog zum oben erläuterten Beispiel die Angabe von Metadaten überhaupt (F2), bis hin zu bestimmten relevanten Attributen (R1), inkl. Nutzungslizenz (R1.2) und Herkunft (R1.2) sichergestellt werden. Dies ist individuell auf Konventionen und Standards der jeweiligen Fachrichtung anpassbar (R1.3), hier der Ingenieurwissenschaften. Zusätzlich sind weiter Spezialisierungen, beispielsweise für den Werkzeugmaschinenbaus sind umsetzbar und bleiben interoperabel.

Die Angabe von Identifiern für Datensätze (F4), sowie speziell die Nutzung von global einzigartigen und persistenten Identifiern wie DOI, Handle, w3id, etc. (F1) lassen sich durch geeignete Metadatenprofile überprüfen. In Verbindung mit geeigneten Repositorien können schließlich die übrigen Aspekte der FAIR-Prinzipien abgedeckt werden: Indexierung und Suche (F3), Abruf via Identifier (A1) und standardisierte, offene Protokolle (A1.1), inkl. Authentifizierung und Autorisierung (A1.2), Zugänglichkeit von Metadaten unabhängig von den Daten selbst (A2).

Natürlich ist dies nicht nur auf Daten und ihre Metadaten anwendbar, sondern gilt in gleichem Maße für die Metadatenprofile selbst: als Ressourcen verstanden, sollen sie den FAIR-Prinzipien ebenso genügen.

3 Datenvalidierung mittels formalisierter semantischer Metadaten

Eine direkte Anwendung findet sich nun z.B. bei der Validierung von Datensätzen. Hier Sensordaten eines Temperatursensors, bestehend aus einer Größe Temperature in der Einheit Celsius, gezeigt auf der rechten Seite der Abbildung 2. Ein zugehöriges Profil, dass allgemein einen Messwert beschreibt ist links dargestellt, es verlangt genau eine Größe mit genau einer Einheit. Nicht gezeigt ist hier die Verknüpfung mit allgemeineren Profilen, die wie im obigen Abschnitt beschrieben allgemeinere Aspekte der FAIR-Prinzipien validieren.

Das in Abbildung 2 links gezeigte Profil validiert auch Daten die von einem Lasertracker generiert werden, hier für die Größe Länge und die Einheit Meter, siehe Abbildung 3.

```
ex:TargetPosition
    qudt:hasQuantityKind quantitykind:Length ;
    qudt:applicableUnit unit:M ;
    schema:value (
        "-0.226959617"^^xsd:float
        "-0.3047850568"^^xsd:float
        "15.10344412"^^xsd:float
    ) .
```

Abbildung 3: Das in Abbildung 2 gezeigte Profil einer Messung validiert auch den hier gezeigten Messwert eines Lasertrackers, da die Eigenschaften (Vorhandensein von Größe und Einheit) ebenso erfüllt sind.

Durch eine hierarchische und modulare Modellierung der Metadatenprofile wird eine Individualisierung von Metadatenprofilen ermöglicht. Abbildung 4 zeigt das Profil eines Temperaturmesswertes, basierend auf dem allgemeineren Messwertprofil von zuvor.

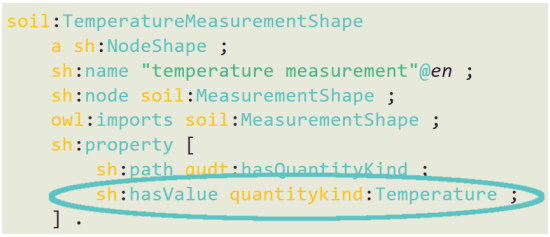

Abbildung 4: Wird zusätzlich die Art der Größe des in Abbildung 1 gezeigten Profils eingschränkt, entsteht ein interoperables aber spezifischeres Profil, das nur noch den in Abbildung 1 gezeigten Messwert, nicht jedoch den des Lasertrackers aus Abbildung 2 validiert.

Hier ist nun zusätzlich vorgeschrieben, dass die physikalische Größe nicht nur angegeben sein muss, sondern außerdem der Art Temperatur sein muss, wodurch das Profil die Temperaturdaten validiert, aber nicht die des Lasertrackers. Es lassen sich also hochspezifische Beschreibungen erstellen, die trotzdem interoperabel sind, da sie gemeinsame Terme und Profile verwenden. Darüber hinaus sind sie maschinennutzbar und so in eine automatisierte Datenvalidierung integrierbar.

Mögliche Ziele kann dabei die Prüfung der Datenqualität anhand von formalen Kriterien bei Datenübergabe innerhalb von Projektstrukturen, vor Erfassung in Repositorien oder Austausch zwischen Forschungsinstitutionen sein, also sowohl noch im Bereich der Forschung, als auch später bei Fragen der Verfügbarmachung und Archivierung.

4 Zusammenfassung & Möglichkeiten in der Zukunft

Metadatenprofile sind auf Basis von RDF und SHACL technologisch umsetzbar und individualisierbar. Sie ermöglichen die automatisierte Überprüfung formaler Kriterien, und bilden damit einen wichtigen Baustein zur breiten Umsetzung der FAIR-Prinzipien und fachspezifischer Standards darüber hinaus. Speziell am Beispiel der Datenvalidierung von Sensorwerten kann gezeigt werden, dass damit hochspezifische Metadaten, wie die Beschreibung physikalischer Größen validierbar erfasst werden können und dabei sowohl untereinander interoperabel, gleichzeitig konform zu bestehenden Normen gehalten werden können. Damit leisten Metadatenprofile auf Basis gemeinsamer kontrollierter Terminologien einen wesentliche Beitrag für die Implementierung der FAIR Prinzipien, insbesondere der Gewährleistung von Interoperabilität.

Das benötigte Hintergrundwissen zur Modellierung, sowie erforderliche Kenntnisse über geeignete Vokabulare, Terminologien und existierende Profile bzw. Informationsmodelle bilden dennoch eine signifikante Einstiegshürde für Forschende die es zu überwinden gilt. Dieser Herausforderung widmen sich aktuell verschiedene Initiativen. Im Projekt AIMS website ist eines der Ziele die Entwicklung einer Plattform zum Gestalten und Teilen von

Metadatenprofilen zur Anwendung in den Ingenieurwissenschaften mit dem expliziten Fokus diese Hürde zu senken.

Eine entsprechende Verbreitung verbunden mit Bemühungen aus den Fachcommunities trägt so zu einer Standardisierung und Harmonisierung bei. Als Beispiel dafür wird die entwickelte Software im Kontext der Metadatendienste des Konsortiums NFDI4Ing (Schwarz und Anthofer 2023) zukünftig als communityweite Dienstleistung angeboten. Bei ausreichender Integration erlauben die Analyse von angewandten Terminologien, Modellierungskonzepten und Akzeptanzverteilung in Zukunft damit noch bessere Anpassungen der Dienste an die Communities und Aussagen über die entscheidenden Kristallisationspunkte einer Standardisierung, die es zu fördern gilt. Die Datenvalidierung ist dabei nur eine mögliche Anwendung, die jedoch großes Potential zu Integration in verschiedene Strukturen im Rahmen der Datenqualitätsanalyse und institutionellen Speicherung bietet. Der Mehrwert reichhaltiger und maschinennutzbarer Metadaten liegt dabei nicht nur bei Einrichtungen wie den Bibliotheken, sondern auch den Forschenden in der Praxis selbst. Niederschwellig gestaltbare Metadatenapplikationsprofile und die zugehörige Infrastrukturlandschaft bilden dabei einen wichtiges Hilfsmittel.

Danksagung

Gefördert durch die Deutsche Forschungsgemeinschaft (DFG) -Projektnummer 432233186. Die Universitäts- und Landesbibliothek Darmstad (ULB), das IT Center der RWTH Aachen University (ITC), der Lehrstuhl für Fluidsystemtechnik (FST), sowie das Werkzeugmaschinenlabor der RWTH Aachen University (WZL) sind Partner in der Umsetzung dieses Projekts.

ORCID:

- Matthias Grönewald https://orcid.org/0000-0002-3480-9102
- Nils Preuß https://orcid.org/0000-0002-6793-8533

Literaturverzeichnis

Albertoni, Riccardo, David Browning, Simon Cox, Alejandra Gonzalez Beltran, Andrea Perego und Peter Winstanley. 2023. „Data Catalog Vocabulary (DCAT) – Version 3“. Besucht am 12. Mai. https://www.w3.org/TR/vocab-dcat-3/.

Coyle, Karen. 2017. „Application Profiles“. In *Advances in Web Technologies and Engineering,* 1–15. IGI Global. DOI: https://doi.org/10.4018/978-1-5225-2221-8.ch001. https://doi.org/10.4018/978-1-5225-2221-8.ch001.

DataCite Metadata Working Group. 2021. „DataCite Metadata Schema Documentation for the Publication and Citation of Research Data and Other Research Outputs v4.4“. DOI: https://doi.org/10.14454/3W3Z-SA82. https://schema.datacite.org/meta/kernel-4.4/.

Deutsche Forschungsgemeinschaft e.V. 2019. *Leitlinien zur Sicherung guter wissenschaftlicher Praxis.* Zenodo. DOI: https://doi.org/10.5281/zenodo.6472827.

Furner, Jonathan. 2019. „Definitions of 'Metadata': A Brief Survey of International Standards“. *Journal of the Association for Information Science and Technology* 71 (6). DOI: https://doi.org/10.1002/asi.24295. https://doi.org/10.1002/asi.24295.

Grönewald, Matthias, Marc Fuhrmans, Nils Preuss, Benedikt Heinrichs, Sousan Homaeipour und Matthias Bodenbenner. 2023a. „Dataset Structured Data | Google Search Central | Documentation | Google Developers“. Besucht am 12. Mai. https://developers.google.com/search/docs/appearance/structured-data/dataset.

———. 2023b. „DFG-Projekt AIMS - Website“. Besucht am 12. Mai. https://www.aims-projekt.de/.

Knublauch, Holger, und Dimitris Kontokostas. 2023. „Shapes Constraint Language (SHACL)“. Besucht am 12. Mai. https://www.w3.org/TR/shacl/.

Schreiber, Guus, und Yves Raimond. 2023. „RDF 1.1 Primer“. Besucht am 12. Mai. https://www.w3.org/TR/rdf11-primer/.

Schwarz, Annett, und Verena Anthofer. 2023. „NFDI4Ing - Website“. Besucht am 12. Mai. https://nfdi4ing.de/.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersbergand, Gabrielle Appleton, Myles Axtonand, Arie Baakand, Niklas Blombergand u. a. 2016. „The FAIR Guiding Principles for scientific data management and stewardship“. *Scientific data* 3 (1): 1–9. DOI: https://doi.org/10.1038/sdata.2016.18.