FAIR-orientiertes Forschungsdatenmanagement auf Basis strukturierter Primärdaten und semantischer Metadaten

Ein methodischer Beitrag am Beispiel einer Radialventilatordatenbank

Zur Erlangung des akademischen Grades eines

DOKTORS DER INGENIEURWISSENSCHAFTEN (Dr.-Ing.)

von der KIT-Fakultät für Maschinenbau des
Karlsruher Instituts für Technologie (KIT)
angenommene

DISSERTATION

von

M. Sc. Matthias Probst
aus Koblenz

Tag der mündlichen Prüfung:	03.07.2026
Hauptreferent:	Prof. Dr.-Ing. Hans-Jörg Bauer
Korreferent:	Prof. Dr.-Ing. Peter F. Pelz

Vorwort des Herausgebers

Der schnelle technische Fortschritt im Turbomaschinenbau, der durch extreme technische Forderungen und starken internationalen Wettbewerb geprägt ist, verlangt einen effizienten Austausch und die Diskussion von Fachwissen und Erfahrung zwischen Universitäten und industriellen Partnern. Mit der vorliegenden Reihe haben wir versucht, ein Forum zu schaffen, das neben unseren Publikationen in Fachzeitschriften die aktuellen Forschungsergebnisse des Instituts für Thermische Strömungsmaschinen am Karlsruher Institut für Technologie (KIT) einem möglichst großen Kreis von Fachkollegen aus der Wissenschaft und vor allem auch der Praxis zugänglich macht und den Wissenstransfer intensiviert und beschleunigt.

Flugtriebwerke, stationäre Gasturbinen, Turbolader und Verdichter sind im Verbund mit den zugehörigen Anlagen faszinierende Anwendungsbereiche. Es ist nur natürlich, dass die methodischen Lösungsansätze, die neuen Messtechniken, die Laboranlagen auch zur Lösung von Problemstellungen in anderen Gebieten - hier denke ich an Otto- und Dieselmotoren, elektrische Antriebe und zahlreiche weitere Anwendungen - genutzt werden. Die effiziente, umweltfreundliche und zuverlässige Umsetzung von Energie führt zu Fragen der ein- und mehrphasigen Strömung, der Verbrennung und der Schadstoffbildung, des Wärmeübergangs sowie des Verhaltens metallischer und keramischer Materialien und Verbundwerkstoffe. Sie stehen im Mittelpunkt ausgedehnter theoretischer und experimenteller Arbeiten, die im Rahmen nationaler und internationaler Forschungsprogramme in Kooperation mit Partnern aus Industrie, Universitäten und anderen Forschungseinrichtungen durchgeführt werden.

Es sollte nicht unerwähnt bleiben, dass alle Arbeiten durch enge Kooperation innerhalb des Instituts geprägt sind. Nicht ohne Grund ist der Beitrag der Werkstätten, der Technik-, der Rechner- und Verwaltungsabteilungen besonders hervorzuheben. Diplomanden und Hilfsassistenten tragen mit ihren Ideen Wesentliches bei, und natürlich ist es der stets freundschaftlich fordernde wissenschaftliche Austausch zwischen den Forschergruppen des Instituts, der zur gleichbleibend hohen Qualität der Arbeiten entscheidend beiträgt. Dabei sind wir für die Unterstützung unserer Förderer außerordentlich dankbar.

Im vorliegenden Band der Schriftenreihe befasst sich der Autor mit dem Thema des Forschungsdatenmanagements (FDM). Das Forschungsdatenmanagement gewinnt durch die zunehmende Datenmenge und Digitalisierung an Bedeutung, um Datenqualität, Nachvollziehbarkeit und Wiederverwendbarkeit zu sichern. Ein effektives FDM ist entscheidend für wissenschaftlichen Erfolg und Wettbewerbsfähigkeit. Herausforderungen bestehen in der Heterogenität von Systemen, Formaten und interdisziplinärer Zusammenarbeit. Andererseits führt unzureichendes Management zu eingeschränkter Wiederverwendbarkeit, Redundanz und hohen Kosten. Ein zunehmender Kulturwandel, der durch Initiativen wie NFDI (Nationale Forschungsdateninfrastruktur) und europäische Open Science Cloud zu beobachten ist, fördert ein systematisches Datenmanagement. Der Autor beschreibt ein umfassendes, FAIR- (Findable, Accessible, Interoperable, Reusable) orientiertes Forschungsdaten-Management für ingenieurwissenschaftliche Studien, das hochperformante Binärdateien im HDF5-Format mit maschinenlesbaren, semantischen Metadaten (RDF, Ontologien) verknüpft. Anhand einer im Rahmen seiner wissenschaftlichen Untersuchungen am Institut für Thermische Strömungsmaschinen vom Autor erzeugten

Radialventilatordatenbank wird das Konzept implementiert, validiert und über SPARQL Abfragen nutzbar gemacht.

Karlsruhe, im Juli 2026 Prof. Dr.-Ing. Hans-Jörg Bauer

Vorwort des Autors

Die vorliegende Arbeit entstand während meiner Tätigkeit als wissenschaftlicher Mitarbeiter am Institut für Thermische Strömungsmaschinen (ITS) des Karlsruher Instituts für Technologie (KIT).

Mein besonderer Dank gilt Herrn Prof. Dr.-Ing. Hans-Jörg Bauer für die Übernahme des Hauptreferats sowie für das mir entgegengebrachte Vertrauen, dieses für das Institut außergewöhnliche Thema bearbeiten zu dürfen. Herrn Prof. Dr.-Ing. Peter Pelz danke ich herzlich für die Übernahme des Korreferats.

Herrn Dr.-Ing. Balázs Pritz verdanke ich meine Faszination für die Strömungsmechanik. Die vertrauensvolle, unterstützende und inspirierende Zusammenarbeit mit ihm war für mich fachlich wie persönlich von großem Wert und hat meine Promotionszeit in besonderer Weise geprägt.

Mit großer Dankbarkeit denke ich an Herrn Prof. Dr.-Ing. Martin Gabi und Herrn Prof. Dr.-Ing. Robert Stieglitz, die den Beginn meiner Promotion begleitet haben. Ihre Unterstützung und die konstruktiven Gespräche mit ihnen werde ich in dankbarer Erinnerung behalten.

Den Kolleginnen und Kollegen des Instituts danke ich für die fachlichen Diskussionen, die herzliche Zusammenarbeit und die gemeinsame Zeit auch außerhalb des Forschungsalltags. Mein herzlicher Dank gilt außerdem den zahlreichen Studierenden, die an meiner Forschung beteiligt waren, sowie den Mitarbeiterinnen und Mitarbeitern der Werkstätten und des Sekretariats für ihre wertvolle Unterstützung.

Für ihre große, unermüdliche und liebevolle Unterstützung danke ich von Herzen meinen Eltern, meinem Bruder und insbesondere meiner Frau Petra.

Denzlingen, im Juli 2026 Matthias Probst

Inhaltsverzeichnis

Abkürzungen

Abkürzung	Beschreibung
ADF	Allotrope Data Format
AF-HDF	HDF Ontologie der Allotrope Foundation
API	Application Programming Interface
ASCII	American Standard Code for Information Interchange
ASDF	Advanced Scientific Data Format
CF	Climate and Forecast
CFD	Computational Fluid Dynamics
CGNS	CFD General Notation System
CSD	Core Scientific Dataset Model
DCAT	Data Catalog Vocabulary
DCMI	Dublin Core Metadata Initiative
DMP	Datenmanagementplans
FDM	Forschungsdatenmanagement
FITS	Flexible Image Transport System
FOAF	Friend of a Friend (Ontologie)
HDF5	Hierarchical Data Format Version 5
IRI	Internationalized Resource Identifier
JSON	JavaScript Object Notation
JSON-LD	JavaScript Object Notation für Linked Data
LD	Linked Data
LES	Large Eddy Simulation
LLM	Large Language Model
M4I	Metadat4ing (Ontologie)
NeXus	Data Format for Neutron, X-ray and Muon Science
netCDF	Network Common Data Form
NFDI	Nationale Forschungsdateninfrastruktur
NFDI4ing	Nationale Forschungsdateninfrastruktur für Ingenieure
ORCID	Open Researcher Contributor ID
OWL	Web Ontology Language
PID	Persistent Identifier
PIV	Particle Image Velocimetry
PIVMeta	PIV Metadata (Ontologie)
PROV-O	Provenance Ontology

PTV	Particle Tracking Velocimetry
RAG	Retrieval-Augmented Generation
RDF	Resource Description Framework
RDFS	Resource Description Framework SCHEMA
RIF	Rule Interchange Format
ROR	Research Organisation Registry
SHACL	Shapes Constraint Language
SKOS	Simple Knowledge Organization System
SNR	Signal-to-Noise-Ratio
SOSA	Sensor Obervation Sample and Actuator (Ontologie)
SPARQL	SPARQL Protocol and RDF Query Language
SSNO	Simple Standard Name Ontology
SSN	Semantic Sensor Network Ontology
TTL	Terse RDF Triple Language (Turtle)
URI	Uniform Resource Identifier
URN	Uniform Resource Name
W3C	World Wide Web Consortium
WWW	World Wide Web
XML	eXtensible Markup Language
XSD	XML Schema Definition

1 Einleitung

Die fortschreitende Digitalisierung und Automatisierung wissenschaftlicher Arbeitsprozesse führt in den Ingenieurwissenschaften zu einer stetig wachsenden Menge komplexer experimenteller und numerischer Forschungsdaten. Insbesondere in datenintensiven Disziplinen wie der Strömungsmechanik entstehen hochdimensionale Datensätze, deren wissenschaftlicher Wert maßgeblich von ihrer langfristigen Nachvollziehbarkeit, Vergleichbarkeit und Wiederverwendbarkeit abhängt. Voraussetzung hierfür ist ein systematisches Forschungsdatenmanagement (FDM), das den gesamten Lebenszyklus der Daten von der Erzeugung über die Analyse bis hin zur Archivierung und Nachnutzung strukturiert unterstützt.

Eine reine Speicherung oder Veröffentlichung von Datensätzen reicht hierfür nicht. Erst durch die konsistente Ergänzung der Primärdaten um Metadaten, wie etwa zu Randbedingungen, Mess- und Simulationsmethoden, eingesetzten Werkzeugen sowie zur Datenprovenienz, wird eine belastbare Interpretation durch Dritte möglich. Neben der menschlichen Nachvollziehbarkeit gewinnt dabei auch die maschinelle Interpretierbarkeit von Daten an Bedeutung, etwa für automatisierte Analysen, Vergleichsstudien oder die Kopplung mit datengetriebenen Methoden. Nachhaltiges Forschungsdatenmanagement ist damit nicht nur eine organisatorische Aufgabe, sondern eine zentrale methodische Voraussetzung moderner Ingenieurwissenschaften.

Diese Notwendigkeit wird zunehmend auch auf institutioneller Ebene anerkannt. Die Leitlinien zur Sicherung guter wissenschaftlicher Praxis der Deutsche Forschungsgemeinschaft (DFG) (Deutsche Forschungsgemeinschaft, 2015) fordern explizit einen verantwortungsvollen, strukturierten und langfristig angelegten Umgang mit Forschungsdaten. Forschungsdaten werden darin als eigenständige wissenschaftliche Ergebnisse verstanden, deren Qualität, Nachvollziehbarkeit und Wiederverwendbarkeit zu sichern sind. Parallel dazu wurde mit der Nationalen Forschungsdateninfrastruktur (NFDI) (Hartl et al., 2021) ein langfristig angelegter institutioneller Rahmen geschaffen, der die nachhaltige Bereitstellung, Vernetzung und Nachnutzung von Forschungsdaten über Disziplin- und Institutionsgrenzen hinweg ermöglichen soll.

Zentrales Leitbild dieser Entwicklungen sind die FAIR-Prinzipien (Findable, Accessible, Interoperable, Reusable) (Wilkinson et al., 2016), die sich als normativer Referenzrahmen für modernes Forschungsdatenmanagement etabliert haben. Sie bilden sowohl die konzeptionelle Grundlage der NFDI als auch einen Maßstab für die Bewertung guter wissenschaftlicher Praxis im Umgang mit Daten. In den Ingenieurwissenschaften zeigt sich jedoch, dass die praktische Umsetzung der FAIR-Prinzipien häufig fragmentarisch bleibt. Während einzelne Aspekte wie Auffindbarkeit oder Zugänglichkeit adressiert werden, fehlen vielfach integrierte technische Konzepte, die FAIR-Prinzipien systematisch und überprüfbar in datengetriebte wissenschaftliche Arbeitsprozesse überführen.

Ein zentrales Defizit besteht insbesondere in der fehlenden Verbindung zwischen etablierten, leistungsfähigen Dateiformaten zur Speicherung großer wissenschaftlicher Datensätze und formalisierten, semantisch eindeutigen Metadatenmodellen. In der ingenieurwissenschaftlichen Praxis werden strukturierte Binärformate wie HDF5 zwar häufig zur effizienten Ablage experimenteller und numerischer Daten genutzt, die enthaltenen Metadaten sind jedoch meist projekt- oder werkzeugspezifisch, uneinheitlich benannt und semantisch nicht eindeutig beschrieben.

Damit bleibt die Interoperabilität der Daten eingeschränkt, und eine maschinelle Nachnutzung über den ursprünglichen Kontext hinaus ist kaum möglich. Die FAIR-Prinzipien werden so zwar anerkannt, jedoch nicht technisch operationalisiert.

Besonders deutlich treten diese Herausforderungen im Umgang mit Validierungsdaten zutage. Validierungsdatensätze verknüpfen experimentelle Messungen mit numerischen Simulationen und bilden eine zentrale Grundlage für die Bewertung, Kalibrierung und Weiterentwicklung physikalischer Modelle. Ihre wissenschaftliche Relevanz geht dabei über einzelne Projekte hinaus, da sie als Referenz für Vergleichsstudien, Methodenbewertungen und regulatorische Nachweise dienen. Gleichzeitig sind Validierungsdaten in hohem Maße erklärungsbedürftig: Ohne eine präzise Beschreibung von Versuchsaufbau, Randbedingungen, Modellannahmen und Auswertemethoden verlieren sie schnell ihre Aussagekraft.

Für Ventilatoren und vergleichbare Strömungsmaschinen kommt hinzu, dass Validierungsdaten auch eine regulatorische Bedeutung besitzen. Die EU-Richtlinie 2009/125/EG (Ökodesign-Richtlinie) (Europäische Kommission, 2009) fordert für „energieverbrauchsrelevante Produkte“ eine transparente Bewertung der Energieeffizienz auf Basis nachvollziehbarer, reproduzierbarer Mess- und Berechnungsverfahren. Dies setzt die Verfügbarkeit qualitativ hochwertiger, vergleichbarer Validierungsdaten voraus, die nicht nur technisch korrekt, sondern auch langfristig überprüfbar dokumentiert sind. Klassische Datenablage- oder Publikationsformate stoßen hier an ihre Grenzen.

Vor diesem Hintergrund adressiert die vorliegende Arbeit eine zentrale Forschungslücke: Trotz klarer wissenschaftspolitischer Vorgaben durch Institutionen wie DFG und NFDI fehlt ein methodisch belastbares, technisch integriertes Forschungsdatenmanagementkonzept, das die FAIR-Prinzipien für komplexe ingenieurwissenschaftliche Daten nicht nur normativ fordert, sondern konkret, überprüfbar und praxisnah umsetzt. Insbesondere mangelt es an Ansätzen, die nicht primär auf großskalige Infrastrukturen oder langfristig angelegte Verbundprojekte ausgerichtet sind, sondern auch in kleinen und mittleren Forschungsprojekten mit überschaubarem Ressourceneinsatz handhabbar bleiben. Es fehlt somit eine Lösung, die strukturierte Primärdaten, semantisch eindeutige Metadaten und maschinelle Validierbarkeit in einem konsistenten Gesamtkonzept zusammenführt und zugleich im alltäglichen ingenieurwissenschaftlichen Forschungsbetrieb einsetzbar ist.

Zur Adressierung dieser Forschungslücke wird in der vorliegenden Arbeit ein generischer Forschungsdatenmanagementansatz entwickelt, der strukturierte wissenschaftliche Daten auf Basis des hierarchischen Dateiformats HDF5 speichert und diese systematisch mit semantischen Metadaten auf Basis von RDF verknüpft. Durch den Einsatz von Ontologien, standardisierten Vokabularen und formalen Validierungsmechanismen wird eine maschinenlesbare, interoperable und überprüfbare Beschreibung von Daten, Prozessen und Parametern ermöglicht. Die FAIR-Prinzipien werden damit von abstrakten Leitlinien in konkrete technische Anforderungen überführt.

Die entwickelte Methodik wird exemplarisch anhand einer Validierungsdatenbank für einen generischen Radialventilator umgesetzt. Dieses Anwendungsbeispiel dient als konkreter Demonstrator für die Anforderungen an nachhaltiges Forschungsdatenmanagement in der Strö-

mungsmechanik. An ihm lässt sich zeigen, wie experimentelle und numerische Daten über Disziplingrenzen hinweg konsistent beschrieben, miteinander verknüpft und langfristig nachnutzbar bereitgestellt werden können.

Abbildung 1.1 fasst den konzeptionellen Ansatz der Arbeit zusammen. Die Validierungsdatenbank bildet das integrative Bindeglied zwischen Simulation, Experiment und Validierung, welches symbolisch für typische ingenieurstechnischen Fragestellungen sowie das konkrete Praxisbeispiel steht. Datenmanagement, FAIR-konforme Standards, Best Practices und geeignete Werkzeuge stellen den methodischen Rahmen dar, innerhalb dessen die Erzeugung, Beschreibung und Nachnutzung der Daten systematisch abgesichert wird. Die Abbildung fungiert damit als konzeptionelle Klammer für den in dieser Arbeit verfolgten Ansatz eines nachhaltigen, semantisch angereicherten Forschungsdatenmanagements.

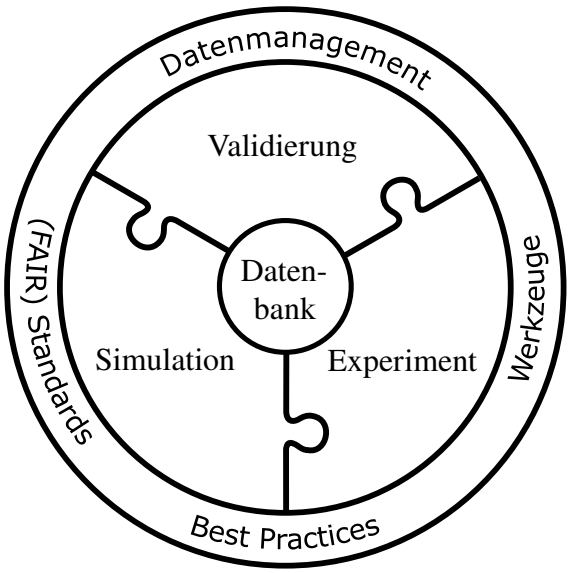


Abbildung 1.1: Konzeptioneller Rahmen der Arbeit zur Realisierung einer FAIR-konformen Validierungsdatenbank. Die Datenbank fungiert als integratives Bindeglied zwischen Simulation, Experiment und Validierung. Datenmanagement, FAIR-Standards, Best Practices und Werkzeuge bilden den methodischen Rahmen, innerhalb dessen die Erzeugung, Beschreibung und Nachnutzung der Daten systematisch abgesichert werden.

1.1 Ziel und Aufbau der Arbeit

Ziel der vorliegenden Arbeit ist die Entwicklung eines methodisch belastbaren und technisch integrierten Ansatzes für ein FAIR-orientiertes Forschungsdatenmanagement auf Basis strukturierter Primärdaten und semantischer Metadaten für komplexe ingenieurwissenschaftliche Forschungsdaten. Der Ansatz zielt darauf ab, die strukturierte und performante Speicherung großer experimenteller und numerischer Primärdaten mit einer formalen, semantisch eindeutigen und maschinenlesbaren Beschreibung von Metadaten zu verbinden, um Auffindbarkeit, Interoperabilität, Nachvollziehbarkeit und Wiederverwendbarkeit der Daten systematisch über den gesamten wissenschaftlichen Datenlebenszyklus hinweg zu gewährleisten.

Ein wesentliches Ziel der Arbeit besteht darin, auf Basis dieses methodischen Ansatzes eine softwaregestützte Lösung zu realisieren, die dessen praktische Anwendung in ingenieurwissenschaftlichen Forschungsprozessen ermöglicht. Die Software bildet die konkrete Umsetzung der entwickelten Methodik ab und dient der konsistenten Erzeugung, Strukturierung, Beschreibung und Validierung FAIR-konformer Forschungsdaten. Sie ist so konzipiert, dass sie von Dritten als Grundlage für eigene Anwendungen, Erweiterungen und Anpassungen genutzt werden kann und damit zur nachhaltigen Etablierung des vorgestellten Ansatzes in der wissenschaftlichen Praxis beiträgt.

Die softwaregestützte Implementierung stellt sicher, dass die aus den FAIR-Prinzipien resultierenden Anforderungen nicht auf einer konzeptionellen Ebene verbleiben, sondern als technische Randbedingungen, standardisierte Schnittstellen und formale Validierungsmechanismen konkret umgesetzt werden. Dadurch wird eine reproduzierbare, überprüfbare und über den spezifischen Anwendungsfall hinaus übertragbare Anwendung der entwickelten Methodik gewährleistet

Im Mittelpunkt der Arbeit steht die Frage, wie für datenintensive Anwendungen eingesetzte Dateiformate systematisch mit Technologien des *Semantic Webs* verknüpft werden können, um semantisch eindeutige und FAIR-konforme Datenstrukturen zu realisieren. Darüber hinaus wird untersucht, wie sich durch maschinenprüfbare, formale Validierungsmechanismen die Qualität, Konsistenz und Nachvollziehbarkeit von Daten und Metadaten systematisch und reproduzierbar absichern lassen.

Die entwickelte Methodik wird exemplarisch anhand einer Validierungsdatenbank für einen generischen Radialventilator aus der Strömungsmechanik demonstriert. Sie dient als repräsentativer Anwendungsfall, um die Anforderungen an Metadatenmodellierung, Datenintegration sowie die langfristige und auch regulatorisch relevante Nachnutzbarkeit ingenieurwissenschaftlicher Validierungsdaten konkret zu untersuchen.

Aus diesen Zielsetzungen ergeben sich folgende zentrale Fragestellungen:

- Wie lassen sich die FAIR-Prinzipien für komplexe, heterogene ingenieurwissenschaftliche Forschungsdaten durch einen softwaregestützten Ansatz technisch, systematisch und überprüfbar operationalisieren?
- Wie können strukturierte Primärdaten in HDF5 mithilfe semantischer Metadaten und Ontologien softwaregestützt so beschrieben und verknüpft werden, dass Interoperabili-

tät und maschinelle Nachnutzbarkeit über den ursprünglichen Erhebungskontext hinaus gewährleistet sind?

- Wie können experimentelle und numerische Validierungsdaten so strukturiert und beschrieben werden, dass sie langfristig vergleichbar, nachvollziehbar und wiederverwendbar bleiben?

Der Aufbau der Arbeit folgt dieser Zielsetzung. Kapitel 2 stellt die Grundlagen des Forschungsdatenmanagements dar und ordnet die Arbeit im Kontext bestehender Konzepte, Standards und verwandter Arbeiten ein. Kapitel 3 analysiert die spezifischen Anforderungen und Herausforderungen des Datenmanagements in der Strömungsmechanik. Kapitel 4 beschreibt das entwickelte generische Forschungsdatenmanagementkonzept auf Basis von HDF5 und RDF sowie die zugrunde liegenden methodischen Prinzipien. Kapitel 5 demonstriert die Anwendung des Konzepts anhand der Validierungsdatenbank für einen generischen Radialventilator. Kapitel 6 bewertet den entwickelten Ansatz mithilfe etablierter FAIR-Reifegradmodelle. Kapitel 7 fasst die zentralen Ergebnisse zusammen und gibt einen Ausblick auf weiterführende Arbeiten.

2 Grundlagen und Stand der Technik von Forschungsdatenmanagement

In diesem Kapitel werden die Grundlagen des Forschungsdatenmanagements (FDM) in den Ingenieurwissenschaften dargelegt. Es vermittelt eine Übersicht über zentrale Anforderungen, aktuelle Entwicklungen sowie relevante technische Konzepte und dient damit als methodische Grundlage für die nachfolgenden Kapitel. Zunächst erfolgt eine Einführung in das Forschungsdatenmanagement und dessen Einordnung in den aktuellen wissenschaftlichen Diskurs. Dabei bilden die sogenannten FAIR-Prinzipien zentrale Leitlinien, die zugleich einen grundlegenden Bezugsrahmen dieser Arbeit darstellen.

Darauf aufbauend werden Ansätze zur Wissensrepräsentation aus dem *Semantic Web* vorgestellt, die für ein technisch integriertes und maschinenlesbares Forschungsdatenmanagement von besonderer Bedeutung sind. Im weiteren Verlauf wird die Thematik im Kontext der Ingenieurwissenschaften verortet und auf das Forschungsgebiet der Strömungsmechanik fokussiert.

2.1 Forschungsdatenmanagement

Ein effektives Forschungsdatenmanagement (FDM) stellt einen wesentlichen Faktor für den wissenschaftlichen Erfolg und die Wettbewerbsfähigkeit dar und etabliert sich zunehmend als eigenständiger Forschungsschwerpunkt in den Ingenieurwissenschaften. Diese Entwicklung ist im Wesentlichen auf zwei Treiber zurückzuführen: Zum einen nimmt die Vernetzung und Kollaboration in Technik und Wissenschaft stetig zu, zum anderen wächst die Menge der im Forschungsprozess anfallenden Daten kontinuierlich. Bereits früh wurde darauf hingewiesen, dass sich das globale Datenvolumen mit hoher Dynamik erhöht und in vielen Bereichen einer nahezu exponentiellen Entwicklung folgt. Verschiedene Beobachtungen und Einschätzungen gehen hier von einer Verdoppelung des Datenvolumens pro Jahr aus (Gray et al., 2005). Die Erschließung dieses Datenpotenzials erfordert jedoch angepasste wissenschaftliche Methoden sowie neue Formen der Datenorganisation und -analyse.

Gray et al. (2005) argumentieren, dass insbesondere die verfügbaren Analysewerkzeuge mit dieser Entwicklung nicht Schritt halten können und identifizieren Defizite an der Schnittstelle zwischen Mensch und Maschine. Sie sehen große Datenzentren in der Verantwortung, Infrastrukturen für datenintensive Wissenschaft bereitzustellen, setzen jedoch zugleich voraus, dass Daten strukturiert, selbsterklärend und maschinenverarbeitbar abgelegt und publiziert werden. Ein zentraler Erfolgsfaktor liegt dabei in der Automatisierung, die Forscher von manuellen und repetitiven Aufgaben entlastet. Vorgeschlagen werden generische Werkzeuge, die eine weitgehend automatisierte Datenexploration und -analyse ermöglichen.

Die Notwendigkeit einer systematischen Auseinandersetzung mit dem Thema Forschungsdatenmanagement ist damit hinreichend begründet. Dennoch besteht in vielen wissenschaftlichen Disziplinen weiterhin ein erheblicher Nachholbedarf hinsichtlich eines nachhaltigen und bewussten Umgangs mit Forschungsdaten. Diese Situation resultiert aus einer Vielzahl von Herausforderungen, die sich unter anderem aus der Heterogenität digitaler und physischer Systeme ergeben,

etwa im Hinblick auf Hardware, Software, Dateiformate sowie experimentelle und numerische Konfigurationen. Hinzu kommt die große Bandbreite wissenschaftlicher Fragestellungen, die häufig interdisziplinäre Zusammenarbeit zwischen Wissenschaft und Industrie erfordert. Weitere Hemmnisse liegen in fehlender Methodenkompetenz und technischer Expertise sowie in einer unzureichenden Sensibilisierung für die langfristige Bedeutung eines konsistenten Datenmanagements (European Commission, 2018). Nicht zuletzt tragen persönliche Präferenzen, Zeitdruck und die hohe Fluktuation des wissenschaftlichen Personals dazu bei, dass Forschungsdatenmanagement im Arbeitsalltag oftmals nachrangig behandelt wird.

Als direkte Folge unzureichender Datenmanagementpraktiken ist die eingeschränkte Wiederverwendbarkeit von Daten und Informationen zu beobachten. Michener et al. (2006) beschreibt diesen Effekt als eine *natürliche Degradierung* des Informations- beziehungsweise Datenwertes für nachnutzende Personen, die mit zunehmendem zeitlichem Abstand zur Datenerhebung einsetzt (vgl. Abbildung 2.1). Fehlende Kontextinformationen, unvollständige Metadaten und eingeschränkte Zugänglichkeit führen dazu, dass Daten an Aussagekraft verlieren oder sogar vollständig unbrauchbar werden. Neben dem Verlust wissenschaftlicher Nachnutzbarkeit begünstigt dies auch redundante Datenerhebungen und vermeidbare Doppelarbeit.

Die Folgen eines unzureichenden Forschungsdatenmanagements sind dabei nicht nur wissenschaftlicher, sondern auch ökonomischer Natur. Die European Commission (2018) schätzt, dass ineffiziente Datenverwaltung jährlich Kosten von mindestens 10,2 Milliarden Euro für die europäische Forschung verursacht, zuzüglich potenzieller weiterer wirtschaftlicher Einbußen von bis zu 16 Milliarden Euro. Durch die konsequente Umsetzung der FAIR-Prinzipien könnten diese Kosten reduziert werden, indem Datenmanagementprozesse effizienter gestaltet, die Qualität von Forschungsergebnissen verbessert und die interdisziplinäre Zusammenarbeit erleichtert werden.

Gleichzeitig ist ein Kulturwandel im Umgang mit Forschungsdaten erkennbar, der mit einer zunehmenden Methodenkompetenz einhergeht. Dieser Wandel wird insbesondere durch nationale und internationale Initiativen und Programme getragen, wie etwa die Nationale Forschungsdateninfrastruktur (NFDI e. V.) (Hartl et al., 2021), die Rahmenbedingungen und Unterstützungsangebote für ein nachhaltiges Forschungsdatenmanagement schafft. Die damit verbundenen Entwicklungen verdeutlichen, dass FDM zunehmend nicht mehr allein als organisatorische Aufgabe verstanden wird, sondern als technisch umsetzbare Voraussetzung für moderne, datengetriebene Wissenschaft.

2.1.1 Hintergründe und Definitionen

In den letzten Jahren hat das Forschungsdatenmanagement deutlich an Aufmerksamkeit gewonnen und wird zunehmend als zentraler Erfolgsfaktor wissenschaftlicher Arbeit erkannt und gezielt gefördert. Ziel dieses Abschnitts ist es, die historische Entwicklung hin zu den heutigen Ausprägungen des Forschungsdatenmanagements in den Ingenieurs- und Naturwissenschaften nachzuzeichnen sowie zentrale Begriffe und Konzepte zu definieren, die für die weitere Argumentation dieser Arbeit relevant sind.

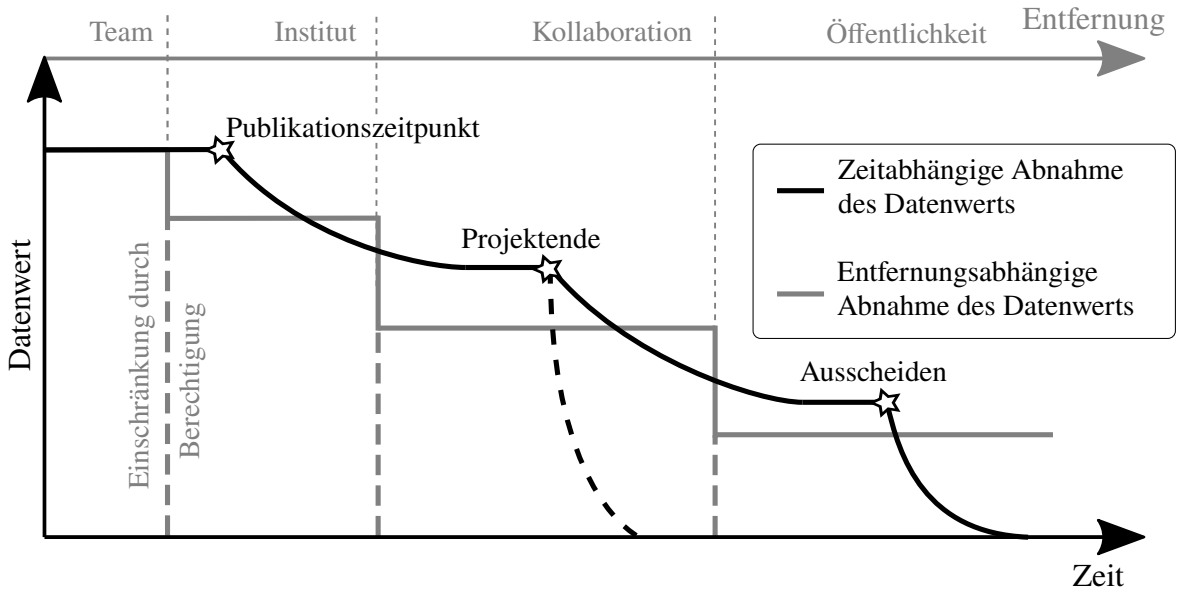


Abbildung 2.1: Abnahme des Informationsgehalts in Anlehnung an Michener et al. (2006). Die Darstellung wurde um den Einfluss der Distanz zur Datenerhebung und -haltung erweitert, welche eine sprunghafte Informationsabnahme infolge verminderter Zugänglichkeit verdeutlicht.

Historische Entwicklung Weder das Thema Datenmanagement noch dessen grundsätzliche Notwendigkeit sind neu. In jüngerer Zeit hat sich jedoch das Bewusstsein für den Umfang, die Bedeutung und die wissenschaftlichen wie gesellschaftlichen Konsequenzen unzureichender Datenpraktiken deutlich geschärft. Zu den wesentlichen Treibern dieser Entwicklung zählen der fortschreitende digitale Wandel, die zunehmende globale Vernetzung und Kooperation in Wissenschaft und Gesellschaft sowie der verstärkte Einsatz datenintensiver Methoden, insbesondere der sogenannten Künstlichen Intelligenz (KI). Letztere wirkt dabei weniger als originärer Auslöser, sondern vielmehr als katalytischer Faktor, der bestehende Defizite in Bezug auf Datenqualität, Dokumentation und Nachnutzbarkeit sichtbar macht und verstärkt.

Verlässlich dokumentierte und langfristig verfügbare Forschungsdaten gelten heute als zentrale Voraussetzung für wissenschaftlichen Fortschritt, Reproduzierbarkeit und nachhaltige Erkenntnisgewinnung (Neuroth, Putnings et al., 2021). Entsprechend betrifft die Forderung nach systematischem Forschungsdatenmanagement längst nicht mehr ausschließlich traditionell datenintensive Disziplinen wie die Klimaforschung oder Hochenergiephysik, sondern erstreckt sich inzwischen auf nahezu alle Bereiche der Wissenschaft.

Bereits im Kodex zur „Sicherung guter wissenschaftlicher Praxis“ aus dem Jahr 1998 wird gefordert, dass „Originaldaten als Grundlage für Veröffentlichungen auf haltbaren und gesicherten Trägern zehn Jahre aufbewahrt werden sollen“ (Deutsche Forschungsgemeinschaft, 2009). Konkrete technische oder organisatorische Handlungsanweisungen werden dort jedoch noch nicht formuliert. In späteren Empfehlungen der Deutschen Forschungsgemeinschaft (Deutsche Forschungsgemeinschaft, 2025) sowie der Allianz der Wissenschaftsorganisationen (Allianz der Wissenschaftsorganisationen, 2010) wird diese Position weiter präzisiert. Neben der Entwick-

lung und Anwendung internationaler sowie fachspezifischer Standards wird insbesondere betont, dass Forschungsdatenmanagement nicht als rein administrative oder technische Zusatzaufgabe zu verstehen ist, sondern als integraler Bestandteil wissenschaftlicher Arbeit, der entsprechend anerkannt und bewertet werden sollte. Ziel ist es, sowohl die Qualität und Nachvollziehbarkeit wissenschaftlicher Ergebnisse als auch die langfristige Produktivität und Wettbewerbsfähigkeit des Wissenschaftssystems zu stärken (Allianz der Wissenschaftsorganisationen, 2010).

Einen wesentlichen Meilenstein stellen die von Wilkinson et al. (2016) formulierten *FAIR (Guiding) Principles* dar. Sie definieren erstmals klar benannte Grundsätze, anhand derer Forschungsdatenmanagementkonzepte systematisch entworfen und bewertet werden können. Die Prinzipien zielen darauf ab, die Auffindbarkeit (**F**indability), Zugänglichkeit (**A**ccessibility), Interoperabilität (**I**nteroperability) sowie die Wiederverwendbarkeit (**R**eusability) digitaler Daten zu verbessern. Eine detaillierte Betrachtung dieser Prinzipien erfolgt in Unterabschnitt 2.1.2.

Die Umsetzung dieser Leitlinien wird seither durch zahlreiche nationale und internationale Initiativen unterstützt. Auf europäischer Ebene ist insbesondere die 2018 initiierte European Open Science Cloud (EOSC) zu nennen, deren Ziel es ist, den Zugang zu datengetriebener Wissenschaft zu erleichtern und interoperable Forschungsinfrastrukturen bereitzustellen (Budroni et al., 2019). In Deutschland wurde auf Initiative des Rats für Informationsinfrastrukturen (RfII) und durch Beschluss der Gemeinsamen Wissenschaftskonferenz (GWK) die Nationale Forschungsdateninfrastruktur (NFDI) ins Leben gerufen (Gemeinsame Wissenschaftskonferenz, 2018). Seit 2020 mit Sitz in Karlsruhe etabliert, soll sie als dauerhaftes Wissens- und Datenökosystem für das deutsche Wissenschaftssystem dienen und die Umsetzung der FAIR-Prinzipien unterstützen (Hartl et al., 2021). Auch die vorliegende Arbeit versteht sich als Beitrag innerhalb dieses wissenschaftspolitischen und infrastrukturellen Rahmens.

Forschungsdatenmanagement Unter Forschungsdatenmanagement wird die systematische Organisation des gesamten Datenlebenszyklus verstanden, von der Erhebung und Speicherung über die Verarbeitung bis hin zur Bereitstellung und Wiederverwendung von Forschungsdaten. Zentrale Zielsetzungen sind dabei die Sicherstellung von Datenqualität, Transparenz, Nachvollziehbarkeit und langfristiger Nachnutzbarkeit. Abhängig von Datenquelle und Anwendung können zusätzlich rechtliche oder ethische Aspekte relevant sein, etwa im Hinblick auf Datenschutz oder Datensicherheit. In den Ingenieurwissenschaften stehen jedoch überwiegend Fragen der Lizenzierung, Zugriffsregelung und technischen Interoperabilität im Vordergrund.

Forschungsdatenzyklus Der Forschungsdatenzyklus, auch als Datenlebenszyklus bezeichnet, beschreibt die zentralen Phasen, die Forschungsdaten von ihrer Entstehung bis zur langfristigen Archivierung und Nachnutzung durchlaufen. Die konkrete Ausprägung einzelner Phasen variiert je nach Disziplin und Anwendung. Abbildung 2.2 fasst fünf wesentliche Schritte zusammen:

1. **Planung und Erhebung**: Daten entstehen durch Messungen, Experimente, Simulationen oder Beobachtungen. Bereits in dieser Phase sind geeignete Datenformate, Metadatenkonzepte und Speicherstrategien festzulegen.

2. **Speicherung und Organisation**: Die strukturierte Ablage der Daten erfolgt in Repositorien, Datenbanken oder spezialisierten Dateiformaten. Eine konsistente Organisation ist Voraussetzung für sicheren Zugriff und langfristige Verfügbarkeit.
3. **Verarbeitung und Analyse**: Daten werden bereinigt, transformiert und analysiert, häufig unter Einsatz spezialisierter Software. Dabei entstehen oftmals weitere Datenprodukte, die ebenfalls dokumentiert werden müssen.
4. **Publikation und Bereitstellung**: Die Veröffentlichung erfolgt über geeignete Plattformen oder Repositorien. Die Einhaltung der FAIR-Prinzipien ist dabei entscheidend für eine effektive Nachnutzung.
5. **Archivierung und Nachnutzung**: Die langfristige Sicherung erfordert standardisierte, interoperable Formate, um eine spätere Wiederverwendung unter veränderten Fragestellungen zu ermöglichen.

Insbesondere in den Ingenieurwissenschaften, etwa bei numerischen Simulationen und experimentellen Untersuchungen, gewinnt ein klar strukturierter Datenlebenszyklus an Bedeutung.

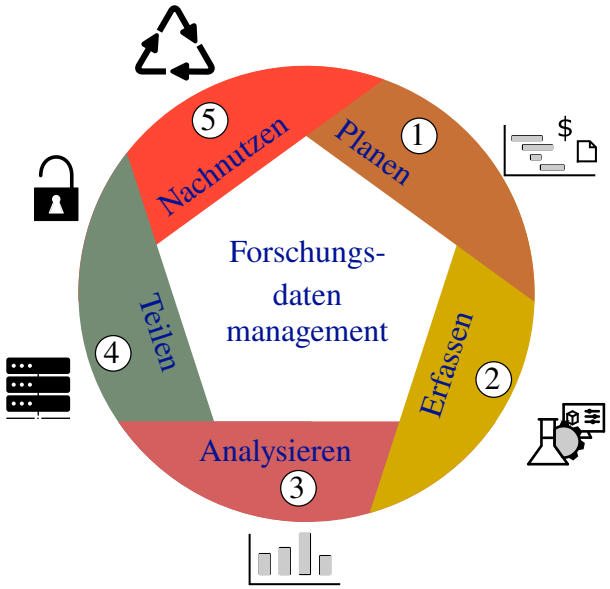


Abbildung 2.2: Illustration des Lebenszyklus von Forschungsdaten. Er beginnt mit der Auswahl eines Dateiformats und eines Metadatenkonzepts (1) und der Durchführung von Qualitätssicherungsmaßnahmen während der Auswahl- und Verarbeitungsphase (2). Im nächsten Schritt werden die Daten für die wissenschaftliche Ausgabe effektiv analysiert (3). Nach der Veröffentlichung sollte die Verfügbarkeit der Daten sichergestellt werden (4). Die (Meta-) Datenqualität definiert sich schließlich durch ihre Auffindbarkeit und damit ihre Wiederverwendbarkeit (5) für spätere Analysen.

Große Datenmengen, komplexe Abhängigkeiten und heterogene Datenformate stellen hohe Anforderungen an Datenqualität, Dokumentation und Standardisierung. Ein konsistent definierter Forschungsdatenzyklus trägt wesentlich zur Effizienzsteigerung, zur interdisziplinären Zusammenarbeit sowie zur Reproduzierbarkeit wissenschaftlicher Ergebnisse bei.

(Forschungs-) Daten und Informationen Der Begriff *Daten* bezeichnet im wissenschaftlichen Kontext häufig quantitative Ergebnisse aus Experimenten, Beobachtungen oder Simulationen. In der Literatur existieren jedoch unterschiedliche Abgrenzungen zwischen Daten und Informationen. So unterscheidet die Informationstheorie nach Shannon (1948) strikt zwischen einer rein syntaktischen Datenrepräsentation und deren semantischer Bedeutung (Voß, 2013). Erst durch die Zuweisung von Bedeutung werden Daten zu Informationen (Floridi, 2005; Kettinger und Li, 2010).

Im wissenschaftlichen Kontext wird daher meist von *Forschungsdaten* gesprochen. Dieser Begriff verdeutlicht sowohl den besonderen Wert dieser Daten als auch die Notwendigkeit eines verantwortungsvollen Umgangs. Forschungsdaten sind zugleich Ergebnis und Grundlage wissenschaftlicher Arbeit und bilden die Basis für Nachvollziehbarkeit, Reproduzierbarkeit und Weiterverwendung von Forschungsergebnissen.

Eine einheitliche Definition des Begriffs *Forschungsdaten* existiert nicht, was vor allem auf die Vielfalt wissenschaftlicher Disziplinen zurückzuführen ist. Entsprechend allgemein ist die Definition der Deutschen Forschungsgemeinschaft formuliert (Deutsche Forschungsgemeinschaft, 2015):

> „Zu Forschungsdaten zählen u. a. Messdaten, Laborwerte, audiovisuelle Informationen, Texte, Surveydaten, Objekte aus Sammlungen oder Proben, die in der wissenschaftlichen Arbeit entstehen, entwickelt oder ausgewertet werden. Methodische Testverfahren, wie Fragebögen, Software und Simulationen können ebenfalls zentrale Ergebnisse wissenschaftlicher Forschung darstellen und sollten daher ebenfalls unter den Begriff Forschungsdaten gefasst werden."

Aus dieser Definition wird deutlich, dass Forschungsdaten heute überwiegend digital vorliegen. Über den Datenbegriff hinaus wird daher häufig der umfassendere Begriff der **digitalen Objekte** verwendet, der neben den Daten selbst auch Software, Hardware, Personen, Prozesse und deren Beziehungen einschließt. Die Verwaltung und Verknüpfung dieser Objekte ist ein zentrales Element moderner Forschungsdatenmanagementkonzepte.

Metadaten Metadaten werden häufig als „Daten über Daten" bezeichnet und liefern den notwendigen Kontext, um Forschungsdaten interpretieren und nachnutzen zu können. Sie wirken der von Michener et al. (2006) beschriebenen „natürlichen Degradierung" des Informationsgehalts entgegen, die mit zunehmendem zeitlichem Abstand zur Datenerhebung einsetzt (vgl. Abbildung 2.1). In der hier verwendeten erweiterten Darstellung wird zusätzlich der Einfluss der Distanz zwischen Datenerzeugenden und Nachnutzenden berücksichtigt, der durch fehlenden direkten Austausch oder eingeschränkten Zugriff zusätzliche Informationsverluste begünstigen kann.

Nach Michener et al. (2006) lassen sich Metadaten nach ihrem Zweck in drei Kategorien einteilen:

1. Auffinden von Daten,
2. Unterstützung von Verständnis und Nutzung durch Menschen,
3. automatisiertes Auffinden, Verarbeiten und Analysieren durch Maschinen.

Die Erhebung strukturierter und konsistenter Metadaten wird von Forschern häufig als aufwendig wahrgenommen, ist jedoch insbesondere in frühen Phasen des Datenlebenszyklus entscheidend für die spätere Nachnutzbarkeit. Eine Schlüsselrolle kommt dabei der Auswahl geeigneter Metadatenstandards zu. Abhängig von Fachgebiet und Anwendung stehen unterschiedliche Konzepte und Werkzeuge zur Verfügung, auf die in den folgenden Kapiteln näher eingegangen wird.

Metadaten lassen sich weiter in verschiedene Kategorien unterteilen (Horsch, Chiacchiera, Cavalcanti et al., 2021):

1. **Technische Metadaten**, die Struktur, Format und Speicherung der Daten beschreiben,
2. **Deskriptive Metadaten**, die der Identifikation und Auffindbarkeit dienen,
3. **Prozessmetadaten**, die Entstehungs- und Verarbeitungsschritte dokumentieren,
4. **Domänenspezifische Metadaten**, die fachliche Inhalte und Kontexte abbilden.

Mit zunehmender fachlicher Spezialisierung sinkt die Verfügbarkeit allgemein akzeptierter Standards, während der manuelle Erfassungsaufwand steigt. Eine besondere Rolle nehmen **semantische Metadaten** ein. Sie stellen eine besondere Ausprägung von Metadaten dar, da sie Daten explizit mit ihrer fachlichen Bedeutung in formalisierter Form verknüpfen, etwa durch den Einsatz kontrollierter Vokabulare oder Ontologien. Nicht alle Metadaten sind per se semantischer Natur. Vielmehr handelt es sich um eine zusätzliche Qualität, die Metadaten durch die Nutzung standardisierter Bedeutungsmodelle erlangen können. Dadurch werden sowohl Daten als auch ihre Metadaten eindeutig beschrieben und maschinenlesbar interpretierbar, was eine automatisierte Verarbeitung, Validierung und Verknüpfung über System- und Disziplingrenzen hinweg ermöglicht.

Die Verwaltung sämtlicher Aspekte entlang des Forschungsdatenzyklus ist eine komplexe Aufgabe, die durch etablierte Verfahren, Standards und geeignete Softwarelösungen unterstützt wird. In vielen geförderten Forschungsvorhaben wird daher die Erstellung eines **Datenmanagementplans** (DMP) gefordert. Ein DMP dokumentiert den Zweck der Datenerhebung sowie die eingesetzten Methoden, Standards und Ressourcen und legt fest, wie Daten gespeichert, geteilt und nach Abschluss eines Projekts für die Nachnutzung bereitgestellt werden sollen. Als überwiegend organisatorisches Instrument schafft der DMP einen wichtigen Rahmen für das Forschungsdatenmanagement, ersetzt jedoch keine technisch integrierten Lösungen zur konsistenten, maschinenlesbaren Umsetzung der beschriebenen Konzepte.

Semantische Metadaten stellen eine besondere Ausprägung von Metadaten dar, da sie Daten explizit mit ihrer fachlichen Bedeutung in formalisierter Form verknüpfen, etwa durch den Einsatz kontrollierter Vokabulare oder Ontologien. Nicht alle Metadaten sind per se semantischer Natur. Vielmehr handelt es sich um eine zusätzliche Qualität, die Metadaten durch die Nutzung

standardisierter Bedeutungsmodelle erlangen können. Dadurch werden sowohl Daten als auch ihre Metadaten eindeutig beschrieben und maschinenlesbar interpretierbar, was eine automatisierte Verarbeitung, Validierung und Verknüpfung über System- und Disziplingrenzen hinweg ermöglicht.

2.1.2 Die FAIR Prinzipien

Das Akronym *FAIR*, eingeführt von Wilkinson et al. (2016), steht für *Findable*, *Accessible*, *Interoperable* und *Reusable* (Auffindbarkeit, Zugänglichkeit, Interoperabilität und Nachnutzbarkeit). Die FAIR-Prinzipien definieren grundlegende Leitlinien für den Umgang mit Daten und Metadaten, mit dem Ziel, deren maschinelle sowie menschliche Auffindbarkeit, Interpretation und Wiederverwendung systematisch zu verbessern. Im Zentrum steht dabei die Steigerung von Transparenz, Reproduzierbarkeit und langfristiger Nachnutzbarkeit wissenschaftlicher Ergebnisse (Nihar et al., 2021).

Hinter den vier Akronymen stehen detaillierte Prinzipien, die heute als Referenzrahmen und Maßstab für modernes und nachhaltiges Forschungsdatenmanagement dienen. Auch die Nationale Forschungsdateninfrastruktur (NFDI) orientiert sich explizit an diesen Leitlinien (Hartl et al., 2021). Darüber hinaus sind die FAIR-Prinzipien fest in Leitlinien zur guten wissenschaftlichen Praxis verankert und finden sich in Förderprogrammen der Europäischen Union, der Deutschen Forschungsgemeinschaft (DFG) sowie zahlreicher Hochschulen wieder (Kailus, 2023). Eine Übersicht der FAIR-Prinzipien und ihrer Konkretisierungen ist in Tabelle 2.1 dargestellt.

Die FAIR-Prinzipien erfahren inzwischen breite Anerkennung über nahezu alle wissenschaftlichen Disziplinen hinweg. Neben Forschungsdaten lassen sie sich auch auf Forschungssoftware anwenden, wie Lamprecht et al. (2020) zeigen. Da viele Datensätze eng mit spezifischer Software verknüpft sind, werden Anforderungen an Software, etwa hinsichtlich Auffindbarkeit, Zugänglichkeit, Lizenzierung und Dokumentation, zur Voraussetzung für die Nachnutzbarkeit der zugehörigen Daten. Mit CodeMeta steht hierfür ein etabliertes Vokabular zur Verfügung, das Software durch standardisierte Metadaten beschreibbar macht (Jones et al., 2023). Die von Gray et al. (2005) erhofften vollständig generischen Analysewerkzeuge existieren bislang jedoch nur in Ansätzen[1], was die Bedeutung strukturierter, maschinenlesbarer Daten- und Metadatenmodelle weiter unterstreicht.

Die konkrete Auslegung der FAIR-Prinzipien ist in der Literatur nicht einheitlich. Diese Offenheit ist bewusst angelegt und spiegelt den Anspruch wider, unterschiedliche Disziplinen, Datenarten und Anwendungsszenarien zu berücksichtigen (FAIR Data Maturity Model Working Group, 2020). In der vorliegenden Arbeit wird die Spezifizierung gemäß Kröger und Wedlich-Zachodin (2020) herangezogen, wie sie in Tabelle 2.1 dargestellt ist. Andere Ansätze wählen teils feinere oder alternative Aufschlüsselungen, was die Vergleichbarkeit erschweren kann, zugleich jedoch den Bedarf an klaren technischen Umsetzungsstrategien verdeutlicht.

[1] Beispielsweise für HDF, netCDF oder FITS.

Unabhängig von der jeweiligen Interpretation lassen sich aus den FAIR-Prinzipien konkrete technische Gestaltungsprinzipien ableiten:

- **Auffindbarkeit**: Verwendung persistenter Identifikatoren (z. B. DOI) sowie standardisierter, indexierbarer Metadatenschemata wie Dublin Core (Dublin Core Metadata Initiative, 2007).
- **Zugänglichkeit**: Nutzung offener und standardisierter Kommunikationsprotokolle (z. B. HTTP) und Sicherstellung der Verfügbarkeit von Metadaten auch bei eingeschränktem Zugriff auf die eigentlichen Daten.
- **Interoperabilität**: Einsatz formaler, maschinenlesbarer Repräsentationen wie dem Resource Description Framework (RDF) (Manola, F. und Miller, E., 2004) sowie kontrollierter Vokabulare und Ontologien, um semantische Verknüpfungen und automatische Integration zu ermöglichen.
- **Nachnutzung**: Bereitstellung eindeutiger Lizenzinformationen, detaillierter Provenienzdaten sowie maschinenlesbarer, standardisierter Metadatenrepräsentationen (z. B. RDF-Serialisierungen wie Turtle) (Beckett und Berners-Lee, 2014), um Wiederverwendbarkeit und Nachvollziehbarkeit sicherzustellen.

Mit dem *FAIR Data Maturity Model* (Bahim et al., 2020; FAIR Data Maturity Model Working Group, 2020) liegt eine weitergehende Spezifizierung der FAIR-Prinzipien vor, die eine systematische und messbare Bewertung der FAIRness erlaubt. Hierzu werden den einzelnen Prinzipien sogenannte Indikatoren zugeordnet, anhand derer der Erfüllungsgrad bestehender oder geplanter Lösungen beurteilt werden kann. So wird beispielsweise das Prinzip F1 in mehrere Indikatoren untergliedert (RDA-F1-01M, RDA-F1-01D, RDA-F1-02M, RDA-F1-02D). Zusätzlich werden Prioritäten und Erfüllungsgrade definiert, die zwischen *nützlich*, *wichtig* und *wesentlich* unterscheiden. Die vollständige Liste ist im Anhang dieser Arbeit dokumentiert (vgl. Abschnitt A.1).

Die FAIR-Prinzipien stellen keinen starren Regelkatalog dar, sondern bewusst allgemein gehaltene Leitlinien (Bahim et al., 2020). Der damit verbundene Interpretationsspielraum führt zu einer Vielzahl unterschiedlicher Werkzeuge, Standards und Methoden, was eine einheitliche Bewertung erschwert (FAIR Data Maturity Model Working Group, 2020). Entgegen der Erwartung einer vollständigen Standardisierung zeigt sich vielmehr eine wachsende Vielfalt an disziplinspezifischen Lösungen. Neben technischen Fragestellungen ergeben sich auch organisatorische Herausforderungen, da die Umsetzung der FAIR-Prinzipien Zeit, Fachwissen und häufig neue Infrastrukturen erfordert. Zudem können etablierte Arbeitsabläufe einer konsequenten Umsetzung entgegenstehen (Safi et al., 2018).

Vor diesem Hintergrund besteht ein klarer Bedarf an unterstützenden Ressourcen, Werkzeugen und Best Practices, die Forscher bei der praktischen Umsetzung der FAIR-Prinzipien begleiten. Eine zentrale Rolle kommt hierbei insbesondere der Nationalen Forschungsdateninfrastruktur (NFDI) zu, die sowohl organisatorische Rahmenbedingungen als auch technische Dienste bereitstellt. Die in Tabelle 2.2 zusammengestellte Auswahl verdeutlicht exemplarisch, dass die Umsetzung der FAIR-Prinzipien nicht durch einzelne Maßnahmen erreicht wird, sondern eine Kombination aus konzeptionellen Rahmenwerken, technischen Werkzeugen und geeigneten Infrastrukturen erfordert.

Prinzip		Beschreibung
Auffindbarkeit	F1	(Meta-) Daten erhalten eine global eindeutige und dauerhafte Kennung.
	F2	Daten werden mit umfangreichen Metadaten beschrieben (siehe R1).
	F3	Metadaten enthalten eindeutig und explizit die Kennung der von ihnen beschriebenen Daten.
	F4	(Meta-) Daten werden in einer durchsuchbaren Ressource registriert oder indiziert.
Zugänglichkeit	A1	(Meta-) Daten können anhand ihrer Kennung unter Verwendung eines standardisierten Kommunikationsprotokolls abgerufen werden.
	A1.1	Das Protokoll ist offen, kostenlos und universell implementierbar.
	A1.2	Das Protokoll ermöglicht bei Bedarf ein Authentifizierungs- und Autorisierungsverfahren.
	A2	Auf Metadaten kann zugegriffen werden, auch wenn die Daten nicht (mehr) verfügbar sind.
Interoperabilität	I1	(Meta-) Daten verwenden eine formale, zugängliche, gemeinsame und allgemein anwendbare Sprache für die Wissensrepräsentation.
	I2	(Meta-) Daten verwenden Vokabulare, die den FAIR-Prinzipien folgen.
	I3	(Meta-) Daten enthalten qualifizierte Verweise auf andere (Meta-) Daten.
Nachnutzbarkeit	R1	(Meta-) Daten werden mit einer Vielzahl genauer und relevanter Attribute ausführlich beschrieben.
	R1.1	(Meta-) Daten werden mit einer eindeutigen und zugänglichen Datennutzungslizenz veröffentlicht.
	R1.2	(Meta-) Daten sind mit detailierten Informationen über die Entstehung versehen.
	R1.3	(Meta-) Daten entsprechen domänrelevanten Community-Standards.

Tabelle 2.1: Die FAIR-Prinzipien nach Wilkinson et al. (2016). Die FAIR Prinzipien im Kontext von Software sind in Lamprecht et al. (2020) aufgelistet.

Zusammenfassend bilden die FAIR-Prinzipien eine unverzichtbare Grundlage für nachhaltiges Forschungsdatenmanagement. Ihre Offenheit ermöglicht die Anpassung an unterschiedliche

wissenschaftliche Disziplinen, macht jedoch zugleich explizite technische Entscheidungen erforderlich, um eine konsistente, überprüfbare und maschinenlesbare Umsetzung zu gewährleisten. Die nachfolgenden Kapitel greifen diese Anforderungen auf und entwickeln darauf aufbauend ein integriertes, FAIR-orientiertes Forschungsdatenmanagementkonzept.

Tabelle 2.2: Auswahl zentraler Ressourcen, Werkzeuge und Infrastrukturen zur Unterstützung eines FAIR-orientierten Forschungsdatenmanagements. Die Tabelle erstreckt sich über mehrere Seiten und gruppiert konzeptionelle Rahmenwerke, technische Werkzeuge, Infrastrukturen und Repositorien.

Name	**Funktion im Kontext FAIR-orientierten FDM**
NFDI4Ing Terminology Service	Zentrales Repositorium für domänenspezifische Ontologien und kontrollierte Vokabulare in den Ingenieurwissenschaften; unterstützt insbesondere Interoperabilität und semantische Eindeutigkeit von Metadaten (Kraft et al., 2023).
FAIRsharing.org	Kuratierte Plattform zur Referenzierung von Datenstandards, Metadatenschemata, Repositorien und Policies; unterstützt die Auswahl FAIR-konformer Standards über Disziplingrenzen hinweg.
FAIR Cookbook	Sammlung praxisnaher Anleitungen zur Umsetzung der FAIR-Prinzipien; primär auf die Lebenswissenschaften ausgerichtet, konzeptionell jedoch übertragbar (Rocca-Serra et al., 2022).
forschungsdaten.info	Zentrale Informationsplattform zum Forschungsdatenmanagement in Deutschland mit Leitfäden, Best Practices und rechtlichen Hinweisen; unterstützt insbesondere organisatorische Aspekte (Kröger und Wedlich-Zachodin, 2020).
Protégé	Open-Source-Software zur Erstellung, Bearbeitung und Validierung von Ontologien; zentrales Werkzeug zur formalen Modellierung semantischer Metadaten (Musen, 2015).
FAIR Data Maturity Model	Bewertungsrahmen zur systematischen und messbaren Einordnung der FAIRness von Daten und Metadaten anhand definierter Indikatoren (FAIR Data Maturity Model Working Group, 2020).

Fortsetzung auf der nächsten Seite

Name	**Funktion im Kontext FAIR-orientierten FDM (Fortsetzung)**
EOSC	Europäische Forschungsinfrastruktur zur Bereitstellung von Diensten, Plattformen und Zugangsmechanismen für datengetriebene Wissenschaft; adressiert insbesondere Auffindbarkeit und Zugänglichkeit (Almeida et al., 2017; Budroni et al., 2019).
GO FAIR Initiative	Internationale Initiative zur Förderung der FAIR-Prinzipien durch Community-getriebene Implementierungsnetzwerke und Referenzkonzepte (Wissel et al., 2020).
Zenodo	Offenes, von der EU unterstütztes Repositorium zur Publikation von Forschungsdaten und Software mit persistenten Identifikatoren; unterstützt Auffindbarkeit, Zitierfähigkeit und langfristige Verfügbarkeit (European Organization For Nuclear Research und OpenAIRE, 2013).
Figshare	Disziplinübergreifendes Repositorium für Forschungsdaten mit Fokus auf Sichtbarkeit und einfache Veröffentlichung; erfüllt grundlegende FAIR-Anforderungen (*figshare* 2023; Singh, 2011).
ing.grid	Open-Access-Journal für FAIR-orientiertes Forschungsdatenmanagement in den Ingenieurwissenschaften; dient der wissenschaftlichen Publikation strukturierter Datensätze (Pimenta et al., 2023).

2.2 Ansätze für Wissensrepräsentation

Dieses Kapitel gibt einen Überblick über grundlegende Ansätze der Wissensrepräsentation und ordnet zentrale Konzepte der semantischen Modellierung ein, die in Wissenschaft und Technik etabliert sind. Sie bilden die methodische Grundlage für die in dieser Arbeit entwickelten Ansätze zu einem nachhaltigen, FAIR-orientierten Forschungsdatenmanagement.

Die Umsetzung der FAIR-Prinzipien erfordert eine strukturierte, konsistente und maschinenlesbare Beschreibung von Daten und Metadaten. Damit rückt die Wissensrepräsentation als methodische Grundlage moderner Forschungsdatenmanagementkonzepte in den Fokus. Sie befasst sich mit der formalen Modellierung von Wissen, um Bedeutung, Beziehungen und Kontextinformationen explizit darzustellen und dadurch Interoperabilität, Nachvollziehbarkeit und Wiederverwendbarkeit zu ermöglichen.

Im Kontext dieser Arbeit wird Wissensrepräsentation nicht als abstraktes Modellierungsproblem verstanden, sondern als technische Voraussetzung für die konsistente, überprüfbare und maschineninterpretierbare Umsetzung der FAIR-Prinzipien. Insbesondere die Anforderungen an Interoperabilität, Wiederverwendbarkeit sowie automatisierte Validierung lassen sich nur durch formale Modelle erfüllen, die über rein syntaktische Metadaten hinausgehen.

In der Informationstechnologie spielen semantische Netze eine zentrale Rolle, da sie Beziehungen und Bedeutungen zwischen Daten explizit abbilden. Zu diesen Ansätzen zählen unter anderem kontrollierte Vokabulare, Thesauri, Ontologien und Wissensgraphen, die jeweils unterschiedliche Zielsetzungen verfolgen und für verschiedene Anwendungsbereiche konzipiert sind. Ontologien wie die *Gene Ontology* (Ashburner et al., 2000) in der Biologie, die *ChEBI Ontology* (Degtyarenko et al., 2007) in der Chemie oder Metadata4Ing (Arndt et al., 2023) in den Ingenieurwissenschaften strukturieren domänenspezifisches Fachwissen. Wissensgraphen wie *DBpedia* (Auer et al., 2007) oder *Wikidata* (Vrandečić und Krötzsch, 2014) verfolgen hingegen eine domänenübergreifende Wissensrepräsentation und dienen insbesondere der Integration heterogener Datenquellen.

Die Verbreitung semantischer Technologien ist disziplinabhängig unterschiedlich ausgeprägt. Während datenintensive Bereiche wie Biologie, Chemie oder Geowissenschaften frühzeitig auf formale Wissensmodelle zurückgegriffen haben, gewinnen entsprechende Ansätze auch in den Ingenieurwissenschaften zunehmend an Bedeutung. Insbesondere vor dem Hintergrund wachsender Datenmengen, komplexer Versuchsaufbauten und numerischer Konfigurationen sowie interdisziplinärer Forschung eröffnen semantische Technologien erhebliche Potenziale zur Effizienzsteigerung, Standardisierung und nachhaltigen Nachnutzung von Forschungsdaten.

Mit der zunehmenden digitalen Datenproduktion, der Vernetzung physischer Systeme im Rahmen des *Internet of Things* (IoT) sowie den Fortschritten in den Bereichen künstliche Intelligenz und maschinelles Lernen vollzieht sich ein grundlegender Wandel in der Informationshaltung des Webs. Während das Internet ursprünglich als *Web of Documents* konzipiert war, beschreibt Berners-Lee et al. (2001) die Vision eines *Web of Data*, in dem Daten nicht nur strukturiert abgelegt, sondern über standardisierte Schnittstellen miteinander verknüpft und mit wohldefinierter Bedeutung versehen sind. Dieses Konzept bildet die Grundlage des *Semantic Web*.

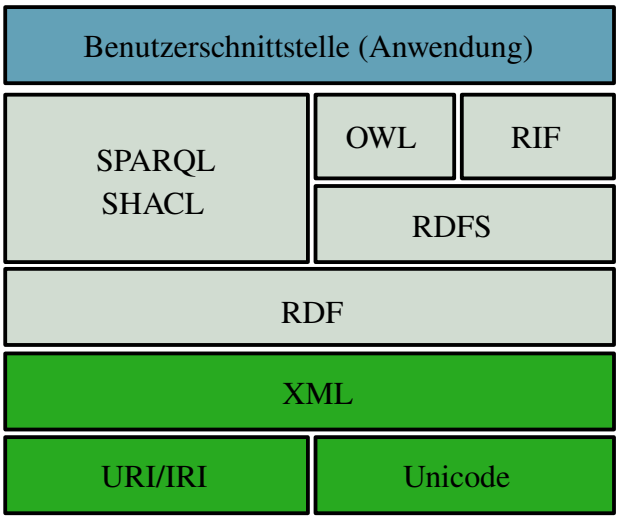


Abbildung 2.3: Ausschnitt des Schichtenmodells des Semantic Webs. Dargestellt sind die für diese Arbeit relevanten Ebenen, die aufeinander aufbauen und eine zunehmend semantische und maschinenlesbare Verarbeitung wissenschaftlicher Daten ermöglichen.

Das Semantic Web basiert auf einer mehrschichtigen Architektur, die eine strukturierte und maschinenlesbare Verarbeitung von Informationen ermöglicht. Das zugrunde liegende Schichtenmodell wurde von Tim Berners-Lee entwickelt und ist nicht als Ablösung, sondern als Erweiterung des klassischen Hypertext-Webs zu verstehen (Berners-Lee et al., 2001). Die Schichten sind dabei nicht statisch, sondern unterliegen einer fortlaufenden Weiterentwicklung. Die in dieser Arbeit verwendete Darstellung (Abbildung 2.3) fokussiert auf die für die Modellierung und Verwaltung wissenschaftlicher Daten relevanten Ebenen.

Die Basis bilden Technologien zur eindeutigen Identifikation von Ressourcen, insbesondere *Uniform Resource Identifier* (URI), ergänzt durch Unicode zur Unterstützung internationaler Zeichensätze. Darauf aufbauend erlaubt die *Extensible Markup Language* (XML) eine flexible Strukturierung von Daten. Das *Resource Description Framework* (RDF) definiert ein standardisiertes Modell zur Beschreibung von Ressourcen und deren Beziehungen in Form von Tripeln. Für die weitergehende semantische Modellierung stellen *RDF Schema* (RDFS) und die *Web Ontology Language* (OWL) formale Vokabulare zur Definition von Klassen, Eigenschaften und logischen Einschränkungen bereit. Ergänzend ermöglicht das *Rule Interchange Format* (RIF) die Spezifikation und den Austausch von Regeln, wird in dieser Arbeit jedoch nicht weiter betrachtet.

Zur Abfrage semantischer Daten dient die *SPARQL Protocol and RDF Query Language* (SPARQL), eine deklarative Abfragesprache für RDF-Datenbestände. Eng damit verbunden ist die *Shapes Constraint Language* (SHACL), eine W3C-Empfehlung zur formalen Spezifikation und Validierung struktureller und semantischer Constraints über RDF-Graphen (Pareti und Konstantinidis, 2022). SHACL dient dabei nicht der ontologischen Modellierung selbst, sondern

der Überprüfung, ob RDF-Daten definierte Struktur-, Typ- und Vollständigkeitsanforderungen erfüllen. In Kombination mit RDF, OWL und SPARQL ermöglichen diese Technologien neben der strukturierten Speicherung und Abfrage von Daten insbesondere die explizite Absicherung von Provenienz- und Prozessinformationen, die für Reproduzierbarkeit und Nachvollziehbarkeit wissenschaftlicher Ergebnisse zentral sind. Mit Ausnahme von RIF kommen alle genannten Technologien im Rahmen dieser Arbeit zum Einsatz und bilden die methodische Grundlage für die entwickelte Standardnamenontologie SSNO (vgl. Abschnitt 4.6) sowie für die Validierungsdatenbank (vgl. Kapitel 5).

Ziel der beschriebenen Ansätze ist die Standardisierung und formale Strukturierung von Wissen. Dabei lassen sich unterschiedliche Formen der Wissensrepräsentation unterscheiden, die eine Hierarchie wachsender Komplexität, Ausdrucksstärke und Funktionalität bilden (Horsch, Chiacchiera, Bami et al., 2020):

1. einfache Listen von Begriffen ohne explizite Struktur,
2. informelle Hierarchien mit grundlegenden Beziehungen,
3. Thesauri mit definierten semantischen Relationen wie Synonymen,
4. Taxonomien mit hierarchischer Klassenstruktur,
5. konzeptuelle Modelle und Schemata (z. B. XSD, RDFS),
6. Ontologien mit formalen Beziehungen, Einschränkungen und Regeln.

Diese Ansätze lassen sich grob in zwei Gruppen einteilen: Kontrollierte Vokabulare, Thesauri und einfache Taxonomien sind primär auf die menschliche Nutzung ausgelegt und unterstützen eine einheitliche Terminologie. Ontologien und konzeptuelle Modelle hingegen zielen auf eine formale, maschinenlesbare und automatisiert auswertbare Wissensrepräsentation ab.

Kontrollierte Vokabulare dienen der Standardisierung von Begriffen innerhalb einer Domäne und reduzieren Mehrdeutigkeiten, etwa durch die Vermeidung von Synonymen oder Homonymen. Sie stellen eine wichtige Grundlage für konsistente Datenbeschreibung dar, stoßen jedoch bei komplexen Abhängigkeiten, logischen Einschränkungen und automatisierter Verarbeitung an ihre Grenzen. Ein bekanntes Beispiel sind die *Climate and Forecast Metadata Conventions* (CF Conventions), die Begriffe aus der Erd- und Klimaforschung standardisieren. Aufgrund inhaltlicher Überschneidungen mit strömungsmechanischen Größen spielen sie auch im Kontext dieser Arbeit eine Rolle und werden in Unterabschnitt 2.3.1 sowie Abschnitt 4.6 vertieft betrachtet.

Ontologien stellen die formalste und ausdrucksstärkste Form der Wissensrepräsentation dar. Sie spezifizieren ein konzeptuelles Modell eines Anwendungsbereichs, das Entitäten, deren Beziehungen sowie logische Einschränkungen explizit beschreibt. Dadurch ermöglichen sie präzise, konsistente und maschineninterpretierbare Beschreibungen komplexer Wissensdomänen. Diese Eigenschaften gehen deutlich über die Möglichkeiten kontrollierter Vokabulare oder Taxonomien hinaus und sind insbesondere für automatisierte Analysen, Validierungen und semantische Integration erforderlich. Auch in den Ingenieurwissenschaften gewinnen Ontologien zunehmend an Bedeutung, wie verschiedene durch die NFDI geförderte Initiativen zeigen, etwa der

NFDI4Ing Terminology Service oder die Ontologie Metadata4Ing. Eine vertiefte Betrachtung von Ontologien erfolgt in Unterabschnitt 2.2.2.

Die Möglichkeit einer expliziten, formalen Wissensbeschreibung ist nicht nur innerhalb einzelner Disziplinen relevant, sondern gewinnt angesichts der stetig wachsenden Daten- und Informationsmengen im World Wide Web weiter an Bedeutung. Die Vielfalt verteilter Datenquellen, Formate und Sprachen stellt eine zentrale Herausforderung für deren Integration und Nachnutzung dar. Insbesondere die Wissenschaft, die auf Austausch, Vergleichbarkeit und Reproduzierbarkeit angewiesen ist, profitiert von standardisierten, semantisch angereicherten Datenmodellen.

Eine Schlüsselrolle bei der Standardisierung und Weiterentwicklung entsprechender Technologien nimmt das World Wide Web Consortium (W3C) ein, das zentrale Standards des Semantic Web entwickelt und pflegt und damit Interoperabilität und langfristige Nutzbarkeit fördert (Hitzler et al., 2008).

Zusammenfassend bilden Ontologien und Semantic-Web-Technologien einen zentralen methodischen Baustein für die strukturierte, interoperable und nachhaltige Beschreibung wissenschaftlicher Daten. Sie schaffen die technische Grundlage für die konsistente Umsetzung der FAIR-Prinzipien und ermöglichen eine über Disziplinen hinweg nutzbare Wissensrepräsentation. Die folgenden Unterkapitel vertiefen diese Konzepte anhand der in dieser Arbeit eingesetzten Technologien. Den Anfang bildet das *Resource Description Framework* (RDF) als grundlegendes Datenmodell für semantische Beschreibungen.

2.2.1 Datenstrukturierung und Verknüpfung mittels RDF

Das *Resource Description Framework* (RDF) bildet einen zentralen Baustein für die strukturierte Beschreibung und semantische Verknüpfung von Daten im *Semantic Web* (Manola, F. und Miller, E., 2004). Es wurde vom World Wide Web Consortium (W3C) entwickelt und stellt ein standardisiertes Datenmodell zur maschinenlesbaren Repräsentation von Informationen bereit. RDF ist dabei nicht als Ersatz klassischer Datenmodelle zu verstehen, sondern als ergänzendes, domänenübergreifend einsetzbares Modell zur Beschreibung von Bedeutung, Beziehungen und Kontextinformationen.

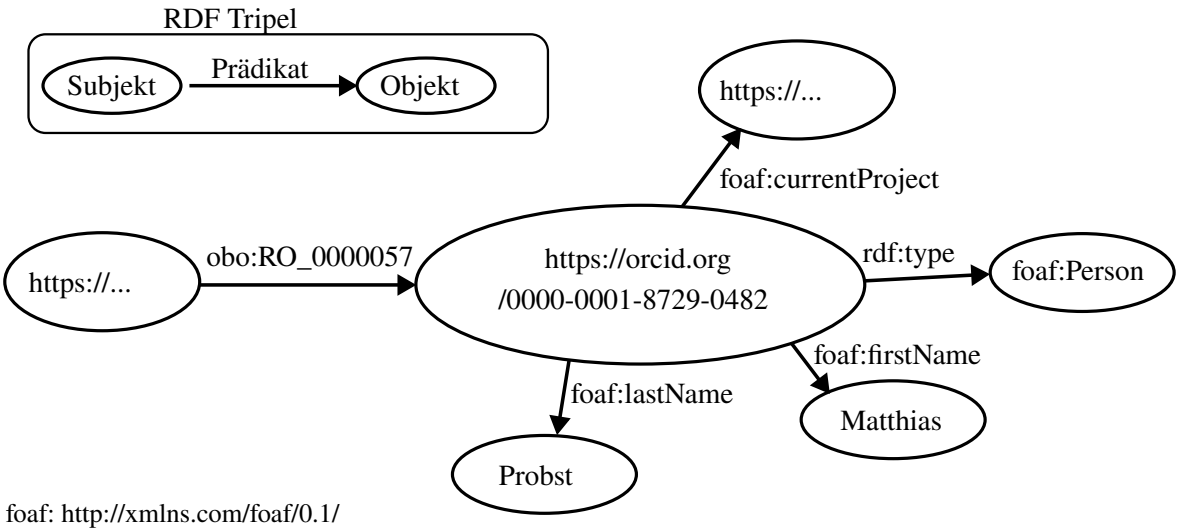
Grundlegendes Konzept von RDF ist die Beschreibung von Aussagen über sogenannte *Ressourcen*. Eine Ressource bezeichnet jedes eindeutig identifizierbare Objekt oder Konzept (*Thing*) (Berners-Lee et al., 2001), das über einen global eindeutigen Bezeichner adressiert wird. In der Regel erfolgt dies über einen *Uniform Resource Identifier* (URI). Da URIs auf ASCII-Zeichen beschränkt sind, wird in der Praxis zunehmend der *Internationalized Resource Identifier* (IRI) verwendet, der auch internationale Schriftzeichen unterstützt. In vielen Anwendungsfällen werden URI und IRI synonym verwendet. Ressourcen können physische Objekte (z. B. Personen, Orte oder Messinstrumente), digitale Artefakte (z. B. Dateien oder Webseiten) oder abstrakte Konzepte (z. B. mathematische Modelle, Ontologien oder wissenschaftliche Methoden) repräsentieren.

Zur formalen Beschreibung von Aussagen verwendet RDF sogenannte *Tripel*. Jedes Tripel besteht aus drei eindeutig definierten Komponenten:

- Das **Subjekt** bezeichnet die Ressource, über die eine Aussage getroffen wird.
- Das **Prädikat** beschreibt die Art der Beziehung oder Eigenschaft, die das Subjekt mit dem Objekt verknüpft.
- Das **Objekt** stellt den Wert der Aussage dar und kann entweder selbst eine Ressource oder ein Literal (z. B. Zahlenwert, Text oder Datum) sein.

Eine Menge solcher Tripel bildet einen RDF-Graphen. In diesem entsprechen Knoten den Ressourcen beziehungsweise Literalen, während gerichtete Kanten die Beziehungen zwischen ihnen darstellen. Diese graphbasierte Struktur erlaubt eine flexible Erweiterung bestehender Datenmodelle, da neue Aussagen durch das Hinzufügen weiterer Tripel ergänzt werden können, ohne vorhandene Strukturen verändern zu müssen. Dadurch eignet sich RDF insbesondere für heterogene und dynamisch wachsende Datenbestände, wie sie im wissenschaftlichen Kontext häufig auftreten.

Abbildung 2.4 zeigt exemplarisch einen Ausschnitt eines RDF-Graphen zur Beschreibung einer Person. Zur besseren Lesbarkeit wird die sogenannte Präfixnotation verwendet, bei der lange IRIs durch kurze, deklarierte Namensraumpräfixe ersetzt werden. So verweist beispielsweise *foaf:Person* auf die Ressource mit der IRI http://xmlns.com/foaf/0.1/Person. Das Präfix „foaf“ steht hierbei für die Ontologie *Friend of a Friend* (FOAF), die häufig zur Beschreibung von Personen und deren Beziehungen eingesetzt wird. Die Verwendung von Präfixen ist ein allgemeines Prinzip in RDF und dient der übersichtlichen und konsistenten Modellierung komplexer Graphstrukturen unabhängig von der konkret verwendeten Ontologie.



foaf: http://xmlns.com/foaf/0.1/
obo: http://purl.obolibrary.org/obo/
rdf: http://www.w3.org/1999/02/22-rdf-syntax-ns#

Abbildung 2.4: Darstellung der Informationen und Beziehungen einer Person im RDF als gerichteter Graph. Jedes Tripel besteht aus Subjekt, Prädikat und Objekt und bildet die grundlegende Struktureinheit des RDF. Der Graph verdeutlicht, wie Daten und Metainformationen explizit verknüpft und semantisch in Beziehung gesetzt werden können.

Serialisierung, Validierung und Durchsuchbarkeit

Die visuelle Darstellung von RDF-Daten als Graphen, wie in Abbildung 2.4 dargestellt, ist lediglich für kleine Datenmengen praktikabel. Mit zunehmender Anzahl von Knoten und Kanten, häufig im Bereich von mehreren hunderttausend oder Millionen Tripeln, verlieren graphische Repräsentationen rasch an Übersichtlichkeit und Handhabbarkeit. Für die praktische Nutzung werden RDF-Daten daher in lineare Zeichenketten überführt, ein Vorgang, der als *Serialisierung* bezeichnet wird (Hitzler et al., 2008). Während RDF das zugrunde liegende Datenmodell definiert, stellen RDF-Serialisierungsformate konkrete syntaktische Repräsentationen dieses Modells dar. Sie sind maschinenlesbar, dateibasiert speicherbar und über Netzwerke übertragbar und bilden damit eine wesentliche Voraussetzung für Interoperabilität im Web.

Zu den am weitesten verbreiteten RDF-Serialisierungsformaten zählen Terse RDF Triple Language (Turtle, Dateiendung *.ttl*) und JSON-LD (Dateiendung *.jsonld*). Beide Formate werden im Rahmen dieser Arbeit verwendet. In den folgenden Kapiteln wird überwiegend das Turtle-Format eingesetzt, da es eine kompakte, gut lesbare Darstellung erlaubt und sich daher besonders für Beispiele, Modellierungsentscheidungen und Diskussionen eignet. Da die im Rahmen dieser Arbeit publizierten Daten grundsätzlich in beiden Formaten bereitgestellt werden und in der Literatur keine eindeutige Präferenz erkennbar ist, werden im Folgenden sowohl Turtle als auch JSON-LD vorgestellt.

JSON (JavaScript Object Notation) ist ein textbasiertes Austauschformat, das insbesondere für die Kommunikation zwischen Webanwendungen entwickelt wurde (Pezoa et al., 2016). Es ist sowohl für Menschen als auch für Maschinen gut lesbar und hat sich disziplinübergreifend als De-facto-Standard etabliert. JSON-LD erweitert dieses Format um Konzepte des Linked Data und ermöglicht damit die Abbildung semantischer Beziehungen innerhalb der vertrauten JSON-Struktur. Dadurch eignet sich JSON-LD besonders für die Integration heterogener Datenquellen sowie für den Einsatz in webbasierten Schnittstellen und Application Programming Interfaces (APIs).

Ein zentrales Element von JSON-LD ist die Kontextdefinition über das Schlüsselwort „@context“. Sie erlaubt die Zuordnung kompakter Schlüssel zu eindeutig referenzierbaren IRIs und stellt sicher, dass Begriffe unabhängig von Sprache oder Benennung konsistent interpretiert werden. Auf diese Weise wird aus einem rein syntaktischen Schlüssel ein semantisch eindeutig definiertes Prädikat, das maschinell ausgewertet und mit externem Wissen verknüpft werden kann.

Listing 2.1 illustriert diesen Unterschied anhand der Beschreibung einer an der Datenerzeugung beteiligten Person. Während die einfache JSON-Darstellung lediglich für Menschen interpretierbar ist, erlaubt JSON-LD durch die explizite Kontextdefinition eine formale, automatisierte Auswertung durch Maschinen.

Zusätzlich zu *@context* spielen in JSON-LD die Schlüsselwörter *@id* und *@type* eine zentrale Rolle. Mit *@id* erhält jede Entität einen global eindeutigen Identifikator, der eine konsistente Referenzierung auch über Datensätze und Systeme hinweg ermöglicht. Im vorliegenden Beispiel erfolgt die Identifikation über die ORCID des Autors (https://orcid.org/0000-0001-8729-0482).

Listing 2.1: JSON-Format eines Datensatzes zur Beschreibung einer Person. Die unterschiedliche Schreibweise und Wahl der Sprache in den Schlüsselworten ist bewusst gewählt, um den Unterschied zu und die Vorteile von JSON-LD zu verdeutlichen.

```
{
  "name": "Probst",
  "Vorname": "Matthias",
  "EMail": "matthias.probst@kit.edu"
}
```

Listing 2.2: Serialisierung des de facto gleichen Datensatzes aus Listing 2.1 im JSON-LD-Format, bei dem die Verwendung von URIs eine eindeutige und explizite Identifizierung der Informationen ermöglicht.

```
{
    "@context": {
        "foaf": "http://xmlns.com/foaf/0.1/"
    },
    "@id": "https://orcid.org/0000-0001-8729-0482",
    "@type": "foaf:Person",
    "foaf:givenName": "Matthias",
    "foaf:familyName": "Probst",
    "foaf:mbox": "matthias.probst@kit.edu"
}
```

Das Attribut *@type* spezifiziert den Typ der Entität und verweist auf eine definierte Klasse innerhalb einer Ontologie, hier *foaf:Person* aus der FOAF-Ontologie.

Das Turtle-Format stellt eine besonders kompakte und explizite Syntax zur Serialisierung von RDF-Daten dar. Im Gegensatz zu JSON-LD, das RDF in der Struktur von JSON-Objekten abbildet, orientiert sich Turtle unmittelbar an der Tripelnotation von RDF und beschreibt Aussagen explizit als Subjekt–Prädikat–Objekt-Ausdrücke. Die vollständige Spezifikation ist in Beckett und Berners-Lee (2014) dokumentiert. Turtle und JSON-LD sind vollständig interoperabel und lassen sich ohne Informationsverlust ineinander überführen; sie unterscheiden sich ausschließlich in der syntaktischen Darstellung. Die entsprechende Turtle-Repräsentation des in Listing 2.1 dargestellten Beispiels ist in Listing 2.3 gezeigt.

Neben der Serialisierung spielt die Sicherstellung von Datenqualität und Konsistenz eine zentrale Rolle. Hierzu dient die *Shapes Constraint Language* (SHACL), eine vom W3C standardisierte Sprache zur Beschreibung und Validierung von RDF-Daten (Knublauch und Kontokostas, 2017). SHACL erlaubt die formale Definition sogenannter *Shapes*, mit denen strukturelle Anforderungen an RDF-Ressourcen spezifiziert werden können, etwa hinsichtlich erlaubter Eigenschaften, Datentypen oder Kardinalitäten.

Ein wesentliches Merkmal von SHACL ist die klare Trennung zwischen Daten- und Validie-

Listing 2.3: RDF-Serialisierung im Turtle-Format.

```
@prefix foaf: <http://xmlns.com/foaf/0.1/> .

<https://orcid.org/0000-0001-8729-0482> a foaf:Person ;
    foaf:givenName "Matthias" ;
    foaf:familyName "Probst" ;
    foaf:mbox "matthias.probst@kit.edu" .
```

rungsebene: Während RDF inhaltliche Aussagen modelliert, definieren SHACL-Shapes formale Regeln zur Überprüfung dieser Aussagen. So kann beispielsweise festgelegt werden, dass jede Instanz der Klasse *foaf:Person* über mindestens einen Vor- und Nachnamen verfügen muss (vgl. Listing 2.4). Auf diese Weise lassen sich RDF-Daten automatisiert validieren und unvollständige oder fehlerhafte Instanzen identifizieren.

Listing 2.4: Beispiel eines SHACL-Shapes zur Validierung einer Person (*foaf:Person*).

```
@prefix sh:   <http://www.w3.org/ns/shacl#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix xsd:  <http://www.w3.org/2001/XMLSchema#> .
@prefix rdf:  <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix ex:   <http://example.org/shapes#> .

ex:PersonShape a sh:NodeShape ;
  sh:targetClass foaf:Person ;
  sh:property [
    sh:path foaf:givenName ;
    sh:nodeKind sh:Literal ;
    sh:minCount 1 ;
    sh:minLength 1 ;
    sh:pattern ".*\\S.*" ;
    sh:or ( [ sh:datatype xsd:string ] [ sh:datatype rdf:langString ] ) ;
  ] ;
  sh:property [
    sh:path foaf:familyName ;
    sh:nodeKind sh:Literal ;
    sh:minCount 1 ;
    sh:minLength 1 ;
    sh:pattern ".*\\S.*" ;
    sh:or ( [ sh:datatype xsd:string ] [ sh:datatype rdf:langString ] ) ;
  ] .
```

Da SHACL unabhängig vom verwendeten Serialisierungsformat arbeitet, können sowohl in Turtle als auch in JSON-LD vorliegende RDF-Daten validiert werden. In der Praxis werden SHACL-Shapes häufig in Turtle formuliert, da sich komplexe Regelstrukturen dort übersichtlich darstellen lassen, während die zu prüfenden Daten in unterschiedlichen Formaten vorliegen

können. SHACL stellt damit ein zentrales Werkzeug zur technischen Operationalisierung von Datenqualität im Rahmen FAIR-orientierter Datenmanagementkonzepte dar.

Nach der Serialisierung und Validierung stellt sich die Frage der effizienten Durchsuchbarkeit und Abfrage semantischer Daten. Hierfür dient die *SPARQL Protocol and RDF Query Language* (SPARQL) (Aranda et al., 2013), die als standardisierte Abfragesprache für RDF-Daten etabliert ist. SPARQL erlaubt die Formulierung komplexer Anfragen über graphbasierte Datenbestände und ist in gängigen RDF-Datenbanken sowie über programmgesteuerte Schnittstellen verfügbar.

Mit SPARQL lassen sich sowohl einfache Abfragen, etwa zur Selektion aller Entitäten eines bestimmten Typs, als auch komplexe Anfragen über mehrere verknüpfte Datenquellen hinweg formulieren. Bekannte SPARQL-Endpunkte sind unter anderem der *Wikidata Query Service* (Malyshev et al., 2018), DBpedia (Auer et al., 2007) oder Europeana (Haslhofer und Isaac, 2011). Diese Dienste verdeutlichen, wie semantisch strukturierte Daten systematisch durchsucht und für wissenschaftliche Analysen nutzbar gemacht werden können.

Das in Listing 2.5 dargestellte Beispiel zeigt eine einfache SPARQL-Abfrage, mit der die Vornamen aller Entitäten des Typs *foaf:Person* selektiert werden. Die Abfrage illustriert das grundlegende Prinzip der Mustererkennung in RDF-Graphen, das SPARQL zu einem leistungsfähigen Werkzeug für die explorative Analyse und Integration semantischer Daten macht.

Listing 2.5: Beispiel einer SPARQL-abfrage, bei der der Vornahme (foaf:firstname) einer Person (Entität vom Typ foaf:Person) gesucht wird.

```
PREFIX foaf: <https://http://xmlns.com/foaf/0.1/>

SELECT ?name
WHERE {
    ?x a foaf:Person
    ?x foaf:firstName ?name
}
```

2.2.2 Ontologien zur semantischen Modellierung von Fachwissen

Während RDF ein flexibles Datenmodell zur Beschreibung und Verknüpfung von Informationen bereitstellt, adressieren Ontologien die formale Modellierung von Fachwissen auf einer tieferen semantischen Ebene. Sie definieren Konzepte (Klassen), deren Eigenschaften sowie die Beziehungen zwischen ihnen innerhalb eines abgegrenzten Anwendungsbereichs und legen damit explizit fest, wie Informationen inhaltlich zu interpretieren sind. Ontologien dienen somit nicht allein der strukturellen Organisation von Daten, sondern der formalen Beschreibung ihrer Bedeutung und ihres Kontexts in einer für Menschen und Maschinen gleichermaßen interpretierbaren Form.

Der Begriff der Ontologie hat seinen Ursprung in der theoretischen Philosophie, wo er zur systematischen Kategorisierung von Entitäten und deren Beziehungen verwendet wurde (Robertson-

von Trotha und R. H. Schneider, 2015). Mit der zunehmenden Digitalisierung und der stark wachsenden Menge heterogener Daten im World Wide Web haben Ontologien jedoch eine ausgeprägte praktische Relevanz erlangt. Heute stellen sie eine Schlüsseltechnologie für Wissensmanagementsysteme, semantische Suchmaschinen und datengetriebene Anwendungen dar (Lanquillon und Schacht, 2023).

Im Kontext des Semantic Web werden Ontologien typischerweise mit der Web Ontology Language (OWL) beschrieben. OWL baut auf RDF auf und erweitert dessen Ausdrucksstärke um die Möglichkeit, Klassenhierarchien, komplexe Relationen, Einschränkungen und logische Axiome formal zu definieren. In Kombination mit RDF entsteht so ein standardisiertes, maschinenlesbares Fundament zur Modellierung komplexer Wissensstrukturen, das sowohl die explizite Beschreibung von Fachwissen als auch automatisierte Schlussfolgerungen erlaubt.

Im Kontext des *Semantic Web*, das von Berners-Lee et al. (2001) als Erweiterung des klassischen Webs beschrieben wird, nehmen Ontologien eine zentrale Rolle ein. Sie definieren die konzeptuellen Strukturen, anhand derer Daten interpretiert, verknüpft und über System- und Domänengrenzen hinweg integriert werden können. Daten werden dadurch nicht länger isoliert betrachtet, sondern in ein semantisches Gefüge eingebettet, das ihre Bedeutung explizit und maschinenlesbar beschreibt.

Auf diese Weise bilden Ontologien die Grundlage für ein maschinenlesbares „Web of Linked Data“, in dem sowohl explizit dokumentiertes als auch implizit erschließbares Wissen aus heterogenen Quellen zusammengeführt werden kann. Dies eröffnet neue Möglichkeiten der Wissensintegration und -exploration, unterstützt die Interoperabilität verteilter Datenbestände und ermöglicht eine kontextsensitive, automatisierte Verarbeitung von Informationen.

Für ein nachhaltiges Forschungsdatenmanagement übernehmen Ontologien eine zentrale methodische Funktion, da sie wesentlich zur langfristigen Verständlichkeit, Nachnutzbarkeit und Interoperabilität von Daten beitragen. Damit adressieren sie zentrale Anforderungen der FAIR-Prinzipien:

- **Auffindbarkeit (Findable)**: Durch die Verwendung persistenter Identifikatoren sowie kontrollierter, formal spezifizierter Begriffe ermöglichen Ontologien eine semantisch fundierte Indizierung, die über rein syntaktische Metadaten oder unstrukturierte Schlagwortvergabe hinausgeht.
- **Zugänglichkeit (Accessible)**: Einheitliche, standardisierte Konzepte erleichtern eine konsistente Dokumentation und Interpretation von Daten über unterschiedliche Systeme hinweg.
- **Interoperabilität (Interoperable)**: Ontologien stellen eine gemeinsame semantische Grundlage bereit, auf deren Basis heterogene Datensätze und Anwendungen über Disziplin- und Systemgrenzen hinweg verknüpft werden können.
- **Wiederverwendbarkeit (Reusable)**: Durch die explizite, maschinenlesbare Beschreibung von Bedeutung und Kontext bleibt die Interpretation von Daten auch über längere Zeiträume und in neuen Anwendungsszenarien hinweg stabil.

Über diese Aspekte hinaus spielt die Fähigkeit zur automatisierten Schlussfolgerung (Inferenz)

eine wesentliche Rolle. Die in Ontologien formalisierten Relationen, Einschränkungen und Axiome erlauben es, aus bestehenden Daten implizite Informationen abzuleiten, die nicht explizit modelliert wurden. Dadurch können Zusammenhänge identifiziert, Inkonsistenzen erkannt und neue Fragestellungen generiert werden. Ontologien erweitern ihre Funktion damit über die reine Wissensorganisation hinaus und werden zu einem aktiven Werkzeug für die Analyse und Weiterentwicklung komplexer, vernetzter Datenbestände.

Beim Entwurf von Ontologien sind sowohl der konkrete Anwendungszweck als auch die Zielgruppe der Nutzerinnen und Nutzer entscheidend. Entsprechend lassen sich grundsätzlich zwei Klassen von Ontologien unterscheiden:

Top-Level-Ontologien (allgemeine Ontologien) definieren grundlegende, domänenunabhängige Konzepte und Relationen, die als semantische Basis für eine Vielzahl von Anwendungen dienen. Sie verfolgen das Ziel, eine konsistente, wiederverwendbare Begriffsgrundlage bereitzustellen und die Integration spezialisierter Ontologien zu erleichtern.

Bekannte Beispiele sind *Dublin Core* (Weibel und Koch, 2000), Schema.org (Payne und Verhey, 2022)[2] oder die *Basic Formal Ontology* (BFO) (Otte et al., 2022). Für die Ingenieurwissenschaften ist insbesondere Metadata4Ing (Schembera und Iglezakis, 2020) hervorzuheben, das zentrale Konzepte technischer Disziplinen formalisiert und als Referenzmodell für weiterführende Ontologien dient.

Spezialisierte Ontologien adressieren hingegen klar abgegrenzte Domänen und konkrete Anwendungsszenarien. Sie bauen in der Regel auf bestehenden Top-Level-Ontologien auf, um grundlegende Konzepte wiederzuverwenden, und erweitern diese um domänenspezifisches Fachwissen. Auf diese Weise lassen sich detaillierte Zusammenhänge modellieren, die nicht unmittelbar aus den Rohdaten selbst hervorgehen. In der Praxis sind spezialisierte Ontologien von besonderer Bedeutung, da sie implizites Expertenwissen explizit formalisieren und dadurch eine einheitliche, maschinenlesbare und langfristig nutzbare Wissensbasis schaffen.

2.2.3 Wissensgraphen als integrative Plattform für Wissensrepräsentation

Ein Wissensgraph (*knowledge graph*) bezeichnet ein semantisches Netzwerk, in dem konkrete Entitäten[3] sowie deren Beziehungen in strukturierter, maschinenlesbarer Form repräsentiert werden. Im Kontext dieser Arbeit wird ein Wissensgraph als RDF-basierter Graph verstanden, der in spezialisierten Graphdatenbanken wie *GraphDB* oder *AllegroGraph* gespeichert und verwaltet werden kann. In der graphbasierten Darstellung repräsentieren Knoten einzelne Entitäten (z. B. Personen, Instrumente oder Publikationen), während gerichtete Kanten die semantischen Beziehungen zwischen ihnen abbilden (vgl. Abbildung 2.4).

[2] Der Begriff Ontologie wird hier in einem pragmatischen, im Semantic-Web-Umfeld gebräuchlichen Sinne verwendet und schließt auch formal spezifizierte Metadatenvokabulare ein, sofern sie maschinenlesbar definiert und semantisch referenzierbar sind.

[3] In Wissensgraphen bezeichnet der Begriff Entität eine eindeutig identifizierbare Ressource der realen oder konzeptuellen Welt.

Obwohl auch Ontologien graphbasiert dargestellt werden können, beschreibt der Begriff Wissensgraph in der Regel nicht das abstrakte Modell selbst, sondern dessen konkrete Instanziierung. Ontologien definieren die konzeptuelle Struktur eines Anwendungsbereichs, während Wissensgraphen diese Struktur mit realen Entitäten, Attributen und Fakten füllen. Wissensgraphen enthalten somit die datengetriebene Ausprägung ontologischer Modelle und stellen die operative Ebene semantischer Wissensrepräsentation dar, auf der Abfragen, Inferenzmechanismen und analytische Auswertungen durchgeführt werden können (Lanquillon und Schacht, 2023). Der Begriff *knowledge graph* wurde insbesondere durch Google im Jahr 2012 geprägt, um strukturierte Wissensrepräsentationen zur Verbesserung von Such- und Informationsdiensten zu beschreiben (Ehrlinger und Wöß, 2016).

Wissensgraphen zeichnen sich durch eine hohe Flexibilität und Erweiterbarkeit aus. Sie erlauben die Integration heterogener Datenquellen, die Abbildung komplexer, mehrstufiger Beziehungen sowie die schrittweise Erweiterung bestehender Wissensbestände ohne grundlegende Strukturänderungen. Diese Eigenschaften machen Wissensgraphen besonders geeignet für datenintensive Anwendungen, in denen Informationen aus unterschiedlichen Domänen, Formaten und Kontexten zusammengeführt werden müssen. Entsprechend finden sie sowohl in wissenschaftlichen als auch in kommerziellen Anwendungsfeldern Verwendung, etwa in Empfehlungssystemen, der Finanzwirtschaft oder im Gesundheitswesen.

Im Kontext eines FAIR-orientierten Forschungsdatenmanagements übernehmen Wissensgraphen eine zentrale integrative Rolle. Sie ermöglichen die semantische Verknüpfung von Forschungsdaten, Metadaten, Provenienzinformationen und externen Referenzen in einer konsistenten Struktur. Dadurch unterstützen sie insbesondere die Interoperabilität und Wiederverwendbarkeit von Daten, da Zusammenhänge explizit modelliert und maschinell auswertbar vorliegen.

Für die wissenschaftliche Praxis sind insbesondere offene, kollaborativ gepflegte Wissensgraphen von Bedeutung. Beispiele hierfür sind Wikidata (Vrandečić und Krötzsch, 2014) und DBpedia (Auer et al., 2007). Wikidata stellt eine frei zugängliche, gemeinschaftlich kuratierte Wissensbasis dar und zählt mit über hundert Millionen Datenobjekten zu den weltweit größten offenen Wissensgraphen (*Wikidata: Statistiken* 2025). Die Plattform erlaubt eine standardisierte, semantisch präzise Beschreibung von Forschungsobjekten und wird auch im Rahmen der vorliegenden Arbeit als Referenz- und Integrationsressource genutzt. So kann beispielsweise ein Neodym-dotierter Yttrium-Aluminium-Granat-Laser eindeutig über das Wikidata-Item Q1110547 referenziert werden.

Wikidata weist eine enge konzeptionelle Nähe zu den FAIR-Prinzipien auf, auch wenn die Plattform nicht als formell FAIR-zertifizierte Forschungsdateninfrastruktur konzipiert ist. Die Bereitstellung offener Schnittstellen sowie maschinenlesbarer Repräsentationen in RDF und JSON-LD unterstützt sowohl die Auffindbarkeit als auch die Zugänglichkeit von Informationen. Darüber hinaus ermöglicht die konsequente Nutzung global eindeutiger Identifikatoren die Verknüpfung mit externen Datenquellen und wissenschaftlichen Repositorien, was eine hohe Interoperabilität gewährleistet. Ontologien und kontrollierte Vokabulare können direkt eingebunden oder erweitert werden, sodass auch domänenspezifische Anforderungen abgebildet werden können.

Ein weiteres zentrales Merkmal von Wikidata ist die kollaborative Pflege der Inhalte. Forscher, Institutionen und Projekte aus unterschiedlichen Disziplinen können gemeinsam an der Erstellung, Korrektur und Erweiterung von Einträgen mitwirken. Dieses gemeinschaftliche Vorgehen führt zu einer kontinuierlichen Verbesserung der Datenqualität und unterstützt den langfristigen Erhalt sowie die Weiterentwicklung wissenschaftlicher Wissensbestände.

Zusammenfassend fungieren Wissensgraphen als integrative Plattform für die Zusammenführung, Verknüpfung und Nachnutzung semantisch strukturierter Informationen. Sie bilden die operative Umsetzung ontologiebasierter Wissensrepräsentation und stellen damit einen zentralen Baustein für eine vernetzte, transparente und FAIR-konforme wissenschaftliche Dateninfrastruktur dar.

2.3 Wissenschaftliche Einordnung und verwandte Arbeiten

Spätestens seit der Formulierung der FAIR-Prinzipien sowie der Etablierung zentraler Institutionen wie der NFDI und der Research Data Alliance (RDA) (Berman und Crosas, 2020) hat sich Forschungsdatenmanagement von einer unterstützenden Zusatzaufgabe zu einem integralen Bestandteil guter wissenschaftlicher Praxis entwickelt. Während auf konzeptioneller Ebene weitgehender Konsens über die Bedeutung von Auffindbarkeit, Interoperabilität und Wiederverwendbarkeit besteht, zeigen sich in der praktischen Umsetzung – insbesondere in den Ingenieurwissenschaften – weiterhin erhebliche Defizite.

Für die Einordnung der vorliegenden Arbeit werden im Folgenden ausgewählte Metadatenkonzepte und Forschungsdatenmanagementlösungen betrachtet. Ziel ist dabei nicht eine vollständige Bestandsaufnahme, sondern eine Analyse bestehender Ansätze im Hinblick auf ihre Eignung zur technisch belastbaren und praktikablen Umsetzung der FAIR-Prinzipien für komplexe ingenieurwissenschaftliche Forschungsdaten. Auf dieser Grundlage wird der spezifische Beitrag dieser Arbeit klar abgegrenzt.

2.3.1 Ausgewählte Metadatenkonzepte für die Ingenieurwissenschaften

Forschungsdatenmanagement wird in den Ingenieurwissenschaften bislang mit stark variierendem Reifegrad praktiziert. Die eingesetzten Methoden und Werkzeuge unterscheiden sich erheblich, was sowohl auf die begrenzte Standardisierbarkeit domänenspezifischer Datenstrukturen als auch auf die heterogenen experimentellen und numerischen Arbeitsweisen zurückzuführen ist (Horsch, Morgado et al., 2021). Entsprechend hängt die Wahl und Durchsetzung von Metadatenstandards eng mit dem Grad der Vernetzung innerhalb der jeweiligen Fachgemeinschaften zusammen.

Domänenspezifische Metadatenmodelle stellen dabei eine notwendige, jedoch keine hinreichende Bedingung für FAIR-konformes Forschungsdatenmanagement dar. Zwar ermöglichen sie eine strukturierte Beschreibung von Daten und Prozessen, bleiben jedoch häufig auf eine bestimmte Abstraktionsebene oder einen begrenzten Anwendungskontext beschränkt. Im Folgenden werden exemplarisch Konzepte vorgestellt, die für numerische und experimentelle Anwendungen relevant sind und im Verlauf dieser Arbeit gezielt aufgegriffen oder erweitert werden.

Grundlegend ist festzuhalten, dass eine belastbare Umsetzung der FAIR-Prinzipien zunehmend auf semantischen Technologien beruht. Diese ermöglichen eine eindeutige, menschen- und maschineninterpretierbare Zuordnung von Bedeutungen und gehen damit über rein syntaktische Metadatenbeschreibungen hinaus (Horsch, Morgado et al., 2021). Viele etablierte Standards sind historisch auf menschliche Lesbarkeit ausgerichtet und nur eingeschränkt für automatisierte Verarbeitung, Validierung und Verknüpfung geeignet. Technologien wie XML-Schemata oder RDF aus dem Semantic-Web-Stack schließen diese Lücke teilweise, adressieren jedoch unterschiedliche Ebenen der Metadatenmodellierung mit jeweils spezifischen Stärken und Einschränkungen.

EngMeta Das Metadatenmodell Engineering Metadata (EngMeta) wurde im Rahmen des DIPL-Projekts entwickelt und richtet sich insbesondere an numerische Simulationen in den Ingenieurwissenschaften (Schembera und Iglezakis, 2020; Schembera, Selent et al., 2019; Selent et al., 2019). Als XML-basiertes Schema stellt EngMeta ein etabliertes, formal strukturiertes Metadatenformat dar, das sowohl menschliche als auch maschinelle Lesbarkeit gewährleistet. Ziel ist die systematische Dokumentation großer Simulationsdatensätze entlang der Phasen *preparation*, *production* und *evaluation*.

Ein wesentlicher Schwerpunkt von EngMeta liegt auf der Beschreibung von Provenienz und Kontextinformationen, etwa zu Autoren, Projekten oder eingesetzten Rechenressourcen. Damit adressiert EngMeta zentrale FAIR-Aspekte wie Nachvollziehbarkeit und Wiederverwendbarkeit. Zugleich zeigt sich jedoch, dass der Ansatz primär auf eine strukturierte Beschreibung abgeschlossener Datensätze ausgerichtet ist. Die semantische Verknüpfung einzelner Metadaten über Projekt- oder Systemgrenzen hinweg sowie deren formale Validierung sind nicht integraler Bestandteil des Modells.

Wie von Horsch, Morgado et al. (2021) hervorgehoben, ist EngMeta weniger als statischer Standard denn als Form „wissenschaftlicher Kommunikation“ zu verstehen. Diese Einordnung unterstreicht den konzeptionellen Wert des Ansatzes, verweist jedoch zugleich auf dessen begrenzte Eignung als technische Grundlage für ein durchgängig maschineninterpretierbares und interoperables Forschungsdatenmanagement.

Metadata4Ing Mit Metadata4Ing wurde im Kontext der NFDI4Ing eine semantisch fundierte, domänenübergreifende RDF-basierte Ontologie entwickelt, die über rein syntaktische Metadatenbeschreibungen hinausgeht. Sie stellt eine semantische Weiterentwicklung des vorgelagerten XML-Schemas EngMeta dar und zielt auf die formale Repräsentation wissenschaftlicher Aktivitäten, ihrer Ergebnisse sowie der zugrunde liegenden Provenienz von Daten, Materialien und Werkzeugen ab (Arndt et al., 2023; Iglezakis et al., 2023).

Metadata4Ing ist von Beginn an als RDF-basierte Ontologie konzipiert und damit explizit auf Interoperabilität, Erweiterbarkeit und maschinelle Verarbeitung ausgelegt. Sie nutzt und erweitert etablierte Ontologien wie PROV (für Aktivitäten und Provenienz) und DCAT (für Datensätze) und schafft damit eine semantische Brücke zwischen domänenübergreifenden Standards und ingenieurwissenschaftlichen Anforderungen. Zentrale Konzepte wie der generalisierte Prozessschritt (*m4i:ProcessingStep*) ermöglichen die semantische Verknüpfung von Methoden, Werkzeugen, Akteuren und quantitativen Größen über unterschiedliche Anwendungskontexte hinweg.

Metadata4Ing ist bewusst nicht an eine spezifische Fachdomäne gebunden, sondern verfolgt einen abstrakten, generischen Ansatz mit modularem Aufbau. Dies sichert eine breite Wiederverwendbarkeit und ermöglicht es, domänenspezifische Spezialisierungen durch Subontologien zu schaffen, etwa für Workflows in High-Performance-Computing-Umgebungen. Durch diese Eigenschaften stellt Metadata4Ing ein zentrales Referenzmodell für semantisch orientiertes Forschungsdatenmanagement in den Ingenieurwissenschaften dar.

Die konzeptionelle Stärke von Metadata4Ing geht jedoch mit einer bewusst hohen Abstrak-

tionsebene einher. Die Ontologie definiert einen semantischen Rahmen für wissenschaftliche Prozesse und Ergebnisse, trifft jedoch noch keine verbindlichen Aussagen zur technischen Integration in konkrete Datenformate, zur strukturellen Kopplung mit Primärdaten oder zur formalen Validierung von Metadaten. Erste praktische Implementierungsleitfäden und SHACL-basierte Validierungsprofile entstehen derzeit durch Community-Projekte, sodass sich die Brücke zwischen ontologischem Modell und operativer Praxis kontinuierlich verschärft. Es bleibt zu diesem Zeitpunkt jedoch noch offen, wie sich die abstrakten Konzepte der Ontologie in konkrete, automatisierbare Forschungsdatenmanagementprozesse überführen lassen, etwa im Umgang mit komplexen und hierarchisch strukturierten Primärdaten.

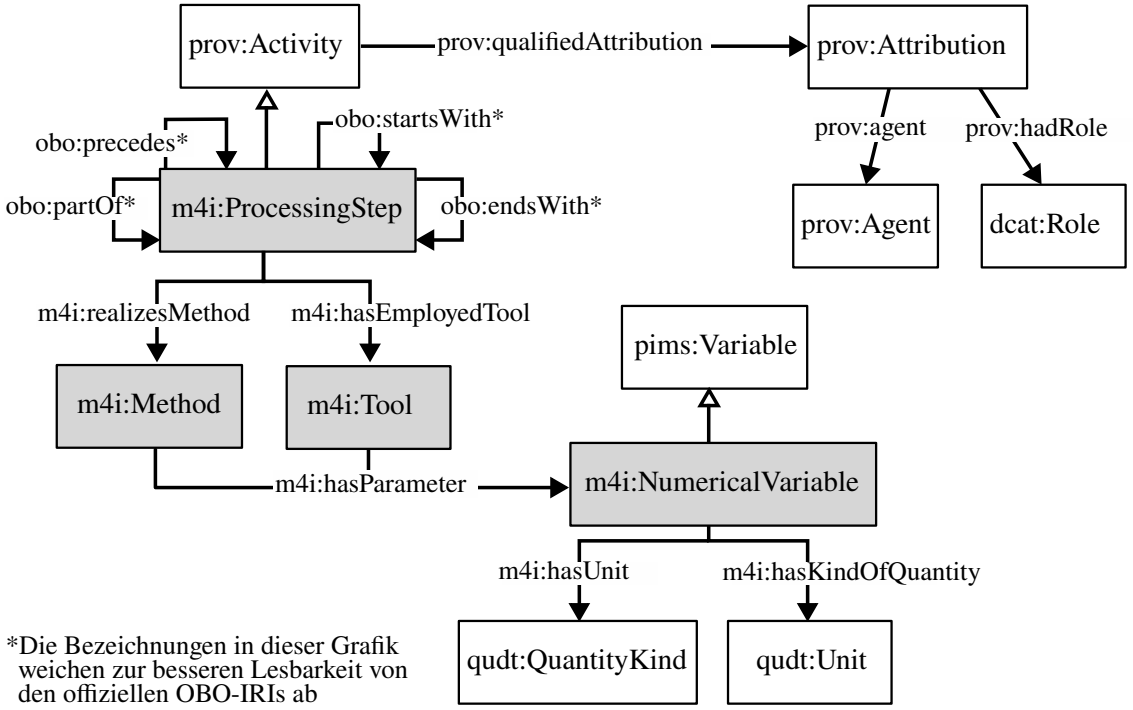


Abbildung 2.5: Schematische Darstellung ausgewählter Kernbausteine der Ontologie Metadata4Ing. Im Zentrum steht der generalisierte Prozessschritt (*m4i:ProcessingStep*), der als semantischer Ankerpunkt zur Verknüpfung von Werkzeugen, Methoden, beteiligten Akteuren und Kenngrößen dient. Klassen aus etablierten Ontologien wie PROV und DCAT sind weiß hinterlegt, während ontologiespezifische Konzepte grau gekennzeichnet sind. Die Visualisierung orientiert sich an Arndt et al. (2023).

Abbildung 2.5 zeigt einen Ausschnitt der zentralen Konzepte von Metadata4Ing sowie deren Relationen zu anderen etablierten Ontologien. Im Mittelpunkt steht der generalisierte Prozessschritt *m4i:ProcessingStep*, der wissenschaftliche Abläufe beschreibt und sowohl eingesetzte Werkzeuge (z. B. Hardware oder Software) als auch angewandte Methoden referenziert. Darüber hinaus erlaubt die Ontologie die Verknüpfung von Prozessen mit beteiligten Akteuren (z. B. Personen über *foaf:Person* oder Organisationen) und integriert zugleich quantitative Größen. Diese werden exemplarisch durch die Klasse *m4i:NumericalVariable* repräsentiert, die numerische

Variablen mit Methoden und Werkzeugen in Beziehung setzt. Ergänzend unterscheidet Metadata4Ing weitere Klassen zur Abbildung textueller Eigenschaften oder kategorialer Merkmale und wird kontinuierlich um domänenspezifische Module erweitert, etwa zur Beschreibung von Workflows und Berechnungsprozessen.

CodeMeta Software ist heute nicht mehr nur ein Hilfsmittel zur Datenverarbeitung, sondern stellt häufig selbst ein zentrales wissenschaftliches Ergebnis dar. Mit CodeMeta steht ein standardisiertes Metadatenmodell zur Verfügung, das die strukturierte Beschreibung, Referenzierung und Zitierbarkeit wissenschaftlicher Software ermöglicht (Jones et al., 2023). CodeMeta definiert hierfür ein kontrolliertes Vokabular, das insbesondere Angaben zu Funktionalität, Versionierung, Lizenzierung, Autorenschaft und Abhängigkeiten umfasst.

Ein zentrales Merkmal von CodeMeta ist die enge Anlehnung an etablierte Metadatenstandards und -ökosysteme wie Dublin Core, DataCite und Schema.org. Dadurch lässt sich Software nahtlos in bestehende Publikations- und Repositorieninfrastrukturen integrieren und als eigenständige, persistent referenzierbare Ressource behandeln. Insbesondere die Serialisierung in JSON-LD unterstützt eine maschinenlesbare Repräsentation und erleichtert die automatisierte Verarbeitung sowie die Verknüpfung mit anderen Forschungsobjekten.

Durch die Möglichkeit, Software explizit mit Datensätzen, Publikationen oder Projekten zu verknüpfen, leistet CodeMeta einen wichtigen Beitrag zur Umsetzung der FAIR-Prinzipien, insbesondere im Hinblick auf Auffindbarkeit und Nachnutzbarkeit wissenschaftlicher Software. Zugleich bleibt der Anwendungsbereich von CodeMeta bewusst auf Softwareartefakte beschränkt. Aspekte der internen Strukturierung von Forschungsdaten, deren semantische Annotation oder die Beschreibung komplexer experimenteller und numerischer Prozesse sind nicht Gegenstand des Modells.

CodeMeta ist daher als spezialisierter Baustein innerhalb eines umfassenderen Forschungsdatenmanagementkonzepts zu verstehen, der die Rolle von Software als zitierbare und vernetzbare Ressource adressiert, jedoch keine Aussagen zur Organisation, Semantisierung oder Validierung der erzeugten Forschungsdaten selbst trifft.

Ein JSON-LD-Beispiel eines im Rahmen dieser Arbeit entwickelten Softwareprodukts ist in Listing A.10 dargestellt.

DCAT Das *Data Catalog Vocabulary* (DCAT) ist ein vom W3C standardisiertes Vokabular zur Beschreibung, zum Austausch und zur Interoperabilität von Datensätzen in offenen Datenkatalogen. Ziel von DCAT ist es, die Auffindbarkeit und den Zugang zu Daten zu verbessern, indem Datensätze sowie deren Bereitstellungen strukturiert und maschinenlesbar in einem RDF-basierten Modell beschrieben werden.

DCAT definiert hierzu zentrale Konzepte wie *Catalog*, *Dataset* und *Distribution*. Ein Katalog beschreibt eine Sammlung von Datensätzen, wie sie beispielsweise in Repositorien wie Zenodo, Figshare oder dem European Data Portal angeboten werden. Ein Datensatz repräsentiert eine logisch zusammengehörige Datenmenge, während die zugehörige *Distribution* die konkrete

technische Bereitstellung beschreibt. Diese umfasst unter anderem Angaben zur Zugriffsadresse, zum Datenformat sowie zum Medientyp. Ergänzend lassen sich über Ontologien wie PROV-O Beziehungen zu Akteuren, Organisationen oder Prozessen modellieren und bei Bedarf erweitern.

Abbildung 2.6 zeigt einen Ausschnitt der zentralen Klassen der DCAT-Ontologie sowie deren grundlegende Beziehungen. Die Darstellung verdeutlicht den Fokus von DCAT auf die Beschreibung von Datensätzen auf Katalog- und Zugriffsebene, unabhängig von der internen Struktur oder dem inhaltlichen Aufbau der zugrunde liegenden Daten.

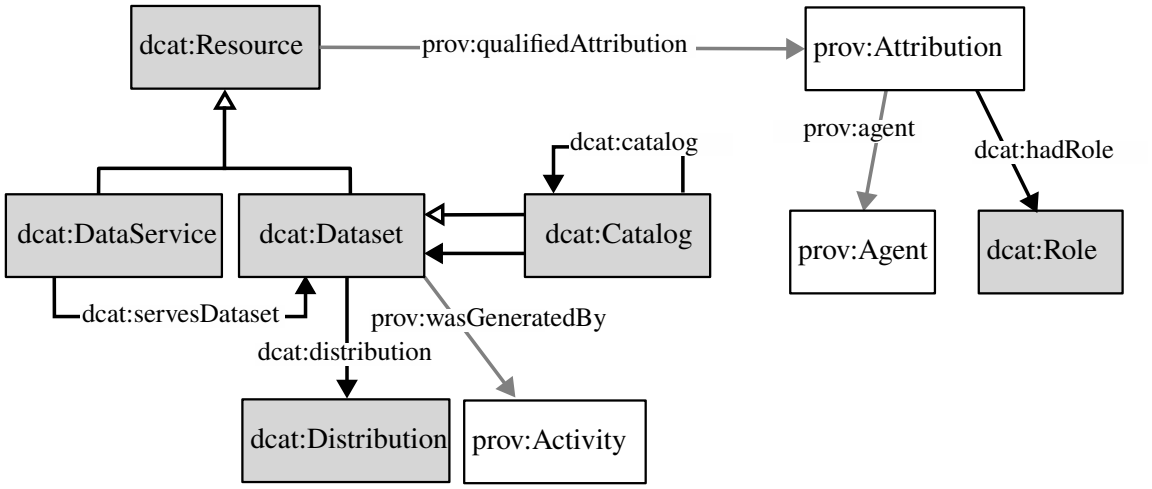


Abbildung 2.6: Auswahl zentraler Klassen der DCAT-Ontologie (grau hinterlegt). In Kombination mit Ontologien wie PROV-O lassen sich Datensätze, deren Bereitstellungen (*dcat:Distribution*) sowie beteiligte Agenten (Personen oder Organisationen) beschreiben.

Aufgrund seiner breiten Akzeptanz in etablierten Repositorien und Dateninfrastrukturen nimmt DCAT eine zentrale Rolle im aktuellen Ökosystem des Forschungsdatenmanagements ein. Insbesondere fungiert das Vokabular als verbindende Beschreibungsebene zwischen datenproduzierenden Anwendungen, Repositorien und nachnutzenden Diensten. Damit bildet DCAT einen faktischen Referenzstandard für die Veröffentlichung, Auffindbarkeit und den standardisierten Zugang zu Forschungsdaten in offenen Infrastrukturen.

Zugleich ist DCAT bewusst auf die Katalogisierung und Bereitstellung von Datensätzen beschränkt. Aussagen zur internen Strukturierung komplexer Forschungsdaten, zu deren semantischer Annotation oder zur formalen Validierung von Metadaten sind nicht Bestandteil des Modells. DCAT ist daher als infrastrukturelles Bindeglied im Forschungsdatenmanagement zu verstehen, nicht jedoch als umfassendes Metadatenmodell für die inhaltliche Beschreibung oder semantische Durchdringung wissenschaftlicher Daten.

Semantic Sensor Network Ontology Die Semantic Sensor Network Ontology (SSN) ist ebenfalls ein vom W3C standardisiertes Ontologiemodell. Es dient zur formalen Beschreibung von Sensoren, Beobachtungen sowie deren Kontext. SSN wurde entwickelt, um das zentrale

Problem der semantischen Interoperabilität in heterogenen Sensorsystemen zu adressieren und die strukturierte Integration von Messdaten in verteilten Anwendungen zu ermöglichen. Typische Anwendungsfelder sind das Internet der Dinge (IoT), die Umweltuntersuchungen sowie wissenschaftliche Experimente.

Ursprünglich wurde die SSN-Ontologie im Jahr 2011 im Rahmen eines *W3C-Incubator*-Projekts vorgestellt. Eine erste umfassende Beschreibung findet sich in der Arbeit von *COMPTON201225*, die das grundlegende Ontologiemodell und dessen Einsatzmöglichkeiten darstellt. Im Zuge der späteren Standardisierung wurde die Ontologie grundlegend überarbeitet und als *W3C Recommendation* veröffentlicht. Die aktuelle Version folgt einem modularen Aufbau und umfasst zwei eng miteinander verbundene Teile: die SSN-Ontologie im engeren Sinne sowie die SOSA-Ontologie (Sensor, Observation, Sample, and Actuator), die gemeinsam im W3C-Standard dokumentiert sind (vgl. (Haller et al., 2017)).

Die SOSA-Ontologie bildet dabei ein leichtgewichtiges Kernmodell, das zentrale Konzepte wie Sensor, Beobachtung, beobachtete Eigenschaft und betrachtetes Objekt bereitstellt. Sie ist domänenneutral gehalten und ermöglicht eine einfache, breit einsetzbare Beschreibung von Messvorgängen. Die SSN-Ontologie erweitert dieses Kernmodell um zusätzliche semantische Schichten, die insbesondere für die Beschreibung von Sensorsystemen, deren Fähigkeiten, Einsatzbedingungen und Messkontext relevant sind.

Ein wesentlicher Bestandteil dieser Erweiterungen ist das sogenannte SSN-System-Modul. Dieses Modul stellt spezialisierte Klassen zur Beschreibung von Systemeigenschaften bereit, darunter Messbereiche, Genauigkeitsangaben, Auflösungen oder Reaktionszeiten. Damit erlaubt SSN-System eine formale, maschinenlesbare Abbildung von technischen Spezifikationen, wie sie typischerweise in Datenblättern oder Herstellerangaben zu finden sind. Im Gegensatz zur ursprünglichen, monolithischen SSN-Inkubator-Ontologie sind diese Aspekte in der aktuellen Version klar modularisiert und können gezielt ergänzt werden.

Durch die Kombination von SOSA, SSN und dem SSN-System-Modul entsteht ein flexibles Ontologiemodell, das sowohl einfache Messbeschreibungen als auch detaillierte Dokumentationen von Sensorsystemen unterstützt. Sensoren können dabei mit ihren Leistungsmerkmalen, etwa Messgenauigkeit oder Messbereich, unter definierten Bedingungen beschrieben und eindeutig mit Beobachtungen verknüpft werden. Dies ermöglicht eine objektive und nachvollziehbare Dokumentation experimenteller Aufbauten sowie eine spätere automatisierte Auswertung der erhobenen Messdaten.

In dieser Arbeit wird die Semantic Sensor Network Ontology (SSN) verwendet, um experimentelle Beschreibungen formal abzubilden. Abbildung 2.7 zeigt die für diese Arbeit relevanten Kernkomponenten der Ontologie sowie deren Beziehungen zueinander in einer vereinfachten Darstellung.

CF Convention Die *Climate and Forecast* (CF) *Convention* ist ein Standard, der speziell für die Klimaforschung entwickelt wurde. Ziel ist die einheitliche Nutzung von Metadaten durch die Bereitstellung eines zentralen, kontrollierten Vokabulars. Auf diese Weise soll ein nachhaltiger Austausch sowie eine effiziente Verarbeitung von Klimadaten ermöglicht werden (Hassell et al.,

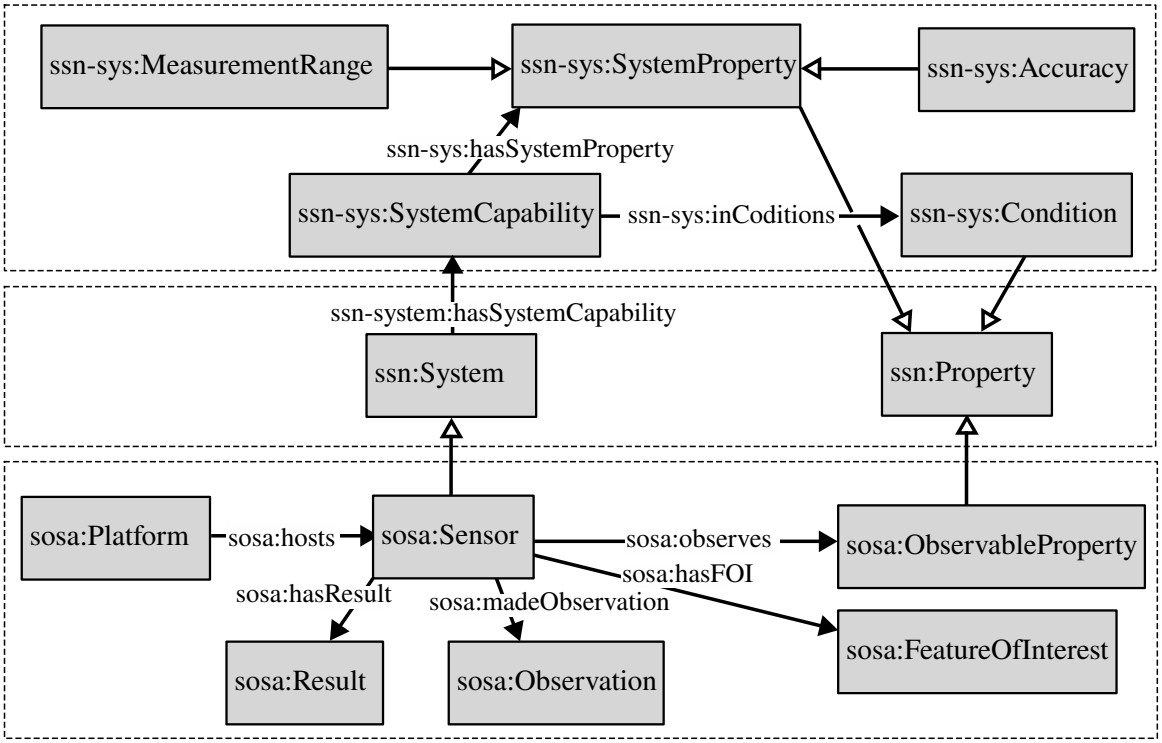


Abbildung 2.7: Zentrale und für die Arbeit relevante Komponenten der Semantic Sensor Network Ontology zur Beschreibung von Beobachtungen und Sensoren sowie deren Integration in einen Messaufbau. Die Boxen grenzen die Module der Ontologie von einander ab.

2017). Der Standard wurde 2003 eingeführt und definiert eine konsistente Struktur für Metadaten in netCDF4 (Network Common Data Form)-Dateien, die insbesondere in der Klimaforschung, Meteorologie und Ozeanografie breite Anwendung finden.

Die vorliegende Arbeit greift dieses Konzept auf. Für weiterführende Details sei an dieser Stelle auf die grundlegenden Ausführungen von Hassell et al. (2017), die offizielle Online-Dokumentation (M. Harris, 2025a) sowie auf Unterabschnitt 4.6.1 verwiesen.

Im Vergleich zu domänenübergreifenden Ansätzen wie Metadata4Ing oder EngMeta, die eine breite Abdeckung ingenieurwissenschaftlicher Anwendungen anstreben, ist die CF Convention stark auf den Bereich der Klimaforschung zugeschnitten. Während Metadata4Ing auf der Basis semantischer Technologien eine flexible Wiederverwendung und Interoperabilität über Disziplingrenzen hinweg ermöglicht und EngMeta insbesondere numerische Simulationen in den Ingenieurwissenschaften adressiert, verfolgt die CF Convention einen klar domänenspezifischen Fokus. Ihr vorrangiges Ziel ist die Vereinheitlichung der Metadatenbeschreibung innerhalb eines definierten Anwendungsfeldes (Klimaforschung, Meteorologie, Ozeanografie).

Die CF Conventions sind damit weniger als universell einsetzbarer Metadatenstandard zu verstehen, sondern vielmehr als etablierte und hoch spezialisierte Community-Lösung. Aufgrund ihrer engen Verzahnung mit dem weit verbreiteten netCDF4-Format besitzen sie jedoch auch über ihr ursprüngliches Anwendungsfeld hinaus Relevanz, etwa in benachbarten Disziplinen

wie der Strömungsmechanik, in denen ähnliche Datenstrukturen und Anforderungen an Metadaten auftreten. Vor diesem Hintergrund greift die vorliegende Arbeit die CF Conventions als Referenzstandard auf und untersucht deren Übertragbarkeit sowie mögliche Erweiterungen für das Forschungsdatenmanagement in der Strömungsmechanik.

Die Betrachtung der verschiedenen Metadatenmodelle und Ontologien zeigt, dass ein nachhaltiges Forschungsdatenmanagement für datenintensive ingenieurwissenschaftliche Anwendungen nicht durch die isolierte Anwendung einzelner Standards erreicht werden kann. Erforderlich ist vielmehr ein integrierter Ansatz, der bestehende Konzepte gezielt kombiniert, deren jeweilige Stärken nutzt und die identifizierten Lücken zwischen Datenstruktur, semantischer Beschreibung und infrastruktureller Einbettung systematisch adressiert.

2.3.2 Verwandte Forschungsdatenmanagementlösungen

In diesem Abschnitt werden ausgewählte Forschungsdatenmanagementlösungen betrachtet, die entweder explizit durch die FAIR-Prinzipien motiviert sind oder das Dateiformat HDF5 als zentrale Grundlage verwenden. Ziel ist es, diese Ansätze im Kontext der zuvor diskutierten Metadatenkonzepte einzuordnen und deren Stärken sowie Grenzen im Hinblick auf datenintensive ingenieurwissenschaftliche Anwendungen herauszuarbeiten.

FAIR-konforme Forschungsdatenplattformen

Die breite Einsetzbarkeit des hierarchischen Dateiformats HDF5 spiegelt sich in einer Vielzahl von Forschungsdatenmanagementlösungen wider, insbesondere in den Natur- und Materialwissenschaften, in denen große Mengen komplex strukturierter Daten entstehen. Viele dieser Plattformen sind explizit an den FAIR-Prinzipien ausgerichtet und kombinieren leistungsfähige Datenformate mit Metadaten- und Infrastrukturkonzepten zur langfristigen Nachnutzung wissenschaftlicher Daten.

Eine prominente Lösung stellt die am Karlsruher Institut für Technologie entwickelte Plattform *Kadi4Mat* dar, die ein modulares und erweiterbares Forschungsdatenmanagement für die Materialwissenschaften bereitstellt (Brandt et al., 2021). Kadi4Mat ermöglicht die strukturierte Erfassung, Verwaltung und Wiederverwendung heterogener Forschungsdaten durch HDF5-basierte Datenformate und strukturierte Metadatenerfassung. Die Plattform weist jedem Datensatz persistente Identifikatoren (DOI) zu und versioniert Datenobjekte, um Nachvollziehbarkeit wissenschaftlicher Ergebnisse und langfristige Referenzierbarkeit zu gewährleisten. Durch die Kombination von Datenmanagement, Metadatenstandards und strukturierten Workflows trägt Kadi4Mat zur maschinellen Auswertbarkeit und Interoperabilität von Forschungsdaten bei.

Ein weiteres Beispiel ist das DFG-geförderte Konsortium FAIRmat (FAIR Data Infrastructure for Condensed-Matter Physics and the Chemical Physics of Solids), das eine fachspezifische, FAIR-konforme Dateninfrastruktur für die Materialwissenschaften entwickelt (Junkes et al., 2022). Zentrale technische Grundlage bildet die Plattform NOMAD (Novel Materials Discovery), ein

offenes Ökosystem zur Speicherung, Annotation und Analyse experimenteller und theoretischer Daten aus der Materialforschung. Zur effizienten Organisation, Verarbeitung und langfristigen Archivierung der entstehenden Datensätze nutzt NOMAD das hierarchische Dateiformat HDF5.

NOMAD stellt standardisierte Metadatenmodelle und Schnittstellen zu etablierten Metadatakatalogen sowie Werkzeuge zur Visualisierung, datengetriebenen Analyse und zur Integration elektronischer Laborbücher bereit (Ghiringhelli et al., 2017). Die Plattform folgt einem föderierten Infrastrukturansatz mit lokalen Datenservern und einem zentralen Metadatenportal. FAIRmat demonstriert damit exemplarisch, wie domänenspezifische Forschungsdatenplattformen in übergeordnete nationale und internationale Datenlandschaften eingebettet werden können.

Über konkrete Plattformimplementierungen hinaus betonen auch Aggour et al. (2024) die zentrale Bedeutung der FAIR-Prinzipien für ein nachhaltiges Forschungsdatenmanagement in den Materialwissenschaften. Sie beschreiben eine integrierte Infrastruktur aus verteilter Datenspeicherung, semantischen Annotations- und Zugriffsebenen sowie nachgelagerten Analysediensten, in der Wissensgraphen als verbindendes Element zur Verbesserung von Auffindbarkeit, Zugänglichkeit und Wiederverwendbarkeit fungieren. Ähnliche disziplinspezifische Umsetzungen FAIR-konformer Dateninfrastrukturen entstehen auch in anderen Fachgebieten: im Biomedizinbereich (Liao et al., 2024), in Geowissenschaften (Kinkade und Shepherd, 2022) und in der Umweltforschung (Queralt-Rosinach et al., 2022).

Im Vergleich zu diesen etablierten Lösungen in angrenzenden Disziplinen steht die Entwicklung FAIR-konformer Dateninfrastrukturen in der Strömungsmechanik noch am Anfang. Insbesondere sind bislang keine Lösungen etabliert, die in vergleichbarer Weise strukturierte Primärdaten, semantisch eindeutige Metadaten und eine FAIR-orientierte Infrastruktur konsistent integrieren.

HDF5-basierte Abfragesysteme und Erweiterungen

Neben den zuvor betrachteten plattformorientierten Forschungsdatenmanagementlösungen, die typischerweise von spezialisierten Konsortien mit breiter Nutzerbasis getragen werden, adressieren zahlreiche Arbeiten das Forschungsdatenmanagement auf einer deutlich niedrigeren Abstraktionsebene. Im Fokus stehen dabei direkte Erweiterungen des Dateiformats HDF5, um steigenden Anforderungen an Effizienz, Skalierbarkeit und flexible Datenabfrage gerecht zu werden. Grundsätzlich lassen sich zwei zentrale Entwicklungsrichtungen unterscheiden: zum einen die Optimierung der internen Dateiorganisation und Zugriffsmuster, zum anderen die Entwicklung spezialisierter Abfragesysteme für große und komplex strukturierte Datensätze.

Ein Beispiel für letztere Kategorie stellt *HDF5-QL* dar, eine Erweiterung, die relationale Abfragekonzepte auf HDF5-Daten überträgt und eine deklarative Abfragesprache einführt (*HDFql - The easy way to manage HDF5 data* 2024). Dadurch werden selektive Zugriffe auf Teilmengen großer Dateien ermöglicht, ohne diese vollständig in den Arbeitsspeicher laden zu müssen. Ähnlich adressiert *HDF5 Fast Query* die performante Selektion in multidimensionalen Datensätzen durch den Einsatz bitmap-basierter Indexstrukturen, die schnelle Filterungen relevanter Datenbereiche erlauben (Gosink et al., 2006). Diese Ansätze verbessern insbesondere die technische Zugänglichkeit und Effizienz der Datenverarbeitung, treffen jedoch keine Aussagen zur semantischen Beschreibung der Daten oder zu deren domänenübergreifender Interoperabilität.

Auch im Bereich des Datenaustauschs existieren HDF5-basierte Konzepte. Das von De Carlo et al. (2014) vorgestellte *Data Exchange Model* zielt darauf ab, die Speicherung und den Austausch wissenschaftlicher Daten durch eine standardisierte, aber flexible Struktur zu vereinfachen. Der Schwerpunkt liegt dabei auf technischer Konsistenz, Nachvollziehbarkeit und Wiederverwendbarkeit von Analyse- und Austauschprozessen, etwa in Synchrotron-Techniken wie der Tomografie oder Fluoreszenzspektroskopie. Eine explizite semantische Repräsentation der Daten oder eine formale Kopplung an domänenübergreifende Metadatenmodelle ist jedoch nicht Bestandteil dieses Ansatzes.

Einen stärker domänenspezifischen Beitrag für numerische Simulationen und experimentelle Studien in der Strömungsmechanik und Thermodynamik liefern Selent et al. (2019) mit der Entwicklung des Metadatenschemas EngMeta. Ziel von EngMeta ist die strukturierte Erfassung technischer, prozessualer und kontextbezogener Informationen, etwa zu Simulationsparametern, numerischen Verfahren, Materialien oder Messumgebungen. Damit adressiert der Ansatz insbesondere Aspekte der Nachvollziehbarkeit und Wiederverwendbarkeit datenintensiver CFD- und Thermodynamikdaten. Die praktische Umsetzung erfolgt durch die Integration von EngMeta in das Repositoriumssystem Dataverse, wofür spezifische Metadatenblöcke konfiguriert wurden. EngMeta stellt damit ein leistungsfähiges, FAIR-orientiertes Metadatenmodell dar, bleibt jedoch auf die Metadatenebene beschränkt und ist nicht direkt mit der internen Struktur oder Organisation der zugrunde liegenden HDF5-Daten gekoppelt.

Eine konkrete Lösung für die Strömungsmechanik wird von Preuß und Pelz (2018) vorgestellt. Die Autoren entwickeln ein modulares, objektorientiertes Datenmodell zur strukturierten Speicherung und Verarbeitung von Messdaten und zugehörigen Metadaten in Ventilatorprüfständen, wobei HDF5 als zentrales Speichersystem eingesetzt wird. Ziel ist die Erzeugung selbstbeschreibender Datenpakete, die sämtliche Informationen zur Versuchsdurchführung, Instrumentierung und Kalibrierung enthalten. Die vorgeschlagene Softwarearchitektur erlaubt eine datengetriebene Verarbeitung, bei der Konfigurationsinformationen direkt aus den Daten extrahiert und automatisiert weiterverarbeitet werden. Trotz der hohen Praxisnähe und Konsistenz bleibt der Ansatz stark auf den jeweiligen Anwendungskontext zugeschnitten und adressiert keine domänenübergreifende Interoperabilität oder FAIR-konforme Nachnutzung über Projektgrenzen hinweg.

Zusammenfassend zeigen die betrachteten HDF5-basierten Abfragesysteme und Erweiterungen, dass auf daten- und formatnaher Ebene leistungsfähige Lösungen für effiziente Speicherung, Abfrage und projektinterne Wiederverwendung existieren. Die Integration semantisch eindeutiger Metadaten, die formale Validierung von Beschreibungen sowie die konsistente Einbettung in FAIR-orientierte Infrastrukturen werden jedoch entweder nicht adressiert oder nur in isolierten Teilaspekten berücksichtigt.

2.3.3 Datenbanken in der Strömungsmechanik

Die experimentelle und numerische Untersuchung strömungsmechanischer Systeme erfordert den Zugriff auf verlässliche Referenzdaten, insbesondere für die Validierung numerischer Simulationsmethoden. Vor diesem Hintergrund wird im Folgenden ein Überblick über bestehende

Datenbankkonzepte und Publikationsmodelle gegeben, die für das Forschungsdatenmanagement in der Strömungsmechanik und angrenzenden Disziplinen relevant sind.

Grundsätzlich lassen sich vier etablierte Ansätze zur Veröffentlichung wissenschaftlicher Forschungsdaten unterscheiden: die Bereitstellung begleitend zu wissenschaftlichen Publikationen, die Ablage in institutionellen Repositorien, die Nutzung disziplinärer Fachdatenbanken sowie die Veröffentlichung über allgemeine, disziplinübergreifende Repositorien. Diese Ansätze spannen ein Spektrum von stark spezialisierten bis hin zu breit zugänglichen Lösungen auf, die jeweils unterschiedliche Anforderungen an Auffindbarkeit, Zugänglichkeit, Standardisierung und fachliche Tiefe erfüllen. Die Wahl des Publikationswegs wird dabei nicht nur durch wissenschaftliche Zielsetzungen bestimmt, sondern auch durch technische Rahmenbedingungen, institutionelle Unterstützung und verfügbare Expertise.

Die Veröffentlichung von Forschungsdaten in direkter Verbindung zu wissenschaftlichen Artikeln ist weit verbreitet und wird zunehmend von Fachzeitschriften gefordert. Der wesentliche Vorteil dieses Ansatzes liegt in der engen Kopplung von Publikation und Datensatz sowie in der möglichen Qualitätssicherung durch den Peer-Review-Prozess. Gleichzeitig sind solche Datensätze häufig in Umfang, Struktur und Metadatenbeschreibung stark eingeschränkt und nicht immer für eine langfristige Archivierung oder systematische Nachnutzung ausgelegt.

Institutionelle Repositorien, etwa *DaRUS* (Universität Stuttgart) oder *KITopen* (Karlsruher Institut für Technologie), bieten eine niederschwellige Möglichkeit zur Veröffentlichung von Forschungsdaten im lokalen institutionellen Kontext. Sie sind u. U. in bestehende Arbeitsabläufe integriert, weisen jedoch Unterschiede hinsichtlich Standardisierung, Interoperabilität und überregionaler Sichtbarkeit auf.

Disziplinäre Repositorien und Fachdatenbanken adressieren gezielt die Anforderungen einzelner wissenschaftlicher Communities und fördern durch standardisierte Datenformate und Metadatenmodelle die Nachnutzbarkeit der Daten. Für die Strömungsmechanik stellt die *ERCOFTAC Database* ein etabliertes Beispiel dar, das seit Mitte der 1990er Jahre experimentelle und numerische Referenzdatensätze für die Validierung von CFD-Methoden bereitstellt. Solche Datenbanken erhöhen die fachliche Vergleichbarkeit, sind jedoch häufig auf spezifische Datentypen beschränkt und bieten nur begrenzte Möglichkeiten zur maschinellen Weiterverarbeitung.

Allgemeine, disziplinübergreifende Repositorien wie *Zenodo* oder *Figshare* zeichnen sich durch hohe Interoperabilität, persistente Identifikatoren und eine klare Orientierung an den FAIR-Prinzipien aus. Durch DOI-Vergabe, standardisierte Metadaten und Indexierung in Suchmaschinen wird eine langfristige Auffindbarkeit und Zitierfähigkeit gewährleistet. Gleichzeitig sind diese Plattformen bewusst generisch ausgelegt und bieten nur eingeschränkte Unterstützung für domänenspezifische Datenstrukturen oder komplexe Abfrageanforderungen.

Ein Großteil der für die Validierung numerischer Strömungssimulationen relevanten Datenbanken wird zentral von wissenschaftlichen Einrichtungen betrieben. Neben *ERCOFTAC* zählen hierzu unter anderem die *John Hopkins Turbulence Database (JHTDB)* sowie institutionelle DNS-Datenbanken, etwa die *Turbulence DNS Database* der TU Darmstadt[4]. Während viele dieser Ressourcen klassische Download-basierte Zugriffe bereitstellen, verfolgt die JHTDB einen

[4] https://www.fdy.tu-darmstadt.de/fdyresearch/dns/direkte_numerische_simulation.en.jsp

weiterentwickelten Ansatz, bei dem umfangreiche DNS-Datensätze über eine Programmierschnittstelle gezielt abgefragt werden können. Dies erleichtert den Umgang mit sehr großen Datenmengen und unterstützt die direkte Integration in numerische Auswertungs- und Simulationsworkflows.

Im Vergleich zu diesen etablierten Lösungen in angrenzenden Disziplinen steht die Entwicklung FAIR-konformer, frei zugänglicher Datenbanken für spezifische Anwendungen der Strömungsmechanik noch am Anfang. Insbesondere fehlen bislang integrierte Ansätze, die experimentelle und numerische Referenzdaten in strukturierter Form bereitstellen und zugleich eine maschinenlesbare, interoperable und langfristig nachnutzbare Beschreibung ermöglichen.

Ein wesentlicher Grund für die begrenzte Verfügbarkeit FAIR-konformer, frei zugänglicher CFD-Datenbanken liegt in den spezifischen Herausforderungen des Datenschutzes und der Industrie-Konfidentialität. Anders als in Disziplinen wie den Materialwissenschaften oder der Bioinformatik, in denen umfangreiche öffentliche Datenbestände etabliert sind, werden strömungsmechanische Simulationsergebnisse in vielen Fällen als geschäftssensitiv betrachtet. Insbesondere in der Industrie – etwa bei der Optimierung von Turbinen, Pumpen, Fahrzeugkomponenten oder Heiz-, Lüftungs- und Klimaanlagen – sind CFD-Simulationsdaten oft an proprietäre Fertigungsprozesse, Konstruktionsdetails oder Leistungscharakteristiken gekoppelt und unterliegen damit dem Schutz von Geschäftsgeheimnissen.

Diese Vertraulichkeitsbeschränkungen erstrecken sich häufig auch auf wissenschaftliche Kooperationen zwischen akademischen Einrichtungen und Industrie. Selbst wenn in solchen Projekten hochwertige Daten entstehen, können diese aufgrund von Geheimhaltungsvereinbarungen nicht öffentlich gemacht werden. Dies führt zu einer systematischen Lücke in den öffentlich verfügbaren Datenressourcen: Während akademische CFD-Datenbanken wie ERCOFTAC oder JHTDB wichtige Benchmarks für Validierungszwecke bieten, decken sie typischerweise nur eine begrenzte Palette von Anwendungsfällen ab (meist Turbulenz, einfache Geometrien, grundlegende Strömungsregimes). Komplexe, anwendungsnahe Szenarien mit technisch relevanten Geometrien und Betriebsbedingungen bleiben hingegen häufig privatisiert.

Hinzu kommt, dass viele kommerzielle CFD-Plattformen und Dienstleistungen (etwa cloudbasierte Simulationslösungen) bewusst proprietäre Datenformate und geschlossene Ökosysteme nutzen, um ihre Geschäftsmodelle zu schützen. Diese Systeme bieten zwar integrierte Datenmanagement-Funktionen, widersprechen aber grundsätzlich den FAIR-Prinzipien zur offenen, maschinenlesbaren und nachnutzbaren Datenbereitstellung. Die Kombination aus wirtschaftlichen Schutzinteressen und technologischer Fragmentierung führt damit zu einer Situation, in der qualitativ hochwertige und praxisrelevante CFD-Daten der wissenschaftlichen Gemeinschaft vielfach nicht zur Verfügung stehen.

Diese strukturelle Herausforderung unterstreicht die Bedeutung von Initiativen, die gezielt auf akademischen oder unternehmerisch-akademischen Kooperationen basieren und dabei Strategien zur Datenfreigabe (etwa durch Anonymisierung, Aggregation oder Publikation nach Ablauf von Sperrfristen) in den Forschungsprozess integrieren. Sie verdeutlicht zugleich die Notwendigkeit, FAIR-konforme Dateninfrastrukturen in der Strömungsmechanik nicht als globale Zentrallösung zu konzipieren, sondern als modulares, dezentralisiertes Ökosystem, das

auch proprietäre und vertrauliche Datenbestände angemessen berücksichtigt und gleichzeitig offene, nachnutzbare Kernbestände kontinuierlich aufbaut.

3 Anforderungen und Herausforderungen an Datenmanagement in der Strömungsmechanik

Die Strömungsmechanik ist durch ein breites Anwendungsspektrum in zahlreichen naturwissenschaftlichen und technischen Disziplinen gekennzeichnet. Untersuchungen reichen von kleinskaligen Phänomenen in der Mikrofluidik über Anwendungen in der Energiewirtschaft bis hin zu großskaligen Prozessen in Geophysik und Astrophysik. Technisch besonders relevante Fragestellungen betreffen unter anderem die Energiewandlung, die Aerodynamik sowie Verbrennungsprozesse.

Die hohe Diversität der Problemstellungen und Teildisziplinen führt zu einer großen Variabilität und Individualität sowohl bei experimentellen Prüfständen als auch bei numerischen Konfigurationen. In der Regel müssen zahlreiche Parameter erfasst werden, wobei unterschiedlichste Messtechniken und Simulationsmethoden zum Einsatz kommen. Änderungen an Randbedingungen oder Forschungsfragen gehen häufig mit Umbauten und Anpassungen einher, deren sorgfältige Dokumentation essenziell für die Nachvollziehbarkeit und Auswertung ist (Heinrichs et al., 2021; Selent et al., 2019). Dies gilt in gleicher Weise für numerische Untersuchungen, bei denen Modellannahmen, Diskretisierungen, Randbedingungen und Lösungsmethoden den Ergebnisraum maßgeblich bestimmen.

Charakteristisch für die Strömungsmechanik ist damit nicht allein die Größe der entstehenden Datenmengen, sondern insbesondere deren strukturelle und semantische Heterogenität. Forschungsdaten umfassen neben hochdimensionalen Feldgrößen auch beschreibende Metadaten, Modellparameter, messtechnische Konfigurationen, Softwareartefakte sowie Auswertungs- und Validierungsschritte, die in engen inhaltlichen Abhängigkeiten zueinander stehen. Diese vielfach vernetzten Beziehungen zwischen Experimenten, Simulationen, Modellen und Verfahren, Messtechnik und Software, wie in Abbildung 3.1 schematisch dargestellt, stellen eine zentrale Herausforderung für ein konsistentes Forschungsdatenmanagement dar.

Gerade an dieser Stelle zeigen sich systematische Grenzen der FAIR-Prinzipien in ihrer bisherigen praktischen Umsetzung in der Strömungsmechanik. Diese werden überwiegend auf Ebene der Datenpublikation adressiert, nicht jedoch auf Ebene der internen Datenentstehung und -strukturierung. Wesentliche Kontextinformationen bleiben dadurch implizit oder ausschließlich in projektspezifischen Konventionen dokumentiert. Die Auffindbarkeit (Findability) ist in der Regel auf projektinterne Strukturen beschränkt, während die Interoperabilität (Interoperability) unter uneinheitlichen Bezeichnungen physikalischer Größen, fehlender semantischer Eindeutigkeit und stark variierenden Datenstrukturen leidet. Besonders kritisch ist die eingeschränkte Wiederverwendbarkeit (Reusability), da ohne formal erfasste Abhängigkeiten zwischen Daten, Modellen und Verfahren, Messtechnik und Software eine reproduzierbare Nutzung durch Dritte kaum möglich ist.

Vor diesem Hintergrund ist es nicht überraschend, dass sich bislang kein einheitliches, disziplinübergreifendes Konzept für das Datenmanagement in der Strömungsmechanik etabliert hat. Gleichwohl besteht Potenzial, die Vielzahl experimenteller und numerischer Methoden auf eine überschaubare Anzahl von Standardansätzen zurückzuführen. Die Identifikation von Gemeinsamkeiten eröffnet die Möglichkeit, Synergieeffekte zwischen Fachbereichen zu nutzen

und durch die Etablierung gemeinsamer Standards die Austauschbarkeit von Daten sowie die Effizienz der Verarbeitung zu verbessern.

In den folgenden Kapiteln werden daher existierende Standards in der Strömungsmechanik, deren Anforderungen sowie geeignete Dateiformate analysiert. Ziel ist es, die besonderen Herausforderungen des Forschungsdatenmanagements in diesem Bereich zu systematisieren und Ansatzpunkte für eine nachhaltige, FAIR-konforme Dateninfrastruktur zu identifizieren.

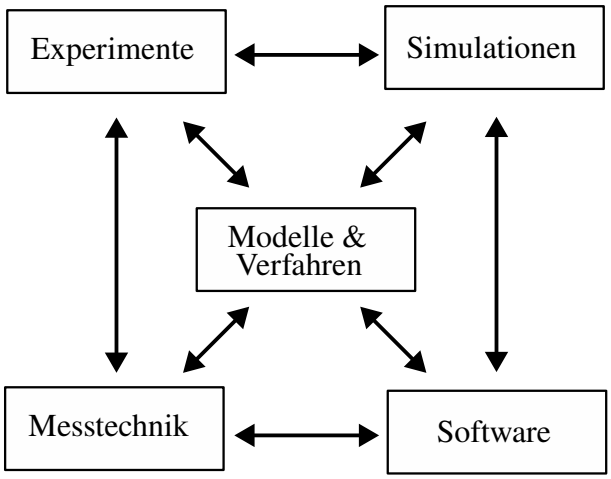


Abbildung 3.1: Vielfältige Abhängigkeiten und Wechselwirkungen zwischen Experimenten, Simulationen, Modellen und Verfahren, Messtechnik sowie Software in der Strömungsmechanik. Aus diesen strukturellen Zusammenhängen und den daraus resultierenden heterogenen Daten ergeben sich die zentralen Herausforderungen für das Forschungsdatenmanagement.

3.1 Standards in der Strömungsmechanik

In der Strömungsmechanik existiert bislang kein einheitlicher Standard, weder für Metadaten und die Benennung von Strömungsgrößen noch für Dateiformate. Während in der numerischen Strömungsmechanik proprietäre Datenstrukturen dominieren, werden in der experimentellen Forschung überwiegend textbasierte Formate genutzt.

Der Bedarf nach einheitlichen Standards für numerische Strömungssimulationen ist seit langem bekannt, insbesondere aufgrund der weit verbreiteten Nutzung geschlossener Formate. Bereits in den 1990er-Jahren wurde daher das *CFD General Notation System* (CGNS) entwickelt, eine standardisierte Datenstruktur und Softwarebibliothek, die aus einer Kooperation zwischen der NASA, der Luftfahrtindustrie und Softwarefirmen hervorging (Legensky et al., 2002). CGNS ermöglicht den Austausch von Simulationsdaten zwischen kommerziellen und offenen CFD-Programmen. Ursprünglich auf dem Advanced Data Format (ADF) basierend, wurde das System später auf HDF5 umgestellt, was einem allgemeinen Trend in der wissenschaftlichen Datenverarbeitung entspricht (Rumsey et al., 2012). Eine CGNS-Kompatibilität setzt voraus, dass entsprechende Software den Standard implementiert. Interoperabilität zwischen unterschiedlichen CFD-Codes ist damit gegeben, eine direkte Einbindung experimenteller Daten bleibt jedoch aus. Dies ist auch nicht zu erwarten, da CGNS keine Technologien des Semantic Webs nutzt und somit nicht im allgemeinen Sinne maschinenverarbeitbar ist.

Experimentelle Messaufbauten bestehen meist aus einer Vielzahl heterogener Komponenten, deren Verwaltung und Dokumentation in der Verantwortung der Wissenschaftler liegt. Häufig kommt dabei kommerzielle Software wie *LabVIEW* von National Instruments zur Anwendung. Da es sich um eine frei konfigurierbare Entwicklungsumgebung handelt, kann das Datenformat individuell angepasst werden. In der Praxis werden Messdaten daher überwiegend in tabellarischen ASCII-Textdateien abgelegt. Die Annotation der Metadaten obliegt den Forschern selbst, wobei Bestrebungen zur teilweisen Automatisierung erkennbar sind (vgl. (Selent et al., 2019)).

Vor diesem Hintergrund stellt sich die grundsätzliche Frage, ob und in welcher Form eine Standardisierung experimenteller Strömungsdaten sinnvoll ist. Besonders naheliegend wäre ein einheitlicher Standard für die weit verbreitete Particle Image Velocimetry (PIV). Gerade für etablierte Messtechniken besteht erhebliches Potenzial, durch standardisierte Dateiformate und Benennungen die Vergleichbarkeit und Austauschbarkeit von Daten zu verbessern. Dies ist besonders relevant, da PIV-Daten häufig als Validierungsgrundlage numerischer Simulationen dienen und präzise Metadaten daher eine wesentliche Voraussetzung für eine transparente und belastbare Nutzung darstellen.

Bereits 1994 wurde der Bedarf an einer Standardisierung für den effizienten Austausch von PIV-Daten erkannt (Willert, 2004). Als Dateiformat wurde zunächst netCDF vorgeschlagen, später von Willert (2004) für PIV-Daten weiterentwickelt und reimplementiert. Das Schema umfasst verpflichtende Attribute zur allgemeinen Beschreibung eines PIV-Datensatzes sowie zur Definition der Metadaten einzelner Variablen. Trotz dieser frühen Bemühungen hat sich der Vorschlag nicht flächendeckend durchgesetzt; eine Ausnahme bildet die Software *PIVview*, die jedoch abgewandelte Benennungen nutzt.

Neben fachspezifischen Ansätzen wie für PIV besteht auch in der Strömungsmechanik insgesamt Bedarf an allgemeinen Metadatenschemata, die experimentelle und numerische Daten beschreibbar, vergleichbar und langfristig nutzbar machen. Hier setzen Modelle wie *EngMeta* (Schembera und Iglezakis, 2020; Selent et al., 2019) oder *Metadata4Ing* an, die speziell für ingenieurwissenschaftliche Experimente und Simulationen entwickelt wurden und etablierte Metastandards integrieren (vgl. Unterabschnitt 2.3.1).

Die Anwendung solcher Schemata und Ontologien eröffnet die Möglichkeit, strömungsmechanische Forschungsdaten besser zu strukturieren, interoperabel bereitzustellen und nachhaltig verfügbar zu machen. Damit leisten sie nicht nur einen entscheidenden Beitrag zur Erfüllung der FAIR-Prinzipien, sondern verbessern zugleich die Vergleichbarkeit und den Austausch zwischen verschiedenen Teildisziplinen der Strömungsmechanik.

3.2 Anforderungen an numerische und experimentelle Datensätze

Experimentelle Datensätze

Experimentelle Daten in der Strömungsmechanik sind in der Regel mehrdimensional. Relativ geringe Datenmengen entstehen, wenn ausschließlich skalare Größen wie Temperatur oder Druck als Funktion der Zeit erfasst werden. Solche Datensätze bewegen sich, abhängig von Messdauer und zeitlicher Auflösung, meist im Bereich von Megabyte bis wenigen Gigabyte und stellen aufgrund ihrer eindimensionalen, tabellarischen Struktur nur geringe Anforderungen an das Dateiformat.

Mit jeder zusätzlichen Dimension steigen sowohl der Speicherbedarf als auch die Komplexität der Datenorganisation. Häufig eingesetzte optische Messtechniken wie Computertomographie oder PIV erzeugen zwei-, drei- oder sogar vierdimensionale Datensätze. Bei PIV etwa werden Geschwindigkeitsfelder in einer Ebene (2D) oder in einem Volumen (3D) bestimmt, häufig ergänzt durch zeitaufgelöste Messungen (4D). Hinzu kommt, dass neben Rohbildern oft auch Referenz- oder Kalibrierungsaufnahmen sowie rekonstruierte Daten gespeichert werden müssen, was den Speicherbedarf erheblich erhöht. Typische Datensätze reichen daher von wenigen bis hin zu mehreren hundert Gigabyte (M. Schneider et al., 2020). Vor diesem Hintergrund sind kompakte Datenstrukturen und effiziente Komprimierung zentrale Anforderungen (Kompenhans, 2000).

Neben den Messdaten selbst müssen die Parameter der eingesetzten Systeme erfasst werden. Diese sind sowohl für die Verarbeitung durch Analyse- und Visualisierungssoftware erforderlich als auch für die Reproduzierbarkeit der Experimente unverzichtbar. Da viele Hard- und Softwaresysteme bereits Ausgabedateien wie Log-Dateien erzeugen, ist eine automatisierte Metadatenextraktion naheliegend. Kompenhans (2000) betonen in diesem Zusammenhang die Bedeutung frei verfügbarer und leicht nutzbarer Programmierschnittstellen, um die Weiterverwendung der Daten überhaupt zu ermöglichen.

Die automatisierte Erfassung sollte jedoch nicht auf technische Parameter beschränkt bleiben. Auch Informationen zu beteiligten Personen, Projekten, Institutionen oder Erfassungszeitpunkten sind essenziell. Sie bilden die Grundlage für Reproduzierbarkeit und erleichtern zugleich die Durchsuchbarkeit der Datensätze auf oberster Ebene. Eine besondere Herausforderung liegt in der Heterogenität experimenteller Systeme, da Messkomponenten oft von unterschiedlichen Herstellern stammen und jeweils eigene Datenformate und Schnittstellen verwenden. Nur in wenigen Bereichen, etwa der medizinischen und biomedizinischen Forschung, gelingt es, durch enge Kooperationen zwischen Herstellern und Forschungseinrichtungen standardisierte Zielformate bereitzustellen (Millecam et al., 2021).

Da experimentelle Aufbauten und deren numerische Analysen in der Strömungsmechanik meist volatil und heterogen sind, muss ein geeignetes Dateiformat flexibel genug sein, um unterschiedlichste Informationsarten zu erfassen. Voraussetzung hierfür ist eine präzise Definition, *welche* Informationen in welcher Form gespeichert werden müssen.

Numerische Datensätze

Die Herausforderungen numerischer Untersuchungen ähneln in vielerlei Hinsicht denen experimenteller Messungen, unterscheiden sich jedoch insbesondere im Umfang und in der Komplexität der erzeugten Daten. Während experimentelle Datensätze oft durch eine begrenzte räumliche und zeitliche Auflösung limitiert sind, erlauben numerische Simulationen eine erheblich höhere Detailtiefe. Direkte Numerische Simulationen (DNS) oder großskalige Large-Eddy-Simulationen (LES) erzeugen nicht nur hochdimensionale, sondern auch extrem umfangreiche Datensätze. Hinzu kommt, dass Randbedingungen, Geometrien und andere Parameter vergleichsweise einfach und schnell variiert werden können. In Kombination mit den stetigen Fortschritten in der Rechenleistung führte dies in den vergangenen Jahrzehnten zu einem explosionsartigen Wachstum der Datenmengen, die heute Größenordnungen von bis zu 10 Terabyte erreichen können (Selent et al., 2019).

Neben den technischen Herausforderungen ergeben sich Anforderungen an die Verwaltung der Daten, insbesondere im Hinblick auf Durchsuchbarkeit, Wiederverwendbarkeit und Interpretierbarkeit. Die Annotation mit aussagekräftigen Metadaten ist hierbei entscheidend. Wie bei experimentellen Untersuchungen ist eine automatisierte Metadatenextraktion wünschenswert, wird jedoch häufig durch die Nutzung proprietärer Software erschwert. Die korrekte Interpretation numerischer Ergebnisse setzt zudem ein tiefgehendes Verständnis der zugrunde liegenden Modellannahmen, Randbedingungen und Diskretisierungsmethoden voraus. Ohne eine einheitliche und maschinenlesbare Beschreibung dieser Metadaten ist es kaum möglich, Simulationsergebnisse objektiv zu vergleichen, mit anderen Datenquellen zu verknüpfen oder für weiterführende Studien zu nutzen.

Ein weiteres Problemfeld betrifft die Langzeitarchivierung. Die enorme Größe numerischer Datensätze sowie die kontinuierliche Weiterentwicklung von Simulationssoftware bergen das Risiko, dass ältere Ergebnisse aufgrund inkompatibler Formate oder veralteter Softwareversionen nicht mehr reproduzierbar sind. Eine nachhaltige Speicherung erfordert daher den Einsatz

offener, standardisierter Datenformate sowie eine umfassende Dokumentation der Simulationsparameter.

Auch die Analyse numerischer Ergebnisse stellt erhebliche Anforderungen. Während experimentelle Daten meist in Form von Zeitreihen oder Punktmessungen vorliegen, liefern numerische Simulationen vollständige[1] dreidimensionale Felder, die zusätzlich zeitabhängig sein können. Die Extraktion relevanter Informationen aus diesen hochdimensionalen Datensätzen erfordert spezialisierte Werkzeuge zur Datenreduktion, Merkmalsextraktion und Visualisierung. Mit der zunehmenden Bedeutung datengetriebener Methoden wie des maschinellen Lernens wächst daher der Bedarf an strukturierten, gut dokumentierten und interoperablen Simulationsdaten, die sowohl für klassische Analysen als auch für innovative Auswertungsverfahren geeignet sind.

3.3 Evaluation ausgewählter wissenschaftlicher Dateiformate

Die Wahl eines geeigneten Dateiformats stellt eine zentrale Herausforderung im wissenschaftlichen Datenmanagement dar. Insbesondere in der Strömungsmechanik, wo sowohl numerische Simulationen als auch experimentelle Messungen regelmäßig mehrdimensionale und datenintensive Ergebnisse erzeugen, muss ein Format die effiziente Speicherung, Verarbeitung und den Austausch der Daten gewährleisten. Dabei treten mehrere Problemfelder auf:

- **Datenvolumen:** Hochaufgelöste Simulationen (z. B. DNS, LES) oder Messverfahren wie die Particle Image Velocimetry (PIV) können mehrere Terabyte an Daten erzeugen.
- **Vielfalt an Formaten:** Zahlreiche, teils proprietäre Formate existieren nebeneinander, oft zugeschnitten auf spezifische Softwarelösungen und ohne standardisierte Schnittstellen.
- **Interoperabilität:** Der Datenaustausch zwischen Forschungsgruppen oder Programmen ist häufig nur über verlustbehaftete Konvertierungen möglich.

Während relationale Datenbanken wie SQL in der kommerziellen IT dominieren (Gray et al., 2005), haben sich in der Wissenschaft andere Lösungen etabliert. SQL-Datenbanken sind für unstrukturierte oder hochdimensionale Daten nur eingeschränkt geeignet. Daher greifen Forscher bevorzugt auf spezialisierte Dateiformate zurück, die für die Anforderungen naturwissenschaftlicher Daten entwickelt wurden.

Proprietäre Formate sind dabei besonders problematisch: Sie erschweren nicht nur den langfristigen Zugriff, sondern können mit dem Ende des Software-Supports unbrauchbar werden. Auch Übergangslösungen wie einfache Textformate (z. B. CSV) bergen Risiken, da sie keine Möglichkeit bieten, komplexe Strukturen oder Metadaten adäquat zu integrieren (Srivastava et al., 2020).

CSV ist als tabellarisches Textformat weit verbreitet, da es leicht lesbar, portabel und kompatibel ist. Für wissenschaftliche Zwecke stößt es jedoch rasch an Grenzen: mehrdimensionale oder hierarchische Strukturen lassen sich nicht darstellen, Metadaten fehlen, und große Datenmengen führen zu erheblichem Speicher- und Rechenaufwand.

[1] Mit vollständig ist hier die gleichzeitige Erfassung mehrerer physikalischer Größen gemeint, wie beispielsweise Temperatur, Geschwindigkeit und Druck, was mit experimentellen Methoden nicht möglich ist.

JSON (JavaScript Object Notation) erlaubt eine hierarchische Strukturierung und ist insbesondere für Metadaten und kleinere Datensätze geeignet. Für umfangreiche numerische n-dimensionale Datensätze ist es allerdings ineffizient, weshalb binäre Formate vorzuziehen sind.

NetCDF (Network Common Data Form) ist ein selbstbeschreibendes, binäres Format, das speziell für wissenschaftliche Anwendungen entwickelt wurde. Die aktuelle Version netCDF4 basiert auf HDF5 und kombiniert effiziente Speicherung großer mehrdimensionaler Datensätze mit leistungsfähigen Funktionen wie „Chunking“ und Komprimierung. Besonders verbreitet ist NetCDF in der Klimaforschung, wo es in Verbindung mit den *Climate and Forecast Conventions* (CF) einen De-facto-Standard darstellt.

CGNS (CFD General Notation System), entwickelt von NASA und Boeing, richtet sich explizit an die numerische Strömungsmechanik. Es basiert auf HDF5 und bietet standardisierte Strukturen für strukturierte und unstrukturierte Gitter. Schnittstellen existieren zu gängigen CFD-Programmen wie ANSYS Fluent, OpenFOAM und STAR-CCM+. Die Spezialisierung auf CFD-Daten macht CGNS jedoch weniger flexibel für experimentelle oder interdisziplinäre Anwendungen.

FITS (Flexible Image Transport System) ist vor allem in der Astronomie verbreitet. Es eignet sich für multidimensionale Daten und unterstützt umfangreiche Metadatenintegration, bleibt aber stark auf bildbasierte Daten ausgelegt und ist daher für CFD oder strömungsmechanische Experimente nur begrenzt nutzbar.

HDF5 (Hierarchical Data Format 5) stellt einen der universellsten und leistungsfähigsten Ansätze dar. Es unterstützt sehr große Datenmengen, ermöglicht effiziente Komprimierung und organisiert Daten hierarchisch in einer Baumstruktur. Als selbstbeschreibendes Format erlaubt es, Metadaten unmittelbar mit den Daten zu speichern. Dank der breiten Unterstützung durch Programmiersprachen und Software-Ökosysteme bildet HDF5 die Grundlage für spezialisierte Formate wie netCDF4, CGNS oder NeXus. Seine Flexibilität macht es sowohl für numerische als auch experimentelle Daten geeignet und prädestiniert es für interdisziplinäre Anwendungen und die Langzeitarchivierung.

Weitere Formate wie das Core Scientific Dataset Model (CSD) (Srivastava et al., 2020), das Advanced Scientific Data Format (ASDF) oder NeXus greifen ähnliche Konzepte auf, sind jedoch stärker domänenspezifisch ausgerichtet. Während ASDF auf FITS basiert und für astronomische Anwendungen optimiert wurde, baut NeXus auf HDF5 auf und adressiert insbesondere die Neutronen- und Synchrotronforschung. CSD nutzt JSON-Strukturen und versucht, Interoperabilität durch semantische Beschreibungen zu fördern.

Eine vergleichende Übersicht der hier diskutierten Formate ist in Tabelle 3.1 dargestellt.

Zusammenfassend lassen sich aus den dargestellten Herausforderungen, bestehenden Standards und Dateiformaten zentrale Anforderungen an ein nachhaltiges Forschungsdatenmanagement in der Strömungsmechanik ableiten. Ein geeignetes System muss sowohl hochdimensionale experimentelle und numerische Primärdaten effizient speichern als auch deren semantischen Kontext maschinenlesbar erfassen. Es muss flexible, hierarchische Datenstrukturen unterstützen, ohne projektspezifische Konventionen zu erzwingen, und zugleich eine formale Beschreibung von

Provenienz, Modellannahmen und Softwareabhängigkeiten ermöglichen. Bestehende Standards und Dateiformate adressieren jeweils Teilaspekte dieser Anforderungen, bieten jedoch kein integriertes Gesamtkonzept. Daraus ergibt sich der Bedarf nach einer methodischen Lösung, die strukturierte Datenhaltung, semantische Beschreibung und FAIR-Operationalisierung konsistent zusammenführt. Die Entwicklung eines entsprechenden Konzepts bildet das Ziel der vorliegenden Arbeit und wird in den folgenden Kapiteln schrittweise ausgearbeitet.

Format	Einsatzgebiet	Vorteile	Nachteile
CSV	Kleine, tabellarische Datensätze	Weit verbreitet, menschenlesbar, einfache Verarbeitung	Keine Metadatenunterstützung, ineffizient für große Datenmengen, keine Hierarchien oder Mehrdimensionalität
JSON	Metadaten, kleine bis mittlere Datensätze	Flexibel, menschen- und maschinenlesbar, hierarchische Struktur	Für große numerische n-dimensionale Datensätze ineffizient, keine native Binärspeicherung
XML	Dokumenten- und Metadatenmodellierung	Selbstbeschreibend, etabliert, unterstützt komplexe Strukturen	Hoher Overhead, wenig effizient für große wissenschaftliche Datenmengen
netCDF4	Klimaforschung, Erdbeobachtung, Ozeanografie	Effiziente Speicherung gitterbasierter Daten, integrierte Metadaten, HDF5-basiert	Primär für Rasterdaten optimiert, eingeschränkte Flexibilität bei unstrukturierten Daten
CGNS	Numerische Strömungsmechanik (CFD)	Standardisiertes CFD-Format, HDF5-basiert, gute Performance für Simulationsdaten	Komplexe Implementierung, kaum verbreitet außerhalb der CFD-Gemeinschaft
FITS	Astronomie, Bild- und Spektraldaten	Umfangreiche Metadatenunterstützung, geeignet für multidimensionale Daten	Kaum Nutzung in Ingenieurwissenschaften, teilweise durch HDF5 ersetzt
HDF5	Breite natur- und ingenieurwissenschaftliche Anwendungen	Skalierbar, effiziente Speicherung, unterstützt Hierarchien und Metadaten, Grundlage für netCDF4, CGNS, NeXus	Höhere Komplexität als einfache Formate, benötigt spezielle Softwarebibliotheken
NeXus	Neutronen- und Synchrotron-Experimente	Standardisierte Metadatenstrukturen, HDF5-basiert, fördert Interoperabilität	Domänenspezifisch, erfordert Einarbeitung in NeXus-Standards
ASDF	Astronomie und Astrophysik	Erweiterbare, JSON-ähnliche Struktur, effizienter als JSON für numerische Daten, gute Metadatenintegration	Kaum verbreitet außerhalb der Astronomie, eingeschränkter Software-Support

Tabelle 3.1: Vergleich ausgewählter Dateiformate für wissenschaftliche Daten in den Natur- und Ingenieurwissenschaften.

4 Einführung eines FDM-Konzepts basierend auf HDF5

In den vorangegangenen Kapiteln wurden die Grundlagen des Forschungsdatenmanagements sowie die spezifischen Herausforderungen strömungsmechanischer Forschungsdaten analysiert. Insbesondere wurde gezeigt, dass bestehende FDM-Ansätze und FAIR-Umsetzungen häufig auf die Ebene der Datenpublikation beschränkt bleiben und die interne Strukturierung, Kontextualisierung und Validierung von Forschungsdaten unzureichend adressieren.

Vor diesem Hintergrund wird im Folgenden ein methodisches Konzept für ein nachhaltiges Forschungsdatenmanagement vorgestellt. Das Konzept versteht sich nicht als weiteres domänenspezifisches Datenformat oder als rein softwaretechnische Lösung, sondern als allgemeiner Rahmen zur operationalisierten Umsetzung der FAIR-Prinzipien für komplexe ingenieurwissenschaftliche Forschungsdaten. Der wissenschaftliche Beitrag liegt in der systematischen Trennung von strukturierter Primärdatenhaltung und semantischer Metadatenbeschreibung sowie in der Integration von Validierungs- und Automatisierungsmechanismen entlang des gesamten Datenlebenszyklus.

Das Konzept wurde ausgehend von einem strömungsmechanischen Anwendungsfall entwickelt, ist jedoch bewusst auf eine möglichst breite, disziplinübergreifende Anwendbarkeit ausgelegt. Die Speicherung primärer und verarbeiteter Daten erfolgt konsequent in HDF5, während Technologien des Semantic Webs zur formalen, maschinenlesbaren Beschreibung von Metadaten eingesetzt werden. Auf diese Weise wird FAIR nicht nur auf Ebene der Datenpublikation, sondern bereits während der Datenentstehung und -verarbeitung technisch verankert.

Dieses Kapitel formuliert zunächst die Zielsetzung des Konzepts, beschreibt anschließend die zugrunde liegenden Methodiken der Daten- und Metadatenmodellierung und schließt mit der Integration des Ansatzes in wissenschaftliche Arbeitsabläufe.

4.1 Ziele und Entwurfsprinzipien

Das vorgestellte Forschungsdatenmanagementkonzept verfolgt das Ziel, Forscher über den gesamten Datenlebenszyklus hinweg zu unterstützen. Dabei geht es nicht nur um konzeptionelle Lösungen, wie die Entwicklung von Standards und Methoden, sondern auch um die Bereitstellung konkreter softwaretechnischer Werkzeuge, die Forscher bei der Interaktion mit Daten und Metadaten unmittelbar unterstützen.

Der ganzheitliche Ansatz umfasst alle Phasen des Datenlebenszyklus von der Planung über die Verarbeitung bis hin zur Veröffentlichung und Wiederverwendung von Forschungsdaten. Ziel ist es, sowohl die maschinelle Verarbeitung und Automatisierung als auch die inhaltliche wissenschaftliche Arbeit zu unterstützen, indem strukturierte Datenhaltung und formal beschriebene Metadaten miteinander verknüpft werden. Die Nutzung etablierter, bewährter Technologien trägt dabei wesentlich zur Akzeptanz des Konzepts bei und ermöglicht eine effiziente Integration in bestehende wissenschaftliche Arbeitsabläufe.

Die Zielsetzung des Konzepts leitet sich aus den Herausforderungen eines nachhaltigen, interoperablen und qualitätsgesicherten Forschungsdatenmanagements ab, wie sie in den Kapiteln 2 und 3 aus unterschiedlichen Perspektiven analysiert wurden. Die folgenden Punkte sind dabei nicht als projektspezifische Anforderungen, sondern als zentrale Entwurfsprinzipien des Konzepts zu verstehen:

1. **FAIR-Konformität**: Die FAIR-Prinzipien bilden die leitende Entwurfsgrundlage des Konzepts und werden technisch so umgesetzt, dass sie bereits während der Datenentstehung und -verarbeitung wirksam sind.
2. **Universelle Anwendbarkeit**: Das Konzept ist domänenübergreifend ausgelegt und nicht an eine spezifische Fachdisziplin gebunden.
3. **Automatisierbarkeit**: Die Annotation, Verarbeitung und Interpretation von Metadaten soll weitgehend maschinell unterstützt oder automatisiert erfolgen.
4. **Prüfbarkeit und Qualitätssicherung**: Datenstrukturen und Metadaten sollen überprüfbar sein und die Einhaltung projektspezifischer Konventionen und Qualitätsanforderungen gewährleisten.
5. **Offenheit**: Sowohl das konzeptionelle Modell als auch die begleitenden Softwarekomponenten sollen offen zugänglich, erweiterbar und mit bestehenden Standards kompatibel sein.

Gerade die Anforderung der universellen Anwendbarkeit steht häufig im Spannungsfeld zu den individuellen Bedürfnissen einzelner Forschungsbereiche. Während andere Ansätze wie NeXus (Klosowski et al., 1997) oder die Arbeiten von Preuß und Pelz (2018) und Preuss et al. (2018) eine starre Dateiorganisation vorgeben, verfolgt die vorliegende Arbeit bewusst ein flexibleres Konzept. Die Universalität wird dabei nicht durch eine festgelegte Datenstruktur erreicht, sondern durch die Trennung generischer Strukturprinzipien von domänenspezifischer Semantik, sodass fachliche Besonderheiten berücksichtigt werden können, ohne die Vorteile einer standardisierten Organisation aufzugeben.

Im Folgenden werden die methodischen Grundlagen zur Umsetzung dieser Entwurfsprinzipien erläutert. Das entwickelte Konzept versteht sich als zukunftssicherer Rahmen, der sich an aktuellen wissenschaftspolitischen Leitlinien und etablierten Technologien orientiert und auf eine breite wissenschaftliche Anwendung ausgelegt ist.

4.2 Methodische Umsetzung

Die Umsetzung des Konzepts folgt einer methodischen Ableitung aus den in Kapitel 3 identifizierten Anforderungen. Die in Abschnitt 4.1 formulierten Ziele und Entwurfsprinzipien dienen dabei als übergeordnete Entwurfslogik für alle konzeptionellen und technischen Entscheidungen. FAIR wird nicht als abstrakter Zielzustand verstanden, sondern als technisch wirksame Leitlinie für die Modellierung von Datenstrukturen, Metadaten und Prozessen entlang des gesamten Datenlebenszyklus. Der Schwerpunkt liegt insbesondere auf der überprüfbaren Wiederverwendbarkeit von Daten, Software und Metadaten.

Die genannten Entwurfsprinzipien definieren die primären Zielsetzungen des Konzepts und bilden den Maßstab für dessen wissenschaftliche Bewertung. Ergänzend werden sekundäre methodische Qualitätskriterien berücksichtigt, die nicht als eigenständige Ziele zu verstehen sind, sondern die praktische Umsetzbarkeit, Akzeptanz und Nachhaltigkeit des Konzepts unterstützen. Im Folgenden werden beide Ebenen getrennt dargestellt.

Umsetzung der primären Entwurfsprinzipien

FAIR-Konformität Die FAIR-Prinzipien bilden die Basis des Konzepts und werden sowohl auf Daten als auch auf wissenschaftliche Software angewandt. Letzteres trägt den Forderungen nach einer stärkeren Berücksichtigung von Software als integralen Bestandteil moderner, „computerbasierter Wissenschaft“ Rechnung (Anzt et al., 2021). Für Softwareversionen wird auf gängige Versionierungssysteme wie Git zurückgegriffen. Da Plattformen wie GitHub oder GitLab jedoch keine dauerhafte Persistenz garantieren[1], ist die zusätzliche Veröffentlichung von Software in Repositorien wie Zenodo vorgesehen, wodurch Versionen mit einer DOI versehen werden können (Lamprecht et al., 2020).

Die semantische Annotation erfolgt über *CodeMeta* (vgl. Unterabschnitt 2.3.1, (Jones et al., 2023)), womit den Empfehlungen der NFDI gefolgt wird. Persistente Identifikatoren (PIDs) wie DOIs für Daten, ORCIDs für Personen oder ROR-IDs für Einrichtungen gewährleisten eine eindeutige Referenzierbarkeit. Diese Maßnahmen sind Voraussetzung dafür, dass Daten und Software langfristig auffindbar, zugänglich und wiederverwendbar bleiben.

Universelle Anwendbarkeit Das Konzept ist auf disziplinübergreifende Anwendbarkeit ausgelegt. Hierzu werden generische Ontologien wie *DCTERMS*, *PROV*, *DCAT* oder *FOAF* als Basisschicht eingesetzt, die in nahezu jedem wissenschaftlichen Kontext nutzbar sind. Gleichzeitig erlaubt die modulare Erweiterbarkeit durch domänenspezifische Vokabulare die flexible Anpassung an spezielle Anforderungen einzelner Disziplinen. Die HDF-Datenstruktur bleibt dabei inhaltlich neutral und wird erst durch semantische Annotationen inhaltlich angereichert, was eine hohe Allgemeingültigkeit sicherstellt.

Automatisierbarkeit Ein wesentlicher methodischer Anspruch liegt in der Automatisierbarkeit von Metadatenprozessen. Durch die eindeutige semantische Annotation von HDF-Inhalten[2] mittels IRIs können Maschinen die Bedeutung der Inhalte interpretieren und in automatisierten Workflows weiterverarbeiten. Dies eröffnet Möglichkeiten für Validierung, Transformation und Integration in größere Datenökosysteme. Die im Rahmen dieser Arbeit entwickelten Softwarebibliotheken bilden hierfür die technische Grundlage.

Prüfbarkeit und Qualitätssicherung Die Qualitätssicherung erfolgt durch den Einsatz formaler Ontologien, die sicherstellen, dass nur valide und erwartungskonforme Werte zur

[1] Das FAIR-Prinzip F1 verlangt langfristige, persistente Identifikatoren; GitHub/GitLab erfüllen dies nicht.
[2] Eine detaillierte Auseinandersetzung mit dem Datenmodell HDF5 erfolgt im anschließenden Kapitel

Beschreibung von Daten genutzt werden. Dadurch wird eine automatisierte Validierung von Metadaten möglich, was Konsistenz und Vergleichbarkeit zwischen Datensätzen stärkt. Ergänzend werden Mechanismen zur Versionskontrolle und Provenienzverfolgung über *PROV* eingebunden, sodass auch die Entstehungskontexte von Daten transparent nachvollziehbar bleiben.

Offenheit Das Konzept verfolgt konsequent das Prinzip der Offenheit. HDF5, RDF, Turtle sowie die eingesetzten Ontologien basieren auf offenen Standards. Darüber hinaus wird im Rahmen dieser Arbeit eine eigene Ontologie entwickelt, die in späteren Kapiteln detailliert vorgestellt wird. Sie dient der Standardisierung zentraler Begriffe und Beziehungen und wird offen dokumentiert und zur Nachnutzung bereitgestellt. Damit schafft das Konzept die Grundlage für eine gemeinschaftliche Weiterentwicklung und eine langfristige Etablierung innerhalb der wissenschaftlichen Community.

Ergänzende methodische Qualitätskriterien

Neben den primären Entwurfsprinzipien werden weitere Qualitätskriterien berücksichtigt, die die praktische Tragfähigkeit und Akzeptanz des Konzepts fördern.

Handhabbarkeit Ein besonderes Augenmerk liegt auf der Handhabbarkeit, also der praktischen Umsetzbarkeit des Konzepts für Forscher. Vollumfängliche Lösungen, die sämtliche Prozesse digital abbilden, dokumentieren und automatisieren, sind zwar langfristig erstrebenswert, aus heutiger Sicht jedoch kaum praktikabel. Ein tragfähiger Ansatz muss daher verständlich sein, klare Vorteile aufzeigen und einen spürbaren Effizienzgewinn bieten. Dazu tragen die Begrenzung auf wenige Systeme und Dateiformate ebenso bei wie die Bereitstellung geeigneter Hilfsmittel in Form von Softwarelösungen. Die Konzentration auf eine relevante Programmiersprache sowie eine umfassende Dokumentation mit Beispielen erhöhen die Zugänglichkeit zusätzlich.

Wiederverwendbarkeit Das Prinzip der Wiederverwendbarkeit betont die konsequente Nutzung bestehender Technologien und Standards, um Effizienz, Interoperabilität und Nachhaltigkeit sicherzustellen. Dazu zählen etwa die Anwendung von ISO 8601 für Zeitangaben, Semantic Versioning (Preston-Werner, 2025) für die eindeutige Kennzeichnung von Softwareversionen oder der Einsatz etablierter Technologien des Semantic Webs wie RDF und bestehender Ontologien. Die Wiederverwendung solcher Bausteine verhindert redundante Entwicklungen und erleichtert die langfristige Pflege und Erweiterung des Systems.

Modularität und Flexibilität Das Managementsystem setzt sich aus verschiedenen Komponenten zusammen, etwa der Verwaltung von Metadatenrichtlinien, der Organisation von Daten und Informationsflüssen oder der Benutzerinteraktion. Diese Teilbereiche sollten modular gestaltet sein, um Abhängigkeiten klar zu strukturieren, Weiterentwicklungen zu erleichtern und

spezifische Anpassungen zu ermöglichen. Modularität steigert zudem die Wiederverwendbarkeit einzelner Komponenten und erleichtert zukünftige Erweiterungen.

Datenreduktion Das Prinzip der Datenreduktion und -optimierung betrifft sowohl die Ebene der Forschungsdaten als auch die administrative Organisation. Auf Datenseite ist zu prüfen, welche Informationen tatsächlich gespeichert werden müssen und wo Komprimierung oder Reduktionsverfahren sinnvoll sind. Manche Daten können zu einem späteren Zeitpunkt aus Metadaten rekonstruiert werden, sodass die Speicherung verzichtbar ist. Auch auf organisatorischer Ebene trägt Reduktion zur Effizienz bei: Die Anzahl eingesetzter Softwarelösungen sollte möglichst gering gehalten werden, um notwendiges Fachwissen zu bündeln und Komplexität zu vermeiden. Gleiches gilt für Dateiformate: Im Rahmen des vorgestellten FDM wird bewusst auf ein einziges primäres Datenformat und ein Metadatenformat gesetzt.

Die Leitprinzipien entlang des Forschungsdatenlebenszyklus sind in Abbildung 4.1 schematisch dargestellt. Die Abbildung gliedert den Datenlebenszyklus in die Phasen Planen, Erfassen, Analysieren, Teilen und Nachnutzen und veranschaulicht, wie strukturierte Daten und formale Metadaten über alle Phasen hinweg integriert werden können. Jede Phase wird dabei durch zentrale methodische Leitbegriffe charakterisiert: In der Planungsphase stehen die Definition von Datenstrukturen und Metadaten als Grundlage aller nachfolgenden Schritte im Vordergrund. Die Phase des Erfassens ist durch die standardisierte Überführung heterogener Eingangsdaten in eine konsistente Datenrepräsentation gekennzeichnet. Während der Analysephase rückt insbesondere

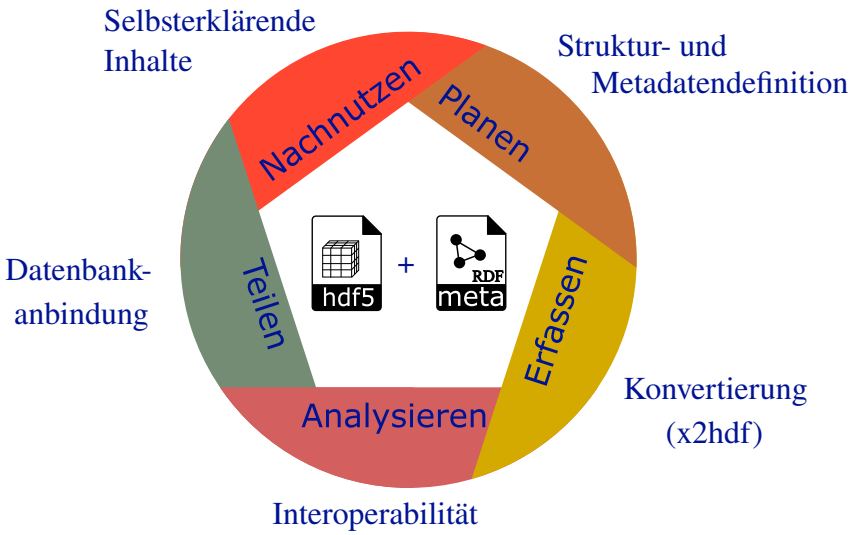


Abbildung 4.1: Der HDF-basierte Ansatz in Kombination mit einer Metadatenstrategie basierend auf RDF erlaubt die vollständige Integration in den Forschungsdatenzyklus. Dies reduziert einerseits die Komplexität und schafft andererseits ein hohes Maß an Nachvollziehbarkeit und Wiederverwendbarkeit.

die Interoperabilität der Daten und Metadaten in den Fokus, während das Teilen von Daten eine strukturierte Anbindung an Datenbanken und Repositorien erfordert. Die Phase der Nachnutzung schließlich stellt Anforderungen an selbstdokumentierende, maschinenlesbare Datenstrukturen, die eine eigenständige Interpretation und Wiederverwendung ermöglichen.

Bereits in dieser konzeptionellen Darstellung sind eine HDF5-basierte Datenrepräsentation sowie eine RDF-basierte Metadatenbeschreibung angedeutet, um die Trennung von strukturierter Datenhaltung und semantischer Beschreibung zu verdeutlichen. Die konkrete Begründung der Wahl von HDF5 als zentrales Datenformat sowie die detaillierte Ausgestaltung der Daten- und Metadatenmodelle erfolgen in den folgenden Abschnitten.

4.3 HDF5 als zentrales Dateiformat

Die Umsetzung eines nachhaltigen Forschungsdatenmanagements gemäß der gesetzten Ziele und eingeführten Prinzipien erfordert die Festlegung auf ein primäres Datenformat, das über den gesamten Forschungsdatenlebenszyklus hinweg konsistent eingesetzt werden kann. Insbesondere die Anforderungen an Handhabbarkeit, Reduktion technischer Komplexität, langfristige Nutzbarkeit und Interoperabilität lassen sich nur dann erfüllen, wenn strukturierte Forschungsdaten nicht in einer Vielzahl unterschiedlicher Formate vorliegen, sondern in einer einheitlichen, stabilen Repräsentation zusammengeführt werden.

Für diese Rolle kommen grundsätzlich verschiedene etablierte wissenschaftliche Datenformate in Betracht. Eine systematische Einordnung und Bewertung ausgewählter Formate wurde bereits in Abschnitt 3.3 vorgenommen. Wie dort gezeigt, adressieren viele dieser Formate jeweils spezifische Teilanforderungen, etwa den effizienten Austausch numerischer Felddaten oder die Speicherung domänenspezifischer Strukturen. Ein Format, das zugleich die flexible Abbildung komplexer Datenstrukturen, die Integration von Metadaten, eine breite Softwareunterstützung sowie eine langfristige Archivierung ermöglicht, ist jedoch nur in wenigen Fällen gegeben.

HDF5 erfüllt diese Anforderungen in ihrer Gesamtheit. Als generisches, binäres Containerformat ist es darauf ausgelegt, große, hochdimensionale und strukturell heterogene Datensätze effizient zu speichern und dabei deren logische Organisation explizit abzubilden. Zugleich erlaubt HDF5 die enge Kopplung von Daten und strukturellen Metadaten innerhalb einer einzelnen Datei und ist plattform- sowie programmiersprachenunabhängig einsetzbar. Damit stellt HDF5 das einzige Dateiformat dar, das alle im Rahmen dieser Arbeit formulierten Entwurfsprinzipien gleichzeitig und konsistent unterstützen kann.

Im vorgestellten Forschungsdatenmanagementkonzept wird HDF5 daher als einziges primäres Format für die Speicherung strukturierter Forschungsdaten verwendet. Andere Datenformate treten lediglich als Eingangs- oder Austauschformate auf, deren Inhalte in eine HDF5-basierte Repräsentation überführt werden. Die konkrete Ausgestaltung der Datenstrukturen innerhalb von HDF5 sowie deren Kombination mit einer separaten, semantischen Metadatenebene werden in den folgenden Kapiteln detailliert ausgearbeitet.

Eine HDF5-Datei ist intern hierarchisch organisiert und besteht aus drei grundlegenden Strukturelementen: Gruppen (*Groups*), Datensätzen (*Datasets*) und Attributen (*Attributes*), wie in

Abbildung 4.2 schematisch dargestellt. Diese Struktur ist funktional mit einem Dateisystem vergleichbar, wobei Gruppen Verzeichnissen und Datensätze Dateien entsprechen. Die oberste Ebene bildet die sogenannte Root-Gruppe, die als Einstiegspunkt der Datei dient und durch den Pfad `/` adressiert wird.

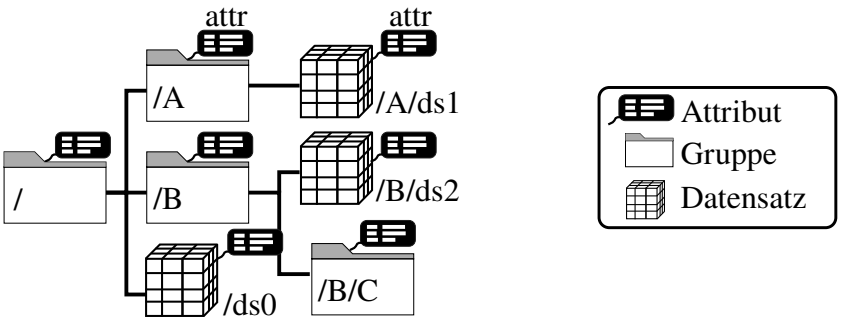


Abbildung 4.2: Interne Dateiorganisation von HDF: Gruppen (Groups) und multidimensionale Datensätze (Datasets) sind hierarchisch angeordnet. Jedes Objekt kann mit Attributen versehen werden, die als Schlüssel-Wert-Paare Metadaten darstellen.

Datensätze repräsentieren die eigentlichen Nutzdaten und können mehrdimensionale Arrays beliebiger Datentypen enthalten. Gruppen dienen der logischen Strukturierung dieser Datensätze und erlauben eine hierarchische Organisation komplexer Datenbestände. Sowohl Gruppen als auch Datensätze können mit Attributen versehen werden, die als Schlüssel-Wert-Paare zusätzliche beschreibende Informationen aufnehmen. Attribute sind integraler Bestandteil des HDF5-Datenmodells und ermöglichen es, strukturbezogene Metadaten direkt an die jeweiligen Datenobjekte zu koppeln.

Durch diese Kombination aus hierarchischer Organisation, klar definierten Datenobjekten und integrierten Attributen stellt HDF5 ein selbstbeschreibendes Datenformat dar. Die Bedeutung und Einordnung der gespeicherten Daten lassen sich somit grundsätzlich aus der Dateistruktur selbst ableiten, ohne auf externe Beschreibungen angewiesen zu sein. Diese Eigenschaft bildet eine wesentliche Grundlage für die im nächsten Kapitel entwickelte Daten- und Metadatenmodellierung sowie für die Trennung von struktureller und semantischer Beschreibung.

4.4 Datenmodellierung basierend auf HDF5 und RDF

Wie im vorhergehenden Kapitel gezeigt, eignet sich HDF5 als zentrales Dateiformat zur Speicherung strukturierter wissenschaftlicher Daten. Offen bleibt jedoch die Frage, wie innerhalb dieses generischen Containerformats eine konsistente, nachvollziehbare und wiederverwendbare Organisation von Daten und Metadaten erreicht werden kann. Insbesondere die Modellierung von Metadaten stellt eine zentrale Herausforderung dar, da HDF5 selbst keine semantischen Vorgaben zur Strukturierung oder Interpretation von Inhalten macht.

Die hohe Flexibilität von HDF5 ist zugleich Stärke und Schwäche des Formats. Als generisches *Containerformat* erlaubt es eine frei gestaltbare hierarchische Organisation von Gruppen und Datensätzen, wobei jede Gruppe und jeder Datensatz mit beliebigen Attributen versehen werden kann. Weder Struktur noch Metadaten unterliegen verpflichtenden Vorgaben, und es existieren keine eingebauten Mechanismen zur Sicherstellung terminologischer Einheitlichkeit oder semantischer Konsistenz. Die Verantwortung für die Ausgestaltung der internen Struktur sowie für die Wahl und Bedeutung von Metadaten liegt vollständig bei den Erstellern der Datei. In der Praxis führt dies häufig zu stark heterogenen Dateiinhalten, deren Wiederverwendbarkeit und Vergleichbarkeit eingeschränkt ist.

Obwohl HDF5 aufgrund der engen Kopplung von Daten und Attributen häufig als selbstbeschreibendes Format bezeichnet wird, ist diese Selbstbeschreibung in der Regel lediglich syntaktischer Natur. Attribute orientieren sich oft an projektspezifischen Konventionen, persönlichen Präferenzen oder an den Bezeichnern von Messinstrumenten und Simulationssoftware, aus denen die Daten hervorgehen. Ohne zusätzliche formale Regeln bleibt die Bedeutung der gespeicherten Informationen für Dritte oder für automatisierte Auswertungen häufig unklar. Diese Problematik ist nicht formatspezifisch, tritt bei HDF5 jedoch aufgrund seiner bewusst offenen Gestaltung besonders deutlich zutage.

Für eine reproduzierbare Nutzung durch andere Forscher sowie für eine maschinelle Weiterverarbeitung bedarf es daher klar definierter Regeln zur Strukturierung von Daten und zur Beschreibung ihrer Bedeutung. Da wissenschaftliche Fragestellungen, Randbedingungen und Datenarten projektspezifisch variieren, kann es dabei kein universelles, starres Datenmodell geben. Erforderlich ist vielmehr ein Ansatz, der einerseits genügend Flexibilität für unterschiedliche Anwendungsfälle bietet, andererseits aber verbindliche Mindestanforderungen an Struktur, Metadaten und Terminologie formuliert und überprüfbar macht.

Zur Schließung dieser Lücke wird im Rahmen des vorliegenden Forschungsdatenmanagementkonzepts ein mehrschichtiges Modell der Daten- und Metadatenmodellierung eingeführt. Dieses trennt klar zwischen der Speicherung strukturierter Primärdaten, der formalen Beschreibung von Metadaten, der Qualitätssicherung sowie dem Zugriff auf Daten und Metadaten. Die Architektur gliedert sich in vier funktional klar abgegrenzte Ebenen:

1. **Basisebene (HDF5 und RDF):** Die Basisebene bildet die technische Grundlage der Datenhaltung. Strukturierte Primärdaten werden in HDF5 gespeichert, während Metadaten in RDF formal beschrieben werden. HDF5 übernimmt die effiziente Ablage großer, komplex strukturierter Daten, während RDF ein standardisiertes, maschinenlesbares Modell zur Repräsentation von Metadaten und deren Beziehungen bereitstellt.
2. **Validierungsschicht:** Zur Sicherstellung struktureller und semantischer Konsistenz wird eine Validierungsschicht eingeführt. Mithilfe der *h5rdmtoolbox* (Probst und Pritz, 2026) können Layoutdefinitionen formuliert werden, die die interne Struktur von HDF-Dateien sowie verpflichtende oder optionale Attribute festlegen. Über sogenannte *Conventions* lassen sich darüber hinaus zulässige Wertebereiche und kontrollierte Vokabulare überprüfen. Die Validierung kann sowohl während der Dateierstellung als auch nachträglich erfolgen und dient der Qualitätssicherung der Datenbestände.

3. **Semantische Schicht (OWL und SPARQL):** Aufbauend auf RDF werden Metadaten durch Ontologien in OWL semantisch angereichert. Dadurch lassen sich domänenspezifische Begriffe, Hierarchien und Relationen formal definieren. Über SPARQL-Abfragen können diese Informationen gezielt durchsucht und ausgewertet werden, etwa im Hinblick auf physikalische Größen, Einheiten oder Randbedingungen.
4. **API-Schicht (Zugriff und Integration):** Die oberste Ebene stellt Schnittstellen für den programmatischen Zugriff auf Daten und Metadaten bereit. Während strukturierte Daten über eine Python-API adressiert werden, erfolgt der Zugriff auf semantische Metadaten über SPARQL. Diese Trennung ermöglicht sowohl klassische numerische Analysen als auch semantisch fundierte Recherchen und unterstützt die Integration in bestehende Softwarelösungen, Repositorien und Forschungsinfrastrukturen.

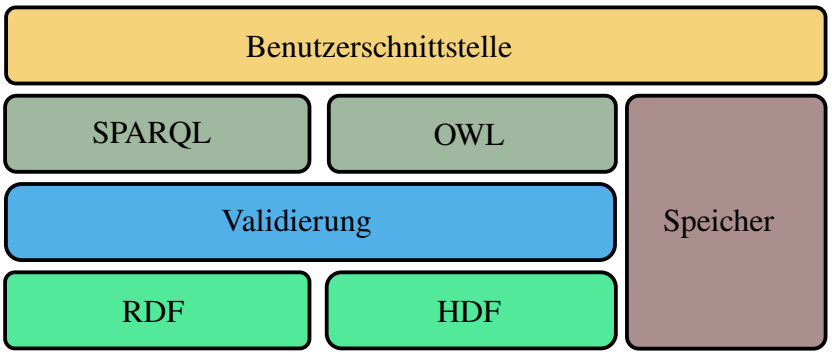


Abbildung 4.3: Schichtenmodell des Forschungsdatenmanagements in Anlehnung an den Semantic Web Stack. Die Architektur gliedert die Verarbeitung wissenschaftlicher Daten in klar abgegrenzte Ebenen: Die Basisschicht kombiniert HDF5 für die Speicherung strukturierter Primärdaten mit RDF zur formalisierten Beschreibung von Metadaten. Darauf aufbauend überprüft die Validierungsschicht mithilfe definierter Layouts die interne Struktur der HDF-Dateien sowie die regelbasierte Vergabe erforderlicher Attribute. Die semantische Schicht mit OWL und SPARQL ermöglicht eine inhaltliche Anreicherung und gezielte Abfrage der Metadaten. Über die API-Schicht wird der programmatische Zugriff auf Daten und Metadaten realisiert. Eine seitlich verlaufende Spalte zur persistenten Datenhaltung verdeutlicht die durchgängige Speicherung über alle Funktionsebenen hinweg.

Ergänzend durchzieht eine persistente Datenhaltung alle Ebenen des Modells. Dabei werden nicht nur die Forschungsdaten selbst, sondern auch Validierungsregeln und Ontologiedefinitionen dauerhaft gespeichert. Abbildung 4.3 veranschaulicht das Konzept in Form eines funktional geschichteten Modells, das in Anlehnung an den Semantic Web Stack entwickelt wurde. Die Darstellung verdeutlicht die klare Trennung der Ebenen sowie deren Zusammenspiel über den gesamten Datenlebenszyklus hinweg. Dieses strukturierte Schichtenmodell schafft einen klaren

Rahmen für wissenschaftlich nachhaltiges Forschungsdatenmanagement. Es verbindet technische Flexibilität mit semantischer Annotation und ermöglicht durch die Kombination von Validierung, Ontologien und Schnittstellen sowohl maschinenlesbare als auch nutzerfreundliche Speicherung und Verwaltung von Daten. Es unterstreicht die zentrale Rolle der langlebigen, integren und zugänglichen Speicherung sowohl der Daten als auch ihrer Beschreibung – unabhängig von der jeweiligen Funktionsebene.

In den folgenden Kapiteln werden die einzelnen Bausteine dieses Modells detailliert ausgearbeitet. Zunächst erfolgt die semantische Beschreibung wissenschaftlicher Metadaten im Kontext von HDF5 mithilfe von Ontologien. Darauf aufbauend wird mit der domänenspezifischen Standardnamenontologie *SSNO* ein kontrolliertes Vokabular für physikalische Größen und deren Repräsentation in Datenstrukturen vorgestellt.

4.5 Metadatenbeschreibung mittels Ontologien

Im vorhergehenden Kapitel wurde HDF5 als technisches Fundament zur strukturierten Speicherung von Forschungsdaten ausgewählt. HDF5 ermöglicht eine flexible Organisation komplexer Datenbestände und ist selbstbeschreibend durch die Möglichkeit, Metadaten als Attribute mit Rohdaten zu speichern. Allerdings trifft HDF5 keine formalen Aussagen über die inhaltliche Bedeutung der gespeicherten Daten oder über die semantische Interpretation ihrer Struktur auf maschinelle Weise. Kontext, Bedeutung und Relationen der Daten bleiben damit explizit außerhalb des Datenformats und sind nicht standardisiert maschinell interpretierbar.

Für eine nachhaltige Nachnutzung, Interoperabilität und automatisierte Verarbeitung ist es daher erforderlich, sowohl die interne Struktur einer HDF5-Datei als auch die Bedeutung der enthaltenen Daten formal und eindeutig zu beschreiben. HDF5 stellt hierfür keine native semantische Beschreibungssprache bereit. Standardisierte semantische Modelle wie RDF oder OWL müssen deshalb ergänzend eingesetzt werden.

Dieses Kapitel behandelt die formale Modellierung struktureller und inhaltlicher Metadaten von HDF5-basierten Datensätzen mithilfe externer Ontologien. Die semantische Beschreibung erfolgt dabei komplementär zur strukturellen Speicherung in HDF5 und bildet eine eigenständige, interoperable Metadatenebene.

Das Kapitel widmet sich dabei zwei zentralen Fragestellungen:

1. Wie lässt sich die interne Struktur einer HDF-Datei präzise und formal beschreiben, um eine konsistente, maschinenlesbare Organisation der Daten zu gewährleisten?
2. Wie können Kontext und Bedeutung der Daten so modelliert werden, dass ihre semantische Aussagekraft für langfristige Nutzung und Interpretation erhalten bleibt?

Bei der Modellierung semantischer Metadaten ist zwischen *technischen* und *kontextuellen* Informationen zu unterscheiden:

- **Technische Informationen** beschreiben Struktur, Format und Speicherorganisation der Daten, etwa Datentypen, Dimensionen oder Speicherlayout. Sie sind notwendig, um Daten korrekt zu interpretieren, zu validieren und softwareseitig weiterzuverarbeiten. Diese

Informationen sind in der HDF5-Struktur bereits implizit vorhanden und können durch Ontologien explizit gemacht werden.

- **Kontextuelle Informationen** beziehen sich auf die inhaltliche Bedeutung der Daten, beispielsweise welche physikalischen Größen erfasst wurden, unter welchen Randbedingungen Messungen oder Simulationen durchgeführt wurden, oder welchem Projekt und welchen Personen die Daten zuzuordnen sind. Diese Informationen sind typischerweise extern zur HDF5-Datei und müssen aktiv (manuell oder automatisiert) erfasst werden.

Diese Differenzierung ist wesentlich für ein nachhaltiges Forschungsdatenmanagement, da sie eine klare Trennung zwischen struktureller Handhabbarkeit und inhaltlicher Interpretation ermöglicht. Ontologien stellen hierfür ein geeignetes Instrument dar, da sie sowohl technische als auch kontextuelle Informationen formal, eindeutig und maschinenlesbar modellieren können.

Strukturelle Beschreibung

Die technische Beschreibung einer HDF-Datei lässt sich aus zwei Perspektiven betrachten: einer *externen* und einer *internen*.

Die *externe Perspektive* richtet sich auf Aspekte der Veröffentlichung, Auffindbarkeit und Zugänglichkeit der Daten. Dazu zählen der Speicherort (z. B. eine persistente URL), das verwendete Dateiformat, Versions- und Lizenzinformationen, Angaben zu verantwortlichen Personen sowie das Erstellungs- und Veröffentlichungsdatum. Eine präzise und interoperable Beschreibung dieser Merkmale sollte auf standardisierte Metadatenvokabulare wie *DCAT* zurückgreifen (vgl. Unterabschnitt 2.3.1).

Demgegenüber fokussiert die *interne Perspektive* auf die Struktur und Organisation der Datei selbst. Sie beschreibt die hierarchische Anordnung von Gruppen, Datensätzen und Attributen sowie deren Datentypen, Dimensionen und Speichereigenschaften. Zur formalen Repräsentation dieser Strukturelemente wurde von der Allotrope Foundation eine Ontologie entwickelt, die grundlegende HDF5-Konzepte abbildet (Allotrope Foundation, 2024; Millecam et al., 2021). Diese Ontologie, im Folgenden AF-HDF genannt, stellt Klassen wie *hdf:File*, *hdf:Group* und *hdf:Dataset* bereit, die sich für eine formale, eindeutige Repräsentation der internen HDF5-Struktur eignen. Die AF-HDF-Ontologie ist als Teil des Allotrope Data Format (ADF) etabliert und wird von einem breiten Konsortium von Pharma- und Wissenschaftsunternehmen gepflegt (Allotrope Foundation, 2024).

Die beiden Perspektiven sind in Abbildung 4.4 schematisch dargestellt. Während die linke Seite (externe Perspektive) die Integration in Datenkataloge und Repositorien mittels *DCAT* abbildet, zeigt die rechte Seite (interne Perspektive) die hierarchische Struktur der Datei auf Grundlage der AF-HDF-Ontologie. Die dazugehörigen RDF-Serialisierungen beider Perspektiven sind in Listing 4.1 und Listing 4.2 im Turtle-Format dargestellt.

Beide Beschreibungen beziehen sich auf dasselbe physische Objekt (die Datei *result.hdf*). Ihre Verknüpfung erfolgt über das im *DCAT*-Vokabular definierte Prädikat *dcat:downloadURL*, das eine logische Verbindung zwischen der konzeptionellen Beschreibung (*dcat:Dataset*) und der

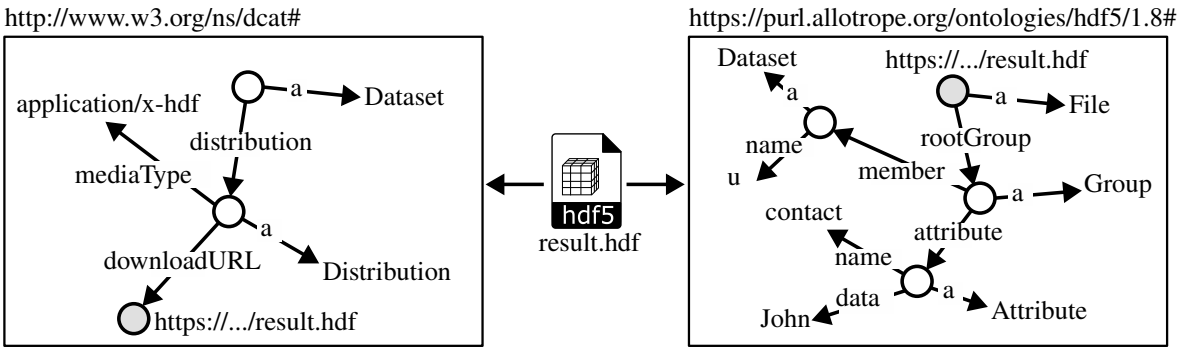


Abbildung 4.4: Schematische Darstellung zweier Perspektiven auf dieselbe HDF-Datei *result.hdf*. Links: externe Beschreibung mit *DCAT* für Auffindbarkeit und Katalogisierung. Rechts: interne Beschreibung der Datei mittels AF-HDF-Ontologie für maschinenlesbare Analyse und Verarbeitung. Beide Graphen sind über die Download-Ressource verknüpft (grauer Knoten, ○). Die Verknüpfung ermöglicht es, dass konzeptionelle und technische Beschreibung nicht isoliert, sondern kohärent interpretiert werden können.

konkreten Datei (*hdf:File*) herstellt. In der Abbildung ist dies über den grauen Knoten (○) verdeutlicht.

Listing 4.1: Externe Beschreibung einer HDF-Datei als *dcat:Dataset* im Turtle-Format.

```
@prefix dcat: <http://www.w3.org/ns/dcat#> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix ex: <https://example.org/> .

ex:resultDS a dcat:Dataset ;
    dcterms:creator <https://orcid.org/0000-0001-8729-0482> ;
    dcterms:title "Simulationsergebnisse"@de ;
    dcat:distribution [ a dcat:Distribution ;
      dcat:downloadURL <https://example.org/result.hdf> ;
      dcat:mediaType "application/x-hdf5" ] .
```

Listing 4.2: Interne Beschreibung der HDF-Struktur mit der AF-HDF-Ontologie. Als Ressource dient dieselbe URL, auf die in der DCAT-Beschreibung verwiesen wird.

```
@prefix hdf: <http://purl.allotrope.org/ontologies/hdf5/1.8#> .

<https://example.org/result.hdf> a hdf:File ;
    hdf:rootGroup [ a hdf:Group ;
        hdf:name "/" ;
        hdf:member [ a hdf:Dataset ;
```

```
            hdf:name "/u" ;
            hdf:datatype hdf:H5T_IEEE_F64LE;
            hdf:dataspace [
                a hdf:Dataspace ;
                    hdf:rank 3 ; ]
        ]
    ] .
```

Die daraus resultierende Beziehungskette lautet:

$$\textit{dcat:Dataset} \rightarrow \textit{dcat:Distribution} \rightarrow \textit{dcat:downloadURL} \rightarrow \textit{hdf:File} \quad .$$

Damit bleibt der semantische Unterschied zwischen der abstrakten Datensatzbeschreibung und der konkreten Repräsentation des Dateiinhalts gewahrt, während zugleich eine semantisch korrekte Verbindung zwischen beiden hergestellt wird. Ein *dcat:Dataset* beschreibt die Forschungsressource auf konzeptueller Ebene, während das über *dcat:downloadURL* referenzierte Objekt als *hdf:File* spezifiziert ist.

Diese Modellierung ermöglicht es automatisierten Systemen, semantische Schlussfolgerungen (Inferenz) zu ziehen und damit die FAIR-Prinzipien technisch zu operationalisieren. Auf der strukturellen Ebene (AF-HDF + DCAT) kann ein System beispielsweise automatisch erkennen, dass mehrere HDF5-Dateien die gleiche technische Struktur aufweisen – etwa die gleichen Datentypen und Dimensionen sowie identische oder kompatible Zeichenkette für Einheiten (z.B. beide „m/s“). Dies ermöglicht eine erste technische Vergleichbarkeit und adressiert Interoperabilität (I) auf struktureller Ebene. Durch Provenienzannotationen (*prov:wasDerivedFrom*, *prov:wasGeneratedBy*) können sog. „Reasoner“ automatisch nachvollziehen, welche Daten und Methoden in ein Ergebnis flossen, was Reproduzierbarkeit und Wiederverwendbarkeit (R) gewährleistet. Maschinelle Validierung gegen semantisch definierte Beschränkungen (wie sie etwa SHACL formalisiert) stärkt zusätzlich Auffindbarkeit (F) durch höhere Datenqualität.

Damit sich der Mehrwert dieser strukturellen Modellierung vollständig entfaltet, müssen die Inhalte der HDF5-Dateien durch domänenspezifische Metadaten ergänzt werden. Erst mit inhaltlich-semantischen Annotationen können automatisierte Systeme fachliche Vergleichbarkeit erkennen – etwa dass zwei Geschwindigkeitsfelder unter ähnlichen Reynolds-Zahlen oder Randbedingungen entstanden sind. Dies eröffnet weiterführende Inferenzmöglichkeiten für fachliche Validierung, Datenintegration und automatisierte Wissenserschließung, wie im folgenden Abschnitt erläutert wird.

Inhaltliche Beschreibung

Technische Modellierungen durch Ontologien wie *DCAT* oder *AF-HDF* erfassen keine domänenspezifischen Wissenskonzepte, da sie als generische, format- bzw. strukturorientierte Vokabulare konzipiert sind. HDF-Dateien fungieren hingegen als Container für Daten, deren Bedeutung inhaltlich einer spezifischen Domäne zugeordnet ist, etwa Strömungsmechanik, Materialwissenschaft oder Chemie. Der semantische Kontext dieser Daten kann nur durch die

Bereitstellung geeigneter domänenspezifischer Attribute und Annotationen erschlossen werden. Dabei ergeben sich zwei zentrale Herausforderungen:

(a) Die Einhaltung von Benennungskonventionen liegt vollständig in der Verantwortung der Anwender, was leicht zu Inkonsistenzen und Missverständnissen führt.

(b) Die semantische Bedeutung natürlichsprachiger Attribute ist für automatisierte Systeme nicht ohne weiteres interpretierbar.

Diese Defizite führen sowohl bei der manuellen Analyse als auch bei der automatisierten Verarbeitung zu Mehrdeutigkeiten und Interpretationsproblemen. Ein fundamentales Problem ist die inhärente Ambiguität natürlicher Sprache: Einzelne Begriffe können je nach Kontext unterschiedliche Konzepte bezeichnen (Polysemie) oder mehrere Bedeutungen haben (Homonymie). So kann der Begriff *Druck* den mechanischen, dynamischen, statischen oder totalen Druck eines Fluids, aber auch psychologischen Druck bezeichnen. Ohne zusätzlichen Kontext bleibt die exakte Bedeutung unklar, was das Risiko von Missinterpretationen erheblich erhöht. Wie in Abbildung 2.1 dargestellt, mindert dies den Informationsgehalt und damit den Mehrwert der Daten. Der erhöhte Aufwand zur Kontextbestimmung wirkt sich zudem negativ auf die Akzeptanz aus und reduziert die Wahrscheinlichkeit einer Nachnutzung in späteren Forschungsarbeiten.

Für eine eindeutige, maschineninterpretierbare Interpretation ist daher eine formalisierte Struktur erforderlich, die jedem Konzept und jedem Attribut eine präzise, referenzierbare Bedeutung zuweist. Ein naheliegender Ansatz besteht in der Ergänzung natürlicher Attribute um zusätzliche Kontextinformationen, etwa „physikalische Einheit“ oder „Fluideigenschaft“. Noch konsistenter und maschineninterpretierbarer gelingt dies durch die Verwendung semantischer Annotationen mithilfe von Internationalen Ressourcen-Identifikatoren (IRIs). Anstatt das Attribut *unit* lediglich textuell zu definieren, kann es beispielsweise mit der standardisierten Eigenschaft *m4i:hasUnit*) verknüpft werden, und der Wert „m/s“ kann durch die standardisierte QUDT-Einheit *qudt:M-PER-SEC* präzisiert werden. Auf diese Weise wird die Bedeutung maschinenlesbar kodiert und in ein standardisiertes semantisches Netzwerk eingebettet, was die Interoperabilität und Wiederverwendbarkeit deutlich stärkt.

Zur Verdeutlichung wird nachfolgend zwischen **natürlichen Attributen** und **semantischen Attributen** unterschieden:

- **Natürliche Attribute** sind benutzerdefinierte Angaben in natürlicher Sprache, wie etwa „Einheit“, „Beschreibung“ oder „Datum der Aufzeichnung“. Auch dann, wenn die Benennungen einer etablierten Konvention folgen, bleibt ihre semantische Eindeutigkeit begrenzt, da natürliche Sprache interpretationsabhängig ist. Diese Attribute sind zwar einfach zu erstellen und zu verstehen, ermöglichen aber nur begrenzte maschinelle Verarbeitung.
- **Semantische Attribute** basieren auf standardisierten Ontologien und werden durch RDF-Tripel mittels IRIs definiert. Jedes Attribut referenziert auf eine formal definierte Eigenschaft einer Ontologie (z.B. *m4i:hasUnit*). Vorausgesetzt, dass die verwendeten Ontologien dokumentiert und für Systeme zugänglich sind, erhalten die Metainformationen dadurch einen klaren Kontext und sind standardisiert maschineninterpretierbar.

In der Praxis existiert ein Spektrum zwischen diesen beiden Polen. Die Anreicherung natürlicher Attribute mit semantischen Annotationen kann die Beschreibungstiefe und Maschineninterpretierbarkeit erhöhen, erfordert jedoch sorgfältiges Vorgehen und zusätzlichen Aufwand bei der Erstellung der HDF5-Datei. Ein ausgewogenes Verhältnis zwischen Detaillierung und Praktikabilität ist daher entscheidend. Als praktische Richtlinie sollten mindestens Schlüsselkonzepte und kritische physikalische Größen semantisch annotiert werden; detailliertere Annotationen können schrittweise ergänzt werden.

Durch die systematische semantische Annotation von HDF-Datensätzen, -Gruppen und -Attributen mittels IRIs wird die Bedeutung der Inhalte innerhalb einer HDF5-Datei eindeutig und formal festgelegt. Da HDF5 und gängige Programmierschnittstellen eine solche semantische Erweiterung bislang nicht nativ unterstützen, wurde im Rahmen dieser Arbeit eine entsprechende Lösung entwickelt. Ihre technische Umsetzung erfolgt mithilfe der Python-Bibliothek *h5rdmtoolbox* (Probst und Pritz, 2024, 2026) und wird in Unterabschnitt A.2.1 detailliert beschrieben. An dieser Stelle wird zunächst das zugrunde liegende Konzept erläutert.

Annotation von HDF5 Attributen mittels IRIs

Abbildung 4.5 stellt die parallele Verwendung natürlicher und semantischer Attribute in einer HDF-Datei anhand eines RDF-Graphen dar. Das Beispiel beschreibt eine HDF-Gruppe mit dem Pfad „/Kontakt“, die zur Dokumentation der Ansprechperson für Rückfragen zum Dateiinhalt dient. Im RDF-Graphen ist diese Gruppe als hellgrauer Knotenpunkt ○ dargestellt.

Der Gruppe wurde beim Erstellen das Attribut „id“ mit dem Wert „0000-0001-8729-0482“ hinzugefügt. Im Graphen ist dies über die Eigenschaften *hdf:name* und *hdf:data* des Attribut-Knotens (○) abgebildet. Ohne weitere Kontextualisierung bleibt jedoch unklar, dass es sich hierbei um eine ORCID-iD handelt – der Wert ist lediglich eine Zeichenkette und damit für Maschinen nicht eindeutig interpretierbar.

Die semantische Anreicherung wird im rechten Teil des RDF-Graphen durch rot gestrichelte Pfeile veranschaulicht. Die HDF-Gruppe „/Kontakt“ repräsentiert inhaltlich eine Person. Entsprechend verweist der rote Knotenpunkt (●) mittels *rdf:type* auf die Klasse *foaf:Person*. Diese Person verfügt über eine eindeutige Identifikationsnummer, die von ORCID vergeben wird. Die Verbindung zwischen der Person und dem Attributwert erfolgt über die Eigenschaft *m4i:orcidId*. Da die Person als Ressource eindeutig identifizierbar ist, kann sie durch die IRI https://orcid.org/0000-0001-8729-0482 repräsentiert werden.

Im RDF-Graphen lassen sich somit sowohl die HDF-Gruppe (○) als auch die inhaltlich beschriebene Person (●) als eigenständige Ressourcen modellieren. Um auszudrücken, dass die Gruppe inhaltlich genau diese Person beschreibt, können beide Ressourcen über eine geeignete Relation – etwa *dcterms:relation* – miteinander verknüpft werden.

Das Beispiel verdeutlicht, wie sich Elemente einer HDF-Datei konsistent in RDF-Strukturen abbilden lassen: HDF Gruppen oder Datensätzen fungieren als Subjekte, Attribute bilden Schlüssel-Wert-Paare und entsprechen damit dem RDF-Prädikat und -Objekt. Eine Übersicht dieser Entsprechungen bietet Tabelle 4.1.

Zur Demonstration der praktischen Umsetzung zeigt Listing 4.3 den entsprechenden Python-Code. Die daraus resultierende RDF-Repräsentation im Turtle-Format ist in Listing 4.4 dargestellt und entspricht inhaltlich dem in Abbildung 4.5 visualisierten Ausschnitt. Weitere technische Details zur semantischen Annotation mit der Python-Bibliothek *h5rdmtoolbox* werden in Unterabschnitt A.2.1 erläutert.

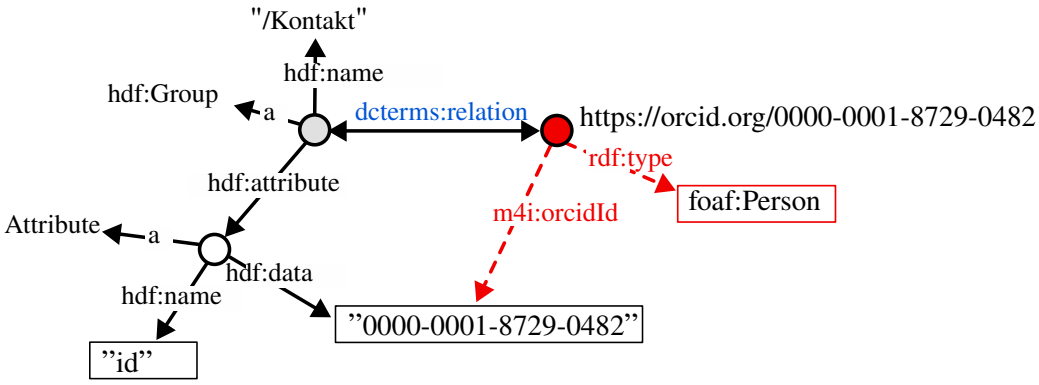


Abbildung 4.5: Semantisches Äquivalent der HDF5 Attributvergabe am Beispiel von Kontaktinformationen. Das Attribut „id“ der Gruppe „/Kontakt“ kann semantisch als Person (*foaf:Person*) annotiert werden, die über eine IRI verfügt und die Eigenschaft *m4i:orcidId* besitzt.

HDF5 Element	**RDF Äquivalent**	**Beispiel**
Dataset / Gruppe	Subjekt (IRI)	<https://orcid.org/0000...>
Attribut-Schlüssel	Prädikat (IRI)	m4i:orcidId
Attribut-Wert	Objekt (Literal/IRI)	"0000-0001-8729-0482"

Tabelle 4.1: Gegenüberstellung bzw. Zuordnung der Bestandteile einer HD5 Datei zu den Elementen der Bestandteile eines RDF Tripels.

Mit der Möglichkeit, semantische Metadaten im Kontext von HDF5 zu nutzen, wurde zunächst eine *technische Lösung* geschaffen. Viele grundlegende Eigenschaften zur Annotation von Beziehungen und Referenzen lassen sich dabei aus etablierten High-Level-Ontologien wie *PROV*, *DCTERMS*, *DCAT* oder *M4I* übernehmen. Für die Beschreibung domänenspezifischen Wissens, z. B. zur Charakterisierung von Strömungsparametern, Messgrößen oder experimentellen Bedingungen, stehen jedoch nicht in allen Fällen geeignete, öffentlich verfügbare Ontologien zur Verfügung.
Eine vollständige Eigenentwicklung einer umfassenden Domänenontologie wäre zwar konzeptionell ideal, geht jedoch mit erheblichem fachlichem Aufwand, hohem Zeitbedarf und kontinuierlichem Pflegebedarf einher. Stattdessen orientiert sich diese Arbeit an der bewährten Praxis der *CF Conventions* (vgl. Unterabschnitt 2.3.1), einem etablierten Standard für die Beschreibung und Dokumentation von Messdaten aus der Klimaforschung. Die CF Conventions definieren standardisierte Vokabulare (sogenannte *Standardnamen*), die Konzepte wie physikalische Grö-

Listing 4.3: Implementierung des Beispiels aus Abbildung 4.5 durch *h5rdmtoolbox*.

```python
import h5rdmtoolbox as h5tbx

with h5tbx.File() as h5:
    g = h5.create_group(
        name="Kontakt",
        rdf_type="http://www.w3.org/ns/prov#Person",
        rdf_subject="https://orcid.org/0000-0001-8729-0482"
    )
    g.attrs["id"] = h5tbx.Attribute(
        value='0000-0001-8729-0482',
        rdf_predicate="http://w3id.org/nfdi4ing/metadata4ing#orcidId"
    )
    print(h5.serialize(fmt="ttl", structural=True))
```

Listing 4.4: Serialisierung der HDF5 Datei aus Listing 4.3

```
@prefix hdf: <http://purl.allotrope.org/ontologies/hdf5/1.8#> .
@prefix m4i: <http://w3id.org/nfdi4ing/metadata4ing#> .
@prefix prov: <http://www.w3.org/ns/prov#> .
@prefix schema: <https://schema.org/> .

<https://orcid.org/0000-0001-8729-0482> a prov:Person ;
    m4i:orcidId "0000-0001-8729-0482" .

[] a hdf:File ;
    hdf:rootGroup [ a hdf:Group ;
        hdf:member [ a hdf:Group ;
            hdf:attribute [ a hdf:StringAttribute ;
                hdf:data "0000-0001-8729-0482" ;
                hdf:name "id" ] ;
            hdf:name "/Kontakt" ;
            schema:about <https://orcid.org/0000-0001-8729-0482> ] ;
        hdf:name "/" ] .
```

ßen und Einheiten konsistent und präzise beschreiben.
Bislang existiert jedoch keine formale Ontologie-Repräsentation der CF Conventions, die eine maschinenlesbare, interoperable Nutzung der Standardnamen ermöglicht. Das folgende Kapitel adressiert diese Lücke: Es entwickelt eine Ontologie, die die für das FDM-Konzept relevanten Kernkomponenten der Standardnamen formalisiert und damit die Brücke zwischen informalen Konventionen und semantischen Web-Technologien schlägt. Auf diese Weise wird ein flexibles, wartbares Rahmenwerk geschaffen, das sowohl die Präzision domänenspezifischen Wissens als auch die praktische Handhabbarkeit und langfristige Wartbarkeit sicherstellt.

4.6 Einführung der Standardnamenontologie SSNO

Die nachhaltige und konsistente Beschreibung wissenschaftlicher Daten im Sinne der FAIR-Prinzipien setzt den gezielten Einsatz von Ontologien voraus. Hierbei entsteht jedoch ein Spannungsfeld: Einerseits erfordert die Entwicklung domänenspezifischer Ontologien erheblichen fachlichen und organisatorischen Aufwand, andererseits ist eine hinreichend präzise und maschinenlesbare Metadatenbeschreibung unverzichtbar.
Mit dem Konzept der **Standardnamen** greift diese Arbeit eine praxisnahe Lösung auf, die zwischen diesen beiden Polen vermittelt. Standardnamen gehen auf die etablierten *Climate and Forecast* (CF) *Conventions* aus der Klimaforschung zurück, die einheitliche Bezeichnungen für physikalische Größen und deren Eigenschaften bereitstellen. Dieses Prinzip lässt sich auf HDF5-Daten übertragen und generalisieren, sodass auch in anderen Fachbereichen eindeutige und interoperable Metadaten gewährleistet werden können.
Da bislang keine formale Ontologie existiert, die das Konzept der Standardnamen im RDF/OWL-Format abbildet, werden in dieser Arbeit die Grundprinzipien der CF Conventions verallgemeinert und in einer eigenen Ontologie modelliert. Ziel ist es, semantische Metadaten domänenübergreifend nutzbar zu machen und gleichzeitig genügend Flexibilität zu wahren, um fachspezifische Erweiterungen zu ermöglichen. Dadurch entfällt die Notwendigkeit, für jedes Projekt oder Fachgebiet individuelle Ontologien neu zu entwickeln und langfristig zu pflegen.

In den folgenden Abschnitten wird das Konzept der Standardnamen, wie es in den CF Conventions definiert ist, zunächst im Detail erläutert und anschließend so erweitert, dass es allgemeiner und disziplinunabhängiger anwendbar ist. Auf dieser Grundlage wird die *Simple Standard Name Ontology* (SSNO) entwickelt.

4.6.1 Zentrale Konzepte von Standardnamen und -tabellen

Die CF Conventions werden zentral unter https://cfconventions.org (M. Harris, 2025a) gepflegt und spezifizieren, welche Attribute für jede Variable in einernetCDF4-Datei[3] verwendet werden sollten. Zentrale Attribute sind die physikalische Einheit (*units*), eine menschenverständliche Kurzbeschreibung (*long_name*) sowie der *standard_name*. Letzterer spielt eine Schlüsselrolle, da er eine exakte, standardisierte Beschreibung der Variablen ermöglicht und damit die Auffindbarkeit und Vergleichbarkeit von Daten entscheidend verbessert.

Standardnamen bestehen aus durch Unterstriche getrennten englischen Begriffen und sind in einer von der Community gepflegten Standardnamentabelle verzeichnet. Diese Tabelle enthält nicht nur eine Liste zulässiger Standardnamen, sondern auch Regeln zur Definition neuer Namen sowie zur Modifikation bestehender Bezeichnungen (M. Harris, 2025a). Ein Beispiel ist:

air_pressure_at_mean_sea_level .

[3]Die CF Conventions wurden ursprünglich für das Format netCDF4 entwickelt, das ab Version 4.0.0 HDF5 als Basisformat nutzt (Unidata, 2024). netCDF4-Variablen entsprechen direkt HDF5-Datasets, und die konzeptionellen Strukturen sind daher unmittelbar übertragbar. Auch wenn die Konventionen formal netCDF4 adressieren, ermöglicht ihre technische Nähe zu HDF5 eine direkte Anwendung auf die in dieser Arbeit betrachteten HDF5-basierten Datenstrukturen.

Ein solcher Standardname kodiert komplexe Informationen in kompakter Form, ist menschen- wie maschinenlesbar und verweist auf eine präzise Definition der enthaltenen Konzepte (z. B. *air_pressure*, *mean_sea_level*). Ergänzend wird für jeden Standardnamen eine kanonische Einheit spezifiziert.
Durch diese Standardisierung können Analyse- und Visualisierungsprogramme automatisch auf Variablen zugreifen und deren Kontext interpretieren. Zahlreiche Softwarepakete (M. Harris, 2025b), darunter auch generische Bibliotheken wie die Python-Bibliothek *xarray* (Hoyer und Hamman, 2017), unterstützen die Verarbeitung solcher standardisierter Attribute.

Die CF Conventions sind in Klimaforschung, Meteorologie und Ozeanografie weit etabliert. Außerhalb dieser Disziplinen existiert jedoch bislang keine systematische Adaption des Ansatzes, obwohl gerade in Bereichen mit fragmentierten Metadatenstrukturen erheblicher Nutzen zu erwarten wäre. Eine disziplinübergreifende Übertragung erfordert die Generalisierung zweier zentraler Aspekte: Erstens müssen die taxonomischen Kategorien von Atmosphären- und Ozeanvariablen auf beliebige Domänen (Strömungsmechanik, Materialwissenschaft, Biologie etc.) verallgemeinert werden. Zweitens muss das Konzept von einer webbasierten Tabelle in eine formale RDF/OWL-Ontologie (Top-Level-Ontologie) überführt werden, um maschinenlesbaren Zugriff, semantische Verknüpfungen und automatisierte Validierung zu ermöglichen.

Bei der Bewertung von Standardnamen als Ansatz zur Datenannotation sind vier zentrale Qualitätskriterien entscheidend:

1. **Aufwand (Praktikabilität)**: Die Angabe von lediglich zwei Attributen (*standard_name* und *units*) verursacht nur geringen, praxisgerechten Mehraufwand.
2. **Informationsgehalt (Semantik)**: Über die Verlinkung zur Standardnamentabelle oder -ontologie sind umfassende Kontextinformationen unmittelbar verfügbar.
3. **Auffindbarkeit (Suchbarkeit)**: Die kompakte, standardisierte Benennung erleichtert die gezielte Suche nach Variablen innerhalb von Dateien und über Datensätze hinweg.
4. **Validierung (Konsistenz)**: Der Eintrag des Standardnamens erlaubt eine automatische Überprüfung der Einheit, indem die in der Tabelle/Ontologie definierte kanonische Einheit mit der in der Datei angegebenen verglichen wird.

4.6.2 Generalisierung und Ontologie-Modellierung

Für die Übertragung der CF Conventions auf weitere wissenschaftliche Disziplinen ist eine Abstraktion ihrer Kernprinzipien erforderlich. Drei zentrale Bausteine bilden dabei die Grundlage der Generalisierung:

1. **Standardnamen**: Kern der CF Conventions ist die Definition von Standardnamen, die jeweils aus einer Bezeichnung, einer kanonischen Einheit sowie einer präzisen Beschreibung bestehen. Im verallgemeinerten Kontext können Standardnamen beliebige domänenspezifische Konzepte repräsentieren.
2. **Standardnamentabelle**: Diese Tabelle dokumentiert sämtliche Standardnamen. Während die CF Conventions auf die Domäne der Klima- und Atmosphärenforschung beschränkt

sind, ist eine generalisierte Form für disziplinübergreifende Anwendungen erforderlich. Die Ontologie-Repräsentation ermöglicht dabei nicht nur die Verwaltung von Standardnamen, sondern auch die Formalisierung von Konstruktionsregeln und Beziehungen.

3. **Dokumentation und Erweiterbarkeit**: Die CF Conventions veröffentlichen die Standardnamentabelle sowie Regeln zur Bildung neuer Namen über eine zentrale Web-Plattform (M. Harris, 2025a). Für eine generalisierte, Ontologie-basierte Lösung muss ein vergleichbares, aber technisch unabhängigeres und flexibleres Konzept etabliert werden, das sowohl Nachvollziehbarkeit als auch dezentralisierte Weiterentwicklung sicherstellt.

Die vorliegende Arbeit beschränkt sich auf diese drei Kernkonzepte und überführt sie in eine Ontologie. Eine eigenständige Ontologie ist erforderlich, da bislang keine formale Ontologie für die CF Conventions existiert. Durch die Einführung eindeutiger Identifikatoren (IRIs) für Standardnamen werden insbesondere die automatische Verarbeitung, semantische Verknüpfungen und die Validierung von Daten erheblich erleichtert. Darüber hinaus können in einer Ontologie nicht nur Namen und Einheiten, sondern auch die logischen Regeln zur Konstruktion neuer Standardnamen konsistent abgebildet werden.

Die nachfolgend vorgestellte *Simple Standard Name Ontology* (SSNO) verfolgt das Ziel, die erfolgreichen Kernelemente der CF Conventions beizubehalten, sie jedoch zu abstrahieren und für den breiteren wissenschaftlichen Kontext zugänglich zu machen. Die zentralen Konzepte und Entwurfsentscheidungen der Ontologie werden im Folgenden dargestellt. Ergänzende technische Details finden sich in Unterabschnitt A.3.3.

Ontologieentwurf

SSNO ist als Ontologie in OWL2 definiert und mit RDF Schema (RDFS) formalisiert (Probst und Pritz, 2025c; W3C OWL Working Group, 2012). Abbildung 4.6 zeigt die Kernkomponenten der Ontologie *StandardName* und *StandardNameTable*. Beide stehen über die inversen Eigenschaften *ssno:standardNames* und *ssno:standardNameTable* in Beziehung. Eine Standardnamentabelle enthält mindestens einen Standardnamen sowie Regeln zu dessen Modifikation, die über *hasModifier* modelliert werden. Ebenso sind die integrierten Klassen externer Ontologien dargestellt.

Die Kernbausteine der *SSNO* repräsentieren konzeptuelle Entitäten und keine real-weltlichen Objekte wie etwa *foaf:Person*. Daher ist die Standardnamentabelle als *skos:ConceptScheme* und der Standardname als *skos:Concept*. Simple Knowledge Organization System (SKOS) ist ein Vokabular zur formalen Abbildung von Wissensorganisationen wie Thesauri, Klassifikationen oder Taxonomien.

Durch die Ableitung der zentralen Klassen *skos:Concept* bzw. *skos:ConceptScheme* können Standardnamen und -tabellen konsistent in semantische Begriffssysteme integriert werden. *SKOS* bringt zudem nützliche Eigenschaften mit, die direkt für die Beschreibung und Verwaltung von Standardnamentabellen genutzt werden können. Eine Auswahl relevanter Eigenschaften ist in Tabelle A.6 dargestellt.

Besonders hervorzuheben sind die Mechanismen zur Modellierung hierarchischer und assoziativer Beziehungen wie *skos:broader*, *skos:narrower* und *skos:related*, die eine strukturierte Einordnung der Konzepte erlauben. Darüber hinaus ist SKOS für dynamische Szenarien geeignet, wie sie auch für Standardnamentabellen typisch sind. So können Änderungen, Erweiterungen oder Verweise etwa über *skos:changeNote* und *skos:editorialNote* dokumentiert werden.

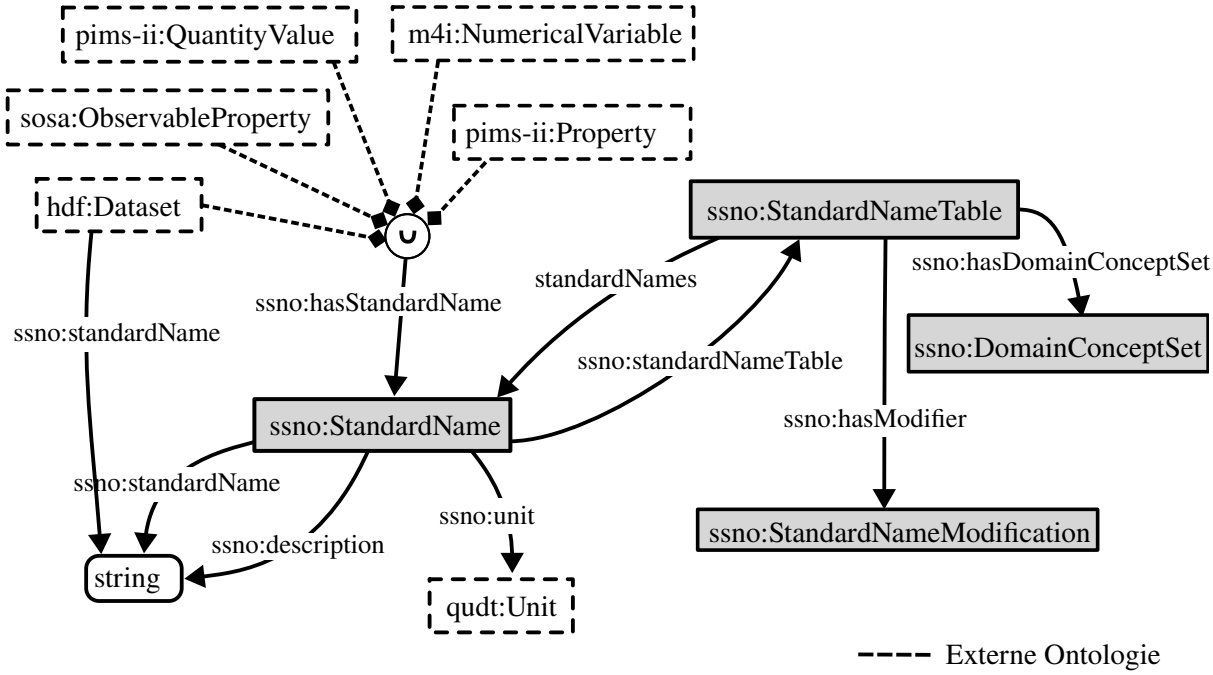


Abbildung 4.6: OWL-Klassendiagramm der *Simple Standard Name Ontology* mit den zentralen Beziehungen. Die Hauptklassen sind *ssno:StandardName* und *ssno:StandardNameTable*. Gestrichelte Klassen und weiß hinterlegte Felder verweisen auf externe Ontologien und verdeutlichen deren Einbettung in die *SSNO*.

Über SKOS hinaus werden weitere zentrale Eigenschaften zur Beschreibung zusätzlicher Metadaten durch das Dublin Core Metadata Terms (DCTERMS) Vokabular abgedeckt. Hierbei spielt die Möglichkeit der Versionierung eine wichtige Rolle. Dublin Core erleichtert die standardkonforme und interoperable Verwaltung von Metadaten, wodurch sich die Integration in bestehende Systeme vereinfacht. Insbesondere in wissenschaftlichen und institutionellen Umgebungen, in denen Metadaten nach den FAIR-Prinzipien strukturiert sein sollen, bieten die Begriffe des Dublin Core eine Grundlage, die sowohl maschinenlesbar als auch inhaltlich verständlich ist. Tabelle A.7 listet im Anhang dieser Arbeit die wichtigsten Dublin Core-Eigenschaften für die Modellierung der Standardnamentabelle auf.

Standardnamentabelle

Die Standardnamentabelle bildet in diesem Konzept ein zentrales Element. Sie wird durch ein Projekt, eine Forschungsgruppe oder ein Institut definiert und gepflegt. Neben der Zuord-

nung von Standardnamen über *ssno:standardNames* umfasst ihre Beschreibung insbesondere die Verknüpfung mit den Ressourcen, die sie nutzen, sowie die Referenzierung der Tabelle als zugängliche Datei.

Obwohl das Konzept der Standardnamen in dieser Arbeit primär für die Anwendung in HDF-Dateien eingeführt wird, ist seine Nutzung nicht darauf beschränkt. Standardnamentabellen können ebenso mit Projekten oder anderen Datensätzen verknüpft werden. Wie in Abbildung 4.7 dargestellt, sieht die Ontologie dafür die Eigenschaften *ssno:usesStandardNameTable* sowie die inverse Beziehung *ssno:standardNameTableUsedBy* vor. Beide sind als Spezialisierungen von *dcterms:relation* definiert, wodurch zusätzliche, in *SSNO* nicht explizit vorgesehene Relationen modellierbar bleiben. Relevante Anwendungsfälle sind insbesondere die Klassen *hdf:File*, *dcat:Dataset* und *schema:Project*[4].

Die Standardnamentabelle muss offen zugänglich und als Datei zum Download bereitgestellt werden. Dazu eignen sich unterschiedliche Serialisierungsformate, die eine Veröffentlichung in wissenschaftlichen Repositorien wie etwa Zenodo ermöglichen. Diese Speicherung kann über das *DCAT*-Vokabular abgebildet werden: Die Beziehung zwischen *ssno:StandardNameTable* und *dcat:Dataset* wird dabei über die Eigenschaft *ssno:dataset* modelliert, die ebenfalls als Untereigenschaft von *dcterms:relation* definiert ist. Ergänzend lässt sich aus Perspektive des Datensatzes die Zugehörigkeit zu einer Standardnamentabelle durch *dcat:theme* ausdrücken.

Standardnamen

Die Klasse *ssno:StandardName* repräsentiert einen durch eine Standardnamentabelle definierten Begriff und bildet damit das zentrale Element des Konzepts. Sie umfasst vier wesentliche Eigenschaften: *ssno:standardName*, *ssno:description*, *ssno:longName* und *ssno:unit*.

Die Eigenschaft *ssno:standardName* spezifiziert den eindeutigen Namen, der als Attributwert in HDF-Dateien genutzt wird, beispielsweise *x_velocity*. Über *ssno:description* wird eine ausführlichere textuelle Erläuterung bereitgestellt, die die Bedeutung des Begriffs präzisiert und seine Komponenten erklärt. Ein Beispiel: „Geschwindigkeit beschreibt die Bewegung von Gasen oder Flüssigkeiten. Der Zusatz *X* verweist auf die x-Koordinate des Koordinatensystems und gibt die Bewegungsrichtung an.“

Die Eigenschaft *ssno:longName* (optional) dient der Angabe eines alternativen, für Anwender verständlicheren Namens, der jedoch nicht für die maschinelle Verarbeitung vorgesehen ist. Schließlich verweist *ssno:unit* auf eine Referenzeinheit, die der SI-Einheit der jeweiligen physikalischen Größe entsprechen sollte. Die in einem Datensatz tatsächlich verwendete Einheit muss nicht identisch sein, muss jedoch eindeutig in die Referenzeinheit konvertierbar sein.

Da physikalische Einheiten in der Praxis in unterschiedlichen Schreibweisen auftreten können (z.B. „m/s“ oder „m s^{-1}“), werden sie nicht als freie Textangaben modelliert. Stattdessen werden Einheiten als IRIs referenziert, die auf die *QUDT*-Ontologie (Quantities, Units, Dimensions, and

[4] In diesem Zusammenhang bezeichnet *hdf:File* nach der HDF-Ontologie den Container für gruppierte, multidimensionale Daten entsprechend der HDF-Spezifikation. Die Datei im Sinne einer Veröffentlichungseinheit sollte hingegen über *dcat:Distribution* modelliert werden.

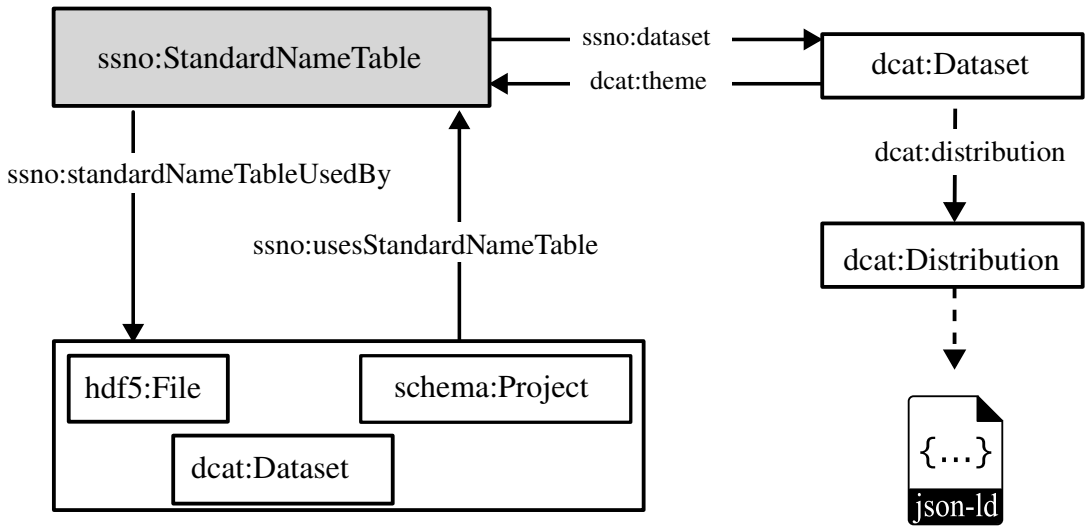


Abbildung 4.7: Eine Standardnamentabelle kann verschiedenen Quellen zugeordnet werden, darunter *hdf:File*, *dcat:Dataset* und *schema:Project*. Die wechselseitige Beziehung zwischen diesen Quellen und der Standardnamentabelle wird durch die Eigenschaften *usesStandardNameTable* bzw. *standardNameTableUsedBy* ausgedrückt. Die physische Bereitstellung der Tabelle kann über *ssno:dataset* mit ihrer semantischen Beschreibung verknüpft werden, ohne dass zusätzliche Eigenschaften erforderlich sind.

Types) verweisen. QUDT bietet eine umfassende, standardisierte Repräsentation physikalischer Größen und Einheiten in maschinenlesbarer Form (Masters et al., o. D.), wodurch eine eindeutige Konvertibilität und Vergleichbarkeit von Einheiten gewährleistet wird.

Definition domainspezifischer Namen

Neben Standardnamen kann es in bestimmten Anwendungsfällen notwendig sein, zusätzliche, domänenspezifische Begriffe in einer Standardnamentabelle zu definieren. Hierfür wird die Klasse *DomainConceptSet* eingeführt, die über die Eigenschaft *hasDomainConceptSet* mit der jeweiligen Standardnamentabelle verknüpft ist.

Die Klasse *DomainConceptSet* dient der Abbildung von Begriffslisten (domain concept sets), die zentrale Termini eines spezifischen Anwendungsgebiets zusammenfassen. Eine Instanz dieser Klasse bündelt mehrere Konzepte, die in einem fachlichen Kontext benötigt werden, etwa für disziplinspezifische Erweiterungen oder Anpassungen bestehender Standardnamen. Dieses Konzept gewinnt insbesondere dann an Bedeutung, wenn Standardnamen für besondere Forschungsdomänen präzisiert oder erweitert werden müssen. Weitere Details hierzu finden sich

Listing 4.5: Beispiel einer Beschreibung eines Projekts („Project X“) im Turtle-Format, welches eine Standardnamentabelle verwendet.

```
@prefix dcat: <http://www.w3.org/ns/dcat#> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix schema: <https://schema.org/> .
@prefix ssno: <https://matthiasprobst.github.io/ssno#> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .

<https://example.org/proj-x> a schema:ResearchProject ;
  ssno:usesStandardNameTable <https://doi.org/1234> ;
  schema:name "Project X"@en .

<https://example.org/snt> a ssno:StandardNameTable ;
  skos:prefLabel "Example Standard Name Table"@en .

<https://doi.org/1234> a dcat:Dataset ;
  dcterms:relation <https://example.org/snt> ;
  dcat:distribution <https://doi.org/1234/files/SNT.jsonld> .

<https://doi.org/1234/files/SNT.jsonld> a dcat:Distribution ;
  dcat:downloadURL <https://doi.org/1234/files/SNT.jsonld> ;
  dcat:mediaType "text/turtle" .
```

Listing 4.6: Beispiel einer Standardnamendefinition im TTL-Format.

```
@prefix ssno: <https://matthiasprobst.github.io/ssno#> .
@prefix ex:   <http://example.org/> .

ex:x_velocity a ssno:StandardName ;
    ssno:standardName     "x_velocity" ;
    ssno:unit             <https://qudt.org/vocab/unit/M-PER-SEC> ;
    ssno:description      "Velocity in X direction in the local
   coordinate system."@en ;
    ssno:standardNameTable ex:MyStandardNameTable .
```

im folgenden Abschnitt sowie in Abschnitt 5.3.

Konstruktion und Modifikation von Standardnamen

Die Regeln zur Konstruktion und Transformation von Standardnamen sind in den CF Conventions bislang auf einer zentralen Webseite dokumentiert. Mit der Entwicklung der *SSNO* wird das Prinzip der Standardnamentabelle jedoch als in sich geschlossene, logisch konsistente Einheit verstanden. Daraus ergibt sich die Anforderung, auch diese Regeln ontologisch zu modellieren. Auf diese Weise wird die Bildung neuer Standardnamen maschinenlesbar, überprüfbar und unabhängig von externen Dokumentationen.

Zu diesem Zweck werden die beiden Klassen *ssno:Qualification* und *ssno:Transformation* definiert, die Unterklassen von *ssno:StandardNameModification* sind. Sie stehen über die Eigenschaft *ssno:definesStandardNameModification* mit der jeweiligen Tabelle in Beziehung.

Das Prinzip lässt sich am Beispiel des Standardnamens *air_pressure_at_mean_sea_level* illustrieren. Der Kernbegriff *pressure* fungiert als Basisstandardname und ist in der Tabelle definiert. Die Zusätze *air* und *at_mean_sea_level* stellen Qualifikationen dar, die den Basisbegriff weiter spezifizieren:

`air`_**pressure**_`at_mean_sea_level` .
„medium“ „at-surface“

Der so gebildete Name ist ein valider Bestandteil der Standardnamentabelle, auch wenn er nicht explizit aufgeführt ist, und kann zur Beschreibung einer physikalischen Größe verwendet werden. Eine Qualifikation definiert ihre zulässigen Werte über die Eigenschaft *hasValidValues*. Für die Qualifikation *medium* sind dies beispielsweise *air* und *water*, wodurch sich Luft- bzw. Wasserdruck eindeutig beschreiben lassen. Die Qualifikation *at_surface* spezifiziert hingegen die Bezugsfläche, etwa die Meereshöhe (*mean_sea_level*).

Qualifikationen können sowohl vor als auch nach dem Standardnamen auftreten. Dies wird durch die Eigenschaften *before* und *after* formalisiert. Eine schematische Darstellung sämtlicher Eigenschaften findet sich in Abbildung 4.8.

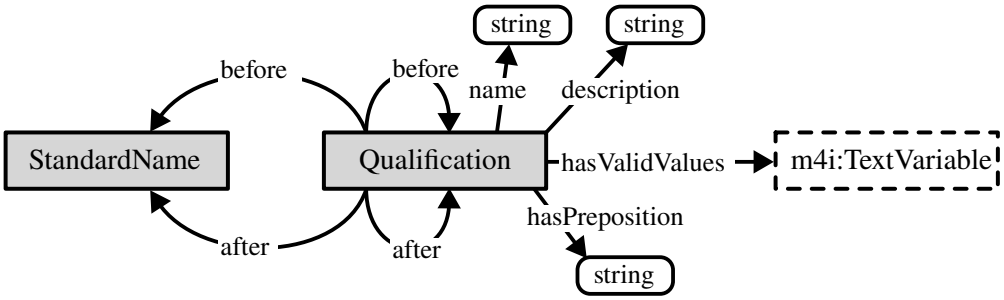


Abbildung 4.8: Die Klasse Qualifikation mit ihren Eigenschaften.

Mit sogenannten Transformatoren (*Transformation*) können neue Standardnamen aus bestehenden Begriffen abgeleitet werden. Sie dienen dazu, mathematische Operationen oder Verknüpfungen explizit innerhalb der Ontologie abzubilden und damit auch komplexe, abgeleitete Größen formal zu beschreiben. Ein einfaches Beispiel ist der Standardname *arithmetic_mean_of_X*. Dabei fungiert *X* als Platzhalter für einen bestehenden Standardnamen, eine Qualifikation oder ein Domainkonzept. Durch die Transformation wird der Mittelwert des gewählten Begriffs eindeutig spezifiziert.

Der zentrale Unterschied zu Qualifikationen besteht darin, dass Transformationen die Einheit eines Standardnamens verändern können. Während Qualifikationen lediglich eine zusätzliche

semantische Spezifizierung vornehmen, kann sich durch Transformationen eine neue physikalische Dimension ergeben. Dies wird in der Ontologie über die Eigenschaft *altersUnit* modelliert. So bleibt beim arithmetischen Mittel die Einheit unverändert (d. h. „[X]“), wohingegen eine Transformation, die das Verhältnis zweier Standardnamen X und Y beschreibt, eine neue Einheit der Form „[X]/[Y]“ erzeugt.

Auf diese Weise ermöglichen Transformationen eine systematische, konsistente und überprüfbare Erweiterung von Standardnamen um mathematisch definierte Größen. Eine Übersicht aller relevanten Eigenschaften und ihrer Beziehungen ist in Abbildung 4.9 dargestellt.

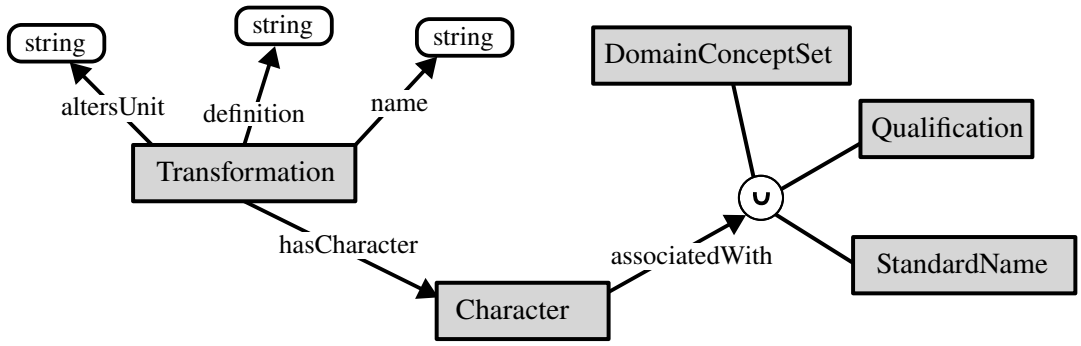


Abbildung 4.9: Die Klasse Transformation mit ihren Eigenschaften.

Durch die Einführung der beiden Klassen ist die systematische Neubildung von Standardnamen formal definiert. Damit lässt sich prinzipiell überprüfen, ob ein in einer Datei verwendeter Standardname den in einer bestimmten Standardnamentabelle festgelegten Regeln entspricht. Um diesen Prozess für Anwender praktikabel zu gestalten und das im Datenmanagementkonzept verankerte Prinzip der Handhabbarkeit zu wahren, wurde eine Softwarelösung entwickelt, die Standardnamentabellen automatisch interpretieren und die Verifizierung übernehmen kann. Detaillierte Ausführungen zu dieser Implementierung dokumentieren Probst (2025) und finden sich in Unterabschnitt A.2.2.

Einbettung in existierende Ontologien

Die Wiederverwendung und Anbindung bestehender Ontologien zählt zu den zentralen Prinzipien der Ontologieentwicklung, da sie die Interoperabilität zwischen unterschiedlichen Datenquellen und Systemen erheblich erleichtert. Die *Simple Standard Name Ontology* orientiert sich daher konsequent an etablierten Vokabularen und integriert bewährte Standards wie *SCHEMA*, *SKOS*, *DCTERMS*, *RDFS* sowie die bereits eingeführten technischen Ontologien *QUDT* und *DCAT*.

Darüber hinaus stellt sich die Frage, in welchem Umfang eine Anbindung an weitere Ontologien sinnvoll und technisch umsetzbar ist. Da Standardnamen in der Regel physikalischen oder numerischen Größen zugeordnet werden, bietet sich insbesondere eine Integration mit Ontologien an, die diese Größen explizit beschreiben. Für die in dieser Arbeit verfolgte Zielanwendung ist vor allem die HDF-Ontologie relevant, da sie eine nahtlose Verknüpfung zwischen semantisch

annotierten Metadaten und der zugrunde liegenden Datenstruktur ermöglicht. Die *SSNO* ist so konzipiert, dass sie flexibel mit weiteren Ontologien kombiniert werden kann, darunter die European Materials Modelling Ontology (*EMMO*), die Ontology *M4I*, *SOSA* sowie *PMS-II* (vgl. Abbildung 4.6). Diese Offenheit gewährleistet die Anschlussfähigkeit der *SSNO* in unterschiedlichsten wissenschaftlichen Kontexten. Im Rahmen dieser Arbeit ergeben sich insbesondere zwei zentrale Anknüpfungspunkte zu anderen Konzepten: Erstens ermöglicht die Eigenschaft *ssno:usesStandardNameTable*, Projekte, HDF-Dateien oder Datensätze (im Sinne einer veröffentlichten Ressource nach *DCAT*) mit einer Standardnamentabelle zu verknüpfen. Auf diese Weise können verwendete Standardnamen mit der referenzierten Tabelle abgeglichen werden. Zweitens können Standardnamen auch direkt Konzepten zugeordnet werden, die numerische Daten repräsentieren. Hierfür sieht die *SSNO* insbesondere Klassen aus gängigen Ontologien wie *EMMO*, *HDF5*, *M4I* oder *SI* vor.

Da sich das in dieser Arbeit entwickelte Datenmanagementkonzept primär auf HDF5 konzentriert, liegt ein besonderer Fokus auf dessen Integration. Die Standardnamentabelle kann mit HDF-Dateien über die Eintrittsgruppe (Root, „/“) verknüpft werden, deren Attribute als Beschreibung der gesamten Datei interpretiert werden. Semantisch entspricht dies einer Zuordnung der Tabelle zur Klasse *hdf:File*. Damit ist es möglich, Standardnamen in den Attributen von HDF-Datasets als einfache Zeichenketten zu hinterlegen, die implizit mit der referenzierten Tabelle verknüpft sind. Abbildung 4.10 illustriert beide Varianten der Zuordnung.

Neben der Integration des Konzepts in HDF-Dateien ist auch die Anschlussfähigkeit an weitere, in den Ingenieurwissenschaften relevante Ontologien von zentraler Bedeutung. Besonders hervorzuheben ist hierbei die Ontologie *M4I*, die im Rahmen von NFDI4Ing entwickelt wird und sich auf die Modellierung numerischer und experimenteller Variablen sowie deren Kontext in ingenieurwissenschaftlichen Datenmodellen konzentriert. Durch die Kombination mit der *SSNO* entsteht ein komplementäres Zusammenspiel. *M4I* modelliert die Variableneigenschaften (Wert, Einheit, Kontext), während *SSNO* den Variablennamen standardisiert. Damit wird *M4I* um eine Ebene erweitert: Eine Variable, die als *m4i:NumericalVariable* mit bestimmtem Wert und Unit definiert wird, kann zusätzlich über *ssno:hasStandardName* auf einen standardisierten Namen referenzieren, der dessen semantische Bedeutung über domänenübergreifende Konventionen dokumentiert.

Das Beispiel in Listing 4.7 verdeutlicht diese Integration. Die Variable *xvel* wird zunächst als *m4i:NumericalVariable* modelliert und mit einem numerischen Wert sowie einer Einheit beschrieben. Über die Eigenschaft *ssno:hasStandardName* wird die Variable zusätzlich mit einem Standardnamen aus der *SSNO* verknüpft. Damit wird nicht nur sichergestellt, dass der Variablenname semantisch eindeutig und maschinenlesbar ist, sondern auch die Grundlage für domänenübergreifende Interoperabilität geschaffen.

Limitierungen

Die *SSNO* formalisiert das Konzept standardisierter Variablennamen und überführt zentrale Prinzipien der CF Conventions in eine maschinenlesbare ontologische Repräsentation. Dabei verfolgt die Ontologie bewusst einen pragmatischen, kompakten Modellierungsansatz, der die

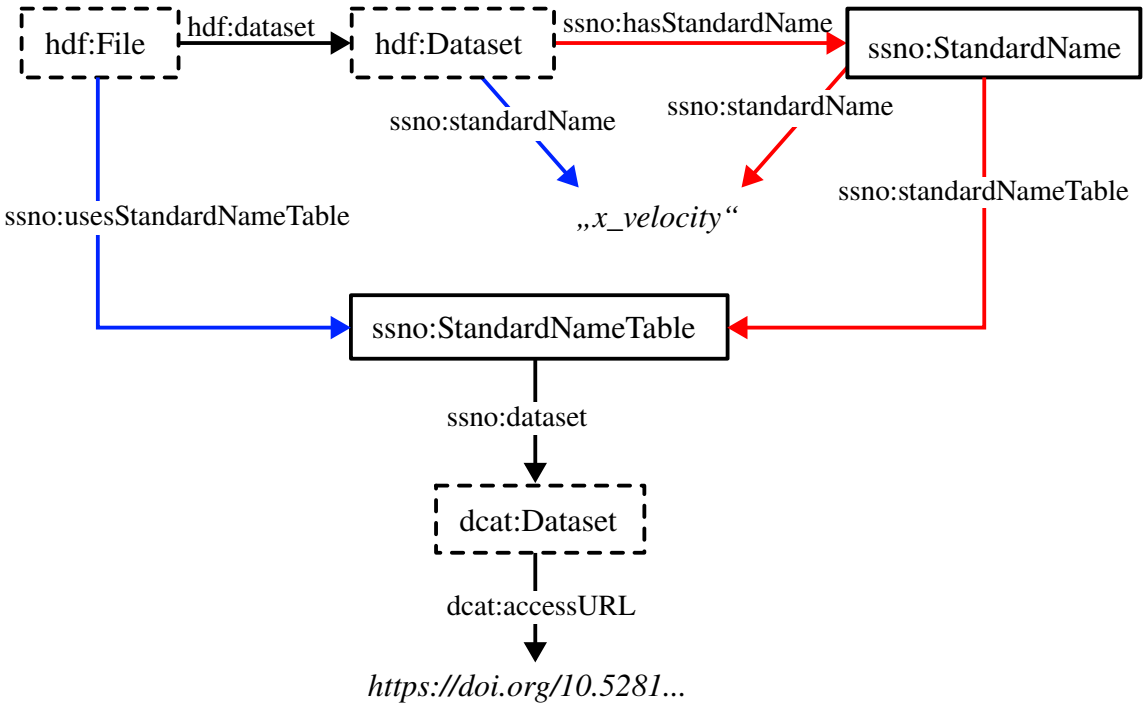


Abbildung 4.10: Beispiel zweier Möglichkeiten, Standardnamen mit ihren Tabellen mittels *SSNO* in einer HDF Datei zu verwenden. Links: Ein HDF5 Dataset wird mit einem Standardnamenobjekt assoziiert, welches auf eine Standardnamentabelle verweist. Rechts: Das HDF5 Dataset zeigt über *ssno:standardName* auf den Standardnamen, der als Text den Attributwert darstellt. Der Rückschluss auf den Standardnamen aus der Tabelle erfolgt über die Beziehung zwischen *hdf:File* und der Tabelle

strukturierte Beschreibung wissenschaftlicher Variablen im Sinne der FAIR-Prinzipien ermöglicht, ohne die Komplexität umfassender Fachontologien zu übernehmen. In ihrer aktuellen Form bestehen jedoch einige konzeptionelle und technische Einschränkungen.

Erstens basiert das Standardnamenkonzept weiterhin auf der linearen Kodierung semantischer Information innerhalb einer Namensstruktur. Obwohl Qualifikationen und Transformationen ontologisch modelliert sind, bleibt die vollständige grammatische Struktur zusammengesetzter Standardnamen nur eingeschränkt formalisiert. OWL2 ist als auf Beschreibungslogik basierende Sprache nicht darauf ausgelegt, rekursive oder positionsabhängige Namensstrukturen vollständig abzubilden (W3C OWL Working Group, 2012).

Dies zeigt sich beispielsweise beim Standardnamen

difference_of_wall_static_pressure_between_point_c_and_point_a.

Dessen Bedeutung ergibt sich nicht allein aus den enthaltenen Begriffen, sondern auch aus deren Reihenfolge innerhalb des Namens. Während Transformationen die systematische Ableitung neuer Standardnamen ermöglichen, kann eine vollständige grammatische Validierung solcher

```
@prefix : <http://example.org/> .
@prefix m4i: <http://w3id.org/nfdi4ing/metadata4ing#> .
@prefix ssno: <https://matthiasprobst.github.io/ssno#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .

:xvel
    rdf:type m4i:NumericalVariable ;
    m4i:hasNumericalValue 4.5 ;
    m4i:hasUnit "km/h" ;
    ssno:hasStandardName :std_x_vel .

:xvel_sn
    rdf:type ssno:StandardName ;
    ssno:standardName "x_velocity" ;
    ssno:unit "m/s" ;
    ssno:description "The velocity in direction of the x coordinate" .
```

Listing 4.7: Beispiel für die Beschreibung einer numerischen Variable aus der *M4I*-Ontologie in Kombination mit *SSNO*.

Strukturen nicht allein durch die Ontologie erfolgen und erfordert ergänzende regelbasierte oder softwaregestützte Verfahren.

Zweitens werden Transformationen konzeptionell definiert und können auch Änderungen der Einheit eines Standardnamens beschreiben. Für Operationen wie *derivative_of_X_wrt_Y* wird beispielsweise erwartet, dass die resultierende Einheit dem Quotienten der Einheiten der Operanden entspricht ($[X]/[Y]$). In OWL2 kann jedoch lediglich modelliert werden, dass X die Einheit U_1, Y die Einheit U_2 und das Ergebnis die Einheit U_3 besitzt. Eine automatische algebraische oder dimensionsanalytische Überprüfung dieser Beziehung ist innerhalb der Ausdrucksmächtigkeit beschreibungslogischer Ontologien nicht möglich, sodass die Einheitenspezifikation deklarativ bleibt.

Darüber hinaus beschränkt sich das Konzept der Standardnamen auf die Beschreibung physikalischer oder mathematischer Größen, für die eine kanonische Einheit angegeben werden kann. Abstraktere Konzepte oder Prozessbeschreibungen (z. B. *transient_simulation*) lassen sich damit nicht direkt als Standardnamen modellieren, obwohl eine solche Bezeichnung in bestimmten Kontexten sinnvoll erscheinen kann. Für derartige Begriffe müssen daher ergänzende Metadatenstrukturen oder Ontologien verwendet werden.

Schließlich wurden die CF Conventions nicht vollständig in die Ontologie überführt. Die Modellierung konzentriert sich bewusst auf die zentralen Komponenten des Standardnamenkonzepts, insbesondere auf Standardnamen, Qualifikationen und Transformationen. Weitere Elemente der CF Conventions, etwa „cell_methods“, „flags“ oder Attribute wie „valid_range“, werden derzeit nicht ontologisch abgebildet. Dadurch bleibt die Ontologie konzeptionell übersichtlich, während spezifische Anwendungsfälle weiterhin durch ergänzende Metadatenstrukturen modelliert werden können.

Die genannten Einschränkungen ergeben sich bewusst aus dem gewählten pragmatischen Entwurf. Ziel der *SSNO* ist nicht die vollständige formale Modellierung physikalischer Prozesse, sondern die Bereitstellung eines schlanken semantischen Referenzsystems für Variablennamen, das eine interoperable und FAIR-konforme Beschreibung wissenschaftlicher Daten ermöglicht.

5 Anwendung auf eine Validierungsdatenbank eines generischen Radialventiators

In den vorangegangenen Kapiteln wurde ein generisches, FAIR-orientiertes Konzept für das Management ingenieurwissenschaftlicher Forschungsdaten entwickelt und dessen methodische Bausteine eingeführt. Ziel dieses Kapitels ist die methodische Validierung dieses Konzepts, indem dessen praktische Anwendbarkeit, Konsistenz und daraus resultierende FAIR-Operationalisierung[1] anhand eines praxisnahen, heterogenen Anwendungsfalls systematisch instanziiert und überprüft wird. Als Validierungsbeispiel dient eine Forschungsdatenbank für einen generischen Radialventilator, die experimentelle und numerische Daten aus unterschiedlichen Quellen in einem gemeinsamen Datenraum zusammenführt.

Der gewählte Anwendungsfall ist in mehrfacher Hinsicht anspruchsvoll und eignet sich daher besonders zur Überprüfung der Tragfähigkeit des entwickelten FDM-Ansatzes. Die betrachteten Daten sind multimodal, da sie auf unterschiedlichen Erhebungs- und Modellierungsverfahren beruhen und sich grundlegend in Struktur, zeitlicher Auflösung und semantischer Bedeutung unterscheiden. Dazu zählen unter anderem skalare Betriebspunktmessungen, feldbasierte PIV-Daten, numerische Simulationsergebnisse aus CFD-Rechnungen sowie geometrische Beschreibungen des Ventilators. Gerade diese Heterogenität stellt eine typische Herausforderung für FAIR-konformes Forschungsdatenmanagement dar, da die Daten unterschiedliche zeitliche und semantische Granularitäten aufweisen, die in klassischen, dateibasierten Ansätzen nur unzureichend gemeinsam adressiert werden können.

Ziel dieses Kapitels ist es zu zeigen, wie die in dieser Arbeit entwickelten methodischen Bausteine, insbesondere HDF5 als zentrales Datenformat, ontologiebasierte Metadatenmodelle, domänenspezifische Standardnamen sowie softwaregestützte Validierungs- und Zugriffskonzepte, unter diesen Randbedingungen konsistent instanziiert, miteinander verknüpft und überprüfbar angewendet werden können. Der Fokus liegt dabei nicht auf der vollständigen Abdeckung aller denkbaren Anwendungsfälle, sondern auf der nachvollziehbaren und reproduzierbaren Umsetzung der FAIR-Prinzipien in einem realen ingenieurwissenschaftlichen Kontext.

Zu Beginn erfolgt eine Einordnung des Anwendungsbeispiels sowie die Darstellung des zugrunde liegenden Umsetzungskonzepts. Diese Vorgehensweise ist erforderlich, da das entwickelte FDM bewusst als generischer Rahmen konzipiert ist und für die Anwendung auf konkrete Forschungsdaten projektspezifische Spezifikationen notwendig sind. Hierzu zählen unter anderem die Definition einer domänenspezifischen Standardnamentabelle, die Festlegung des strukturellen Aufbaus der HDF5-Dateien sowie gegebenenfalls die Entwicklung geeigneter Konvertierungsstrategien für bestehende Datenbestände. In den folgenden Abschnitten werden diese Aspekte getrennt nach den relevanten Datenquellen betrachtet, um die praktische Umsetzbarkeit des Konzepts über verschiedene Datentypen hinweg systematisch zu überprüfen.

[1] FAIR-Operationalisierung bezeichnet hier die konkrete, technisch überprüfbare Umsetzung der FAIR-Prinzipien in Form von Datenstrukturen, Metadatenmodellen, Validierungsregeln und Zugriffsmechanismen.

5.1 Einordnung und Umsetzungskonzept

Numerische Strömungssimulationen basieren auf einer Vielzahl von Annahmen, insbesondere im Bereich der Turbulenzmodellierung, die eine der zentralen Herausforderungen der Strömungsmechanik darstellen.
Strömungsmaschinen, die als Energiewandler eine Schlüsselrolle im Kontext von Energieeffizienz und Klimaschutz einnehmen, stehen dabei besonders im Fokus. Da experimentelle Methoden häufig an ihre Grenzen stoßen, wenn komplexe Strömungsphänomene innerhalb von Bauteilen untersucht werden sollen, bleibt die Validierung numerischer Simulationen trotz steigender Rechenkapazitäten unverzichtbar.

Für hydraulische Strömungsmaschinen existieren derzeit nur wenige öffentlich zugängliche Datenquellen. Speziell für Radialventilatoren fehlen frei zugängliche Validierungsdaten vollständig. Zwar stellt die ERCOFTAC-Datenbank verwandte Fälle bereit, diese entsprechen jedoch weder den Anforderungen moderner, FAIR-orientierter Forschungsdatenbanken noch sind alle Inhalte uneingeschränkt zugänglich (vgl. Unterabschnitt 2.3.3).
Um die Genauigkeit und Aussagekraft numerischer Simulationen für Radialventilatoren fundiert bewerten zu können, wurde daher am Institut für Thermische Strömungsmaschinen ein spezieller Radialventilatorprüfstand entwickelt, aufgebaut und eingehende experimentell untersucht.

Die Umsetzung des zuvor entwickelten FDM-Konzepts erfolgt auf Dateiebene entlang dreier zentraler Aspekte unter Berücksichtigung der unterschiedlichen Anwendungskontexte wie Betriebspunktvermessungen, PIV-Daten, CFD-Simulationen oder geometrische Entwurfsdaten:

1. der strukturellen Organisation und formalen Beschreibung heterogener Primärdaten in HDF5 unter Berücksichtigung der jeweiligen Anwendungskontexte,
2. der semantisch eindeutigen Modellierung physikalischer Größen, Randbedingungen und Ergebnisparameter mittels Ontologien und Standardnamen sowie
3. der konsequenten Publikation der Daten- und Softwareartefakte mit DOIs unter Einbezug semantischer Vernetzung mittels Wikidata.

Zur semantischen Beschreibung physikalischer Parameter und Ergebnisgrößen wird die Standardnamenontologie *SSNO* herangezogen. Für jeden Anwendungsbereich sind hierzu geeignete Standardnamentabellen zu spezifizieren.
Für experimentelle Bildfeldmessverfahren wie Particle Image Velocimetry (PIV) reicht die alleinige Nutzung von *SSNO* jedoch nicht aus, da zusätzlich methodische, apparative und prozessuale Informationen formal beschrieben werden müssen. Zu diesem Zweck wurde ergänzend die Ontologie *PIVMeta* entwickelt. Da PIV-Daten nicht Bestandteil der in dieser Arbeit betrachteten Validierungsdatenbank sind und eine detaillierte Darstellung der Ontologie den Rahmen der vorliegenden Ausführungen überschreiten würde, wird an dieser Stelle auf eine weitergehende Behandlung verzichtet. Die Ontologie ist eigenständig dokumentiert und veröffentlicht und wird hier lediglich als konzeptionell verwandter, jedoch bewusst ausgeklammerter Baustein benannt. Insbesondere komplexe experimentelle Verfahren wie PIV zeigen ein hohes Potenzial für eine zukünftige FAIR-orientierte Formalisierung. Die im Rahmen der Ventilatordatenbank entwi-

ckelten Ontologie *PIVMeta* (Probst und Pritz, 2025b) kann hierbei als Ausgangspunkt für eine zukünftige systematische FAIR-Operationalisierung experimenteller Strömungsdaten dienen.

Abbildung 5.5 visualisiert das Konzept entlang des Forschungsdatenzyklus für den Anwednungsfall dar. Im Zentrum stehen fünf klassische Phasen, beginnend mit der Planungsphase, die die obigen Standards und Regeln (bspw. SHACL) für die Untersuchung definieren sowie allgemeine Metadaten bereitstellen. Im Schritt „Erfassen“ werden die heterogenen Primärdaten durch experimentelle Datenerfassungen und Simulationen erzeugt, die im Schritt „Konvertieren“ durch Harmonisierung in HDF Primärdaten und RDF Metadaten übersetzt werden. Mittels der spezifizierten Validierungsregeln wird nachfolgend gewährleistet, dass alle Informationen für eine Analyse der Daten vorhanden sind. Diese Datengrundlage kann so mit persistenten Identifikatoren publiziert werden, im Rahmen des Praxisbeispiels auf Zenodo. Dieser Forschungsdatenlebenszyklus ist von der Planung bis zur Publikation eingebettet in Layer, die semantische Daten und Lösungen aus dem „Web of Data“ bereitstellen sowie in Softwarelösungen, die sowohl Individualsoftware als auch öffentliche Lösungen sein können. Letzteres ist im Rahmen dieser Arbeit insbesondere durch die *h5rdmtoolbox* ein zentrales Element.

5.2 Beschreibung des experimentellen Aufbaus

Der untersuchte Aufbau entspricht einem geschlossenen Ventilatorprüfstand. Das prinzipielle Schema ist in Abbildung 5.1 dargestellt. Die Luft wird aus einer Beruhigungskammer, in der sich auch Hilfsventilator und Strömungsgleichrichter befinden, über eine Düse angesaugt. Das für die numerische Simulation relevante Gebiet ist durch „Einlass“ und „Auslass“ gekennzeichnet. An diesen Stellen werden die Temperaturen T_e (Einlasstemperatur) und T_a (Auslasstemperatur) erfasst. Die vom Ventilator erzeugte statische Druckdifferenz wird mittels Wandlochbohrungen an denselben Querschnitten bestimmt und in Abbildung 5.1 als Δp_2 ausgewiesen. Eine weitere Druckdifferenzmessung erfolgt zwischen dem Auslassquerschnitt und einer Messstelle im Inneren der Beruhigungskammer (Δp_1). Der Volumenstrom $\dot{V}$ wird indirekt über die Druckdifferenz $\Delta p_{\dot{V}}$ an einer Messblende bestimmt. Ergänzende Messgrößen sind die Umgebungstemperatur, die Drehzahl sowie das Drehmoment der Antriebswelle. Mechanische Komponenten des Prüfstandes sind in der Abbildung nicht enthalten.

Die Einstellung des Betriebspunktes bei konstanter Ventilatordrehzahl erfolgt durch den kombinierten Einsatz einer Drossel und eines Hilfsventilators. Die Drossel erhöht die Systemverluste und ermöglicht die Einstellung von Betriebspunkten mit geringen Volumenströmen. Der Hilfsventilator hingegen kompensiert Verluste und ermöglicht die Untersuchung von Betriebspunkten mit hohen Volumenströmen.

Mit den beschriebenen Messtechniken lässt sich die charakteristische Ventilatorkennlinie aufnehmen. Zur Erfassung detaillierterer Informationen über Strömungsphänomene wird zusätzlich die planare PIV als optisches Messverfahren eingesetzt. Hierfür sind ein doppelt gepulster Laser und eine Kamera orthogonal zueinander angeordnet. Da sowohl das Spiralgehäuse als auch die Tragscheibe des Ventilators aus Plexiglas gefertigt sind, können zweidimensionale Geschwindigkeitsfelder im Schaufelkanal oder im Spiralgehäuse vermessen werden.

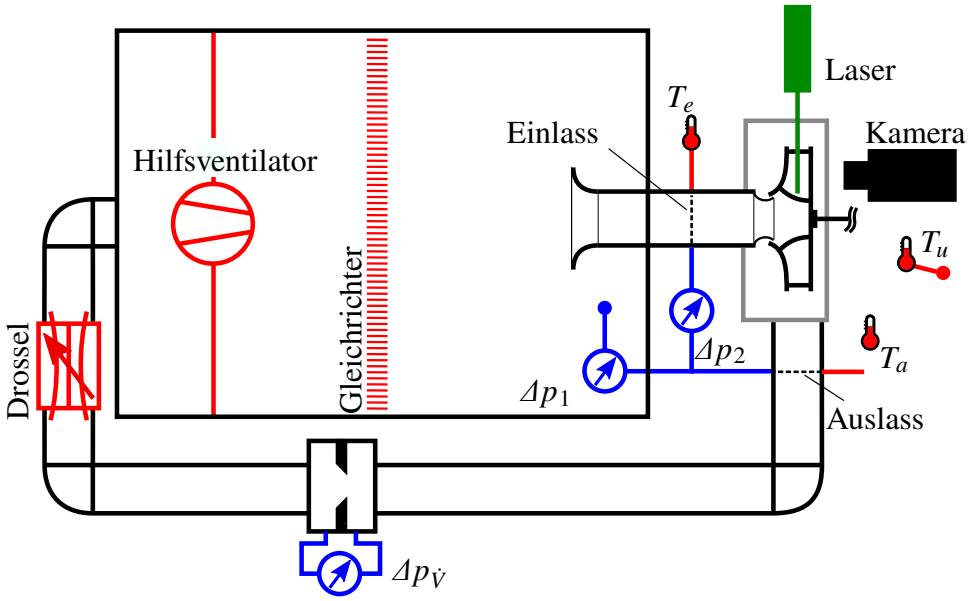


Abbildung 5.1: Prinzipskizze des experimentellen Aufbaus zur Charakterisierung des Radialventilators. Neben Temperatur und Druckdifferenzen ist ein PIV-Aufbau installiert, welcher die Erfassung des Strömungsfelds im Schaufelkanal des Ventilators und der Spirale ermöglicht.

Die Eigenschaften des eingesetzten generischen Radialventilators sind in Tabelle 5.1 zusammengefasst. Eine konstruktive Besonderheit liegt in der Ausführung als Radialventilator mit mitrotierendem Diffusor. Im Unterschied zu konventionellen Radialventilatoren endet der Schaufelkanal hierbei nicht bündig mit der Tragscheibe. Der Durchmesser am Schaufelkanalausgang D_2 entspricht folglich nicht dem Durchmesser der Tragscheibe $D_{2'}$.

Dieses Konstruktionsprinzip entspricht dem einer sogenannten Diffusorpumpe, bei der der Schaufelaustritt ebenfalls nicht mit dem Außendurchmesser des Laufrads übereinstimmt. In der Pumpentechnik wird in diesem Zusammenhang vom *getrimmten Laufraddurchmesser* gesprochen. Dabei werden in der Regel die Schaufelkanten abgedreht, während die Tragscheibe unverändert bleibt Gülich, 2020. Der abweichende Außendurchmesser $D_{2'}$ führt zudem zu unterschiedlichen Auslassbreiten der Strömungskanäle (b_2 bzw. $b_{2'}$), was in strömungsmechanischen Analysen zwingend zu berücksichtigen ist.

An diesem Detail wird exemplarisch die Relevanz einer präzisen, expliziten Parametrisierung und formalen Metadatenbeschreibung im Zuge der Datenerfassung deutlich, da implizite Konventionen oder uneindeutige Annahmen die Vergleichbarkeit und Nachnutzbarkeit der Daten erheblich einschränken können. Sowohl für die theoretische Berechnung des Betriebspunktes im Entwurfsprozess als auch für die spätere Analyse der Ventilatorkennlinie ist es entscheidend, welcher Durchmesser für die Berechnung charakteristischer Kenngrößen herangezogen wird. Da der mitrotierende Diffusor zur Energieübertragung im Laufrad beiträgt, können bei Verwendung von D_2 anstelle von $D_{2'}$, wie in der Literatur oftmals üblich, Abweichungen oder Missverständnisse auftreten. Kenngrößen, in denen der Außendurchmesser eingeht, sind die Reynolds-Zahl,

Größe	Symbol	Wert
Schaufelanzahl	z	9
Schaufelwinkel am Einlass	β_{S1}	30,2°
Schaufelwinkel am Auslass	β_{S2}	40,7°
Schaufelhöhe am Einlass	b_1	58,5 mm
Schaufelhöhe am Auslass	$b_{2'}$	38,2 mm
Kleinster Durchmesser der Düse	D_0	102,5 mm
Durchmesser Schaufeleintritt	D_1	138,0 mm
Durchmesser Schaufelaustritt	D_2	306,0 mm
Laufrad Außendurchmesser	$D_{2'}$	325,0 mm
Spiralbreite	L	146,7 mm
Spiraltiefe	B	147,0 mm
Volumenzahl	$\varphi_{2'}$	0,071
Druckzahl	$\psi_{2'}$	0,958
Schnelllaufzahl	σ	0,275
Durchmesserzahl	$\delta_{2'}$	3,717
Totaldruckerhöhung	Δp_{tot}	390,0 Pa
Totaler Wirkungsgrad	η_{tot}	85,0 %
Nenndrehzahl	n	1500 U/min
Expansionszahl	B_2/b_2	3,85

Tabelle 5.1: Entwurfsparameter und Eigenschaften des untersuchten Radialventilators. Die geometrischen Größen sind mit ihren Symbolen in Abbildung 5.2 referenziert. Der Index 2′ verweist auf die Verwendung von $D_{2'}$ zur Berechnung einer Größe hin. Berechnete Größen basieren auf Gleichungen nach Carolus (2013).

die Druckzahl, die Volumenzahl sowie die Leistungszahl. Diese sind in Tabelle 5.2 aufgeführt, wobei der Außendurchmesser bewusst ohne Index angegeben ist. Der jeweils zugrunde gelegte Durchmesser ist im Rahmen der Betrachtung explizit auszuweisen.

5.3 Konstruktion der Standardnamentabelle

Aufbauend auf der in Abschnitt 4.5 eingeführten Simple Standard Name Ontology (SSNO) wird im Folgenden die Standardnamentabelle für den Radialventilator entwickelt. Sie bildet die Grundlage für die semantische Annotation der physikalischen Parameter des untersuchten Ventilators sowie der im Rahmen von Messungen, Simulationen und Analysen erhobenen Größen. Die Standardnamentabelle dient dabei nicht nur der Vereinheitlichung von Bezeichnungen, sondern stellt ein zentrales Instrument zur technisch überprüfbaren Umsetzung der FAIR-Prinzipien dar. Durch die formale, eindeutige und maschinenlesbare Beschreibung physikalischer Größen

Dimensionslose Kennzahl	**Formel**
Volumenzahl	$\varphi_x = \frac{\dot{V}}{\frac{\pi^2}{4} D_x^3 n}$
Druckzahl	$\psi_{t,x} = \frac{\Delta p_{tot}}{\frac{\pi^2}{2} \rho D_x^2 n^2}$
Schnelllaufzahl	$\sigma_x = \frac{\varphi_x^{\frac{1}{2}}}{\psi_x^{\frac{3}{4}}}$
Durchmesserzahl	$\delta_x = \frac{\psi_{t,x}^{\frac{1}{4}}}{\varphi_x^{\frac{1}{2}}}$
Reynoldszahl	$Re_x = \frac{\pi n D_x^2}{\nu} = \frac{\pi \rho n D_x^2}{\mu}$

Tabelle 5.2: Wichtige Kennzahlen zur Charakterisierung von Strömungsmaschinen und Betriebszuständen. Für den Durchmesser D_x ist der Laufrad- oder Schaufelkanaldurchmesser einzusetzen. Im Falle des mitrotierenden Diffusors ist typischerweise $D_{2'}$ zu wählen.

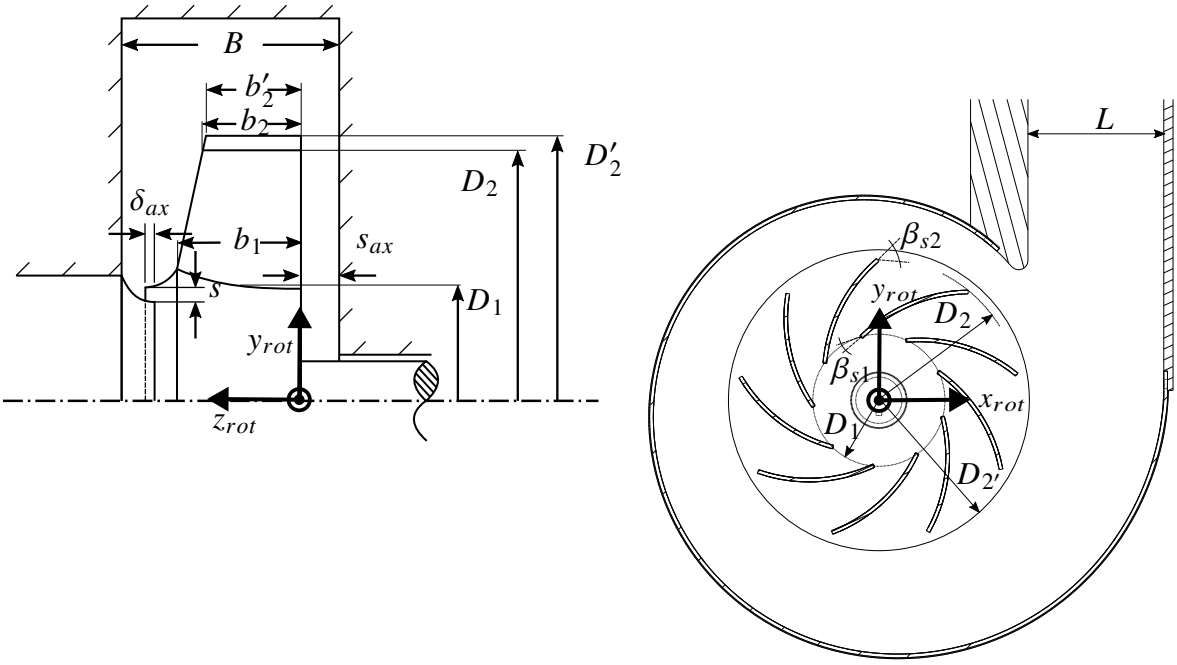


Abbildung 5.2: Schnitte durch den untersuchten Radialventilator. Die charakterisierenden Kenngrößen des Ventilators sind Tabelle 5.1 zu entnehmen.

werden insbesondere Interoperabilität und Nachnutzbarkeit der erzeugten Forschungsdaten sichergestellt.

Die vollständige Standardnamentabelle ist im JSON-LD- und Turtle-Format auf Zenodo veröffentlicht (Probst und Pritz, 2025d). Sie enthält die Standardnamen zur Beschreibung des Ventilators, der Betriebspunkt- und PIV-Messungen sowie der Auswertung von CFD-Simulationen.

Im Folgenden wird zunächst die allgemeine Modellierung dargelegt; eine Auseinandersetzung mit den spezifischen Anforderungen der einzelnen Disziplinen erfolgt anschließend in separaten Unterkapiteln.

Vor der Definition konkreter Standardnamen sind die zugrunde liegenden Konstruktionsregeln festzulegen. Dazu gehören die Identifikation relevanter Anforderungen sowie die Auswahl der physikalischen Größen, die durch standardisierte Metadaten beschrieben werden sollen. Im vorliegenden Anwendungsfall betrifft dies insbesondere thermodynamische und strömungstechnische Kenngrößen, deren Einordnung im Messaufbau (Abbildung 5.1) positions- und bezugssystemabhängig erfolgt.

Es ist daher erforderlich, skalare Größen an charakteristischen Stellen wie definierten Punkten, Flächen oder Schnittstellen zwischen Teilsystemen zu spezifizieren. Ebenso müssen Unterschiede zwischen solchen Messstellen, beispielsweise Druckdifferenzen, abgebildet werden können. Für strömungstechnische Untersuchungen kommt darüber hinaus vektoriellen Größen besondere Bedeutung zu. Hierbei ist sicherzustellen, dass Vektoren einschließlich ihrer Richtungskomponenten im Raum eindeutig benannt und zugeordnet werden können.

Die Gesamtheit der zu beschreibenden Größen und Konzepte ist schematisch anhand eines generischen Strömungsgebietes in Abbildung 5.3 dargestellt.

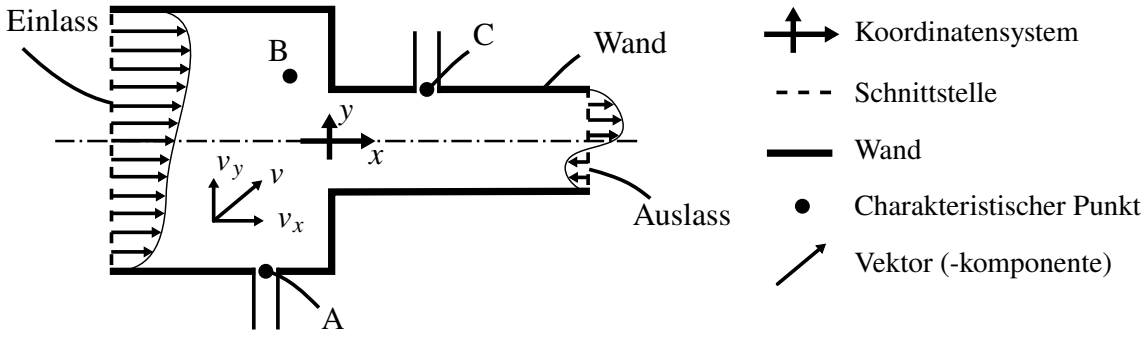


Abbildung 5.3: Schematische Darstellung eines generischen Strömungsgebiets mit charakteristischen Stellen (Punkte, Flächen, Schnittstellen) und einem exemplarischen Geschwindigkeitsvektor v im lokalen Koordinatensystem $(x \mid y)$. Diese Angaben sind sowohl für numerische als auch experimentelle Studien gleichermaßen relevant.

Vor der Definition konkreter Standardnamen sind die zugrunde liegenden Konstruktionsregeln festzulegen, da nur so sichergestellt werden kann, dass die resultierenden Bezeichnungen konsistent, eindeutig interpretierbar und maschinenlesbar sind.

Konstruktionsregeln zur Qualifikation von Standardnamen

Aus den vorangegangenen Überlegungen ergibt sich für den betrachteten Anwendungsfall die folgende Struktur zur Konstruktion von Standardnamen:

```
[surface]
    [component]
        standard_name
            [at location]
                [in medium]
                    [in reference frame]
                        [assuming condition]
```

Die fett hervorgehobene Komponente kennzeichnet den Basis-Standardnamen, während die weiteren Terme als optionale Qualifikationen fungieren. Durch diese modulare Struktur können standardisierte Begriffe je nach Anwendungsfall präzisiert werden. Die Syntax orientiert sich konzeptionell an den CF Conventions, wird jedoch durch die semantische Modellierung mittels *SSNO* erweitert und so um maschinenlesbare und interoperable Eigenschaften ergänzt.

Die Struktur folgt im Kern den „Guidelines for Construction of CF Standard Names“[2]. Aufgrund der inhaltlichen Nähe wurden Konzepte wie die Qualifikationen *surface* oder *component* übernommen. Da sich der vorliegende Anwendungsfall jedoch auf Strömungsmaschinen bezieht, wurden die Qualifikationen erweitert und an die spezifischen Anforderungen der Domäne angepasst. Im Folgenden werden die einzelnen Qualifikatoren näher erläutert:

surface ...

Eine Fläche ist eine geometrische Größe, die nicht durch eine eindeutige Lage im Raum spezifiziert wird. So bezeichnet der Begriff „wall“ eine allgemeine Wand. Diese Angabe steht dem Standardnamen, z. B. einer Geschwindigkeit, voran und bestimmt damit den Wert an dieser Fläche, ohne die exakte Verortung zu spezifizieren. Flächenbezeichnungen bestehen aus einzelnen Wörtern (z. B. *wall_*, *inlet_*, *outlet_*) und stehen am Anfang des Standardnamens. Beispiele sind *wall_velocity* oder *inlet_static_pressure*. Flächen, die durch mehrteilige Phrasen beschrieben werden und eindeutig lokalisiert sind, werden über `[at location]` dargestellt, z. B. *temperature_at_top_of_settling_chamber*.

component ...

Die räumliche Richtung einer Vektorgröße wird durch Koordinatenrichtungen angegeben. In der Standardnamentabelle sind X, Y, Z als mögliche Komponenten definiert. Diese entsprechen den Achsen des als Standard angenommenen Koordinatensystems. Wird ein spezielles Koordinatensystem verwendet, so ist dies durch den Zusatz `[in reference frame]` zu kennzeichnen.

... at location

Charakteristische Stellen im Anwendungsfall können Punkte, Flächen oder Grenzflächen sein. Die Angabe einer Größe an einer solchen Stelle ermöglicht im Gegensatz zu einer allgemeinen

[2] https://cfconventions.org/Data/cf-standard-names/docs/guidelines.html, abgerufen am 13.04.2025

Flächenbeschreibung eine exakte Lokalisierung (vgl. *surface*). Die Standardnamentabelle definiert die für das betrachtete Problem gültigen Punkte. Die konkrete Positionierung muss über ergänzende Metadaten erfolgen, da der Standardname eine konzeptuelle Variable beschreibt. Die Bedeutung bleibt daher in unterschiedlichen Problemfällen identisch, während die konkrete Position im Experiment oder in der Simulation variieren kann.

... in reference frame

Wie bei der Qualifikation *component* erläutert, beziehen sich Vektorkomponenten stets auf ein Koordinatensystem. Ohne Zusatz wird das allgemeine, globale Koordinatensystem vorausgesetzt, das sich in der Regel eindeutig aus dem Kontext ergibt. Soll hingegen ein spezielles Koordinatensystem angegeben werden, etwa ein rotierendes Bezugssystem, erfolgt dies durch ein entsprechendes Suffix. Im Anwendungsfall des Radialventilators ist dies besonders relevant, da Strömungsgrößen im mitrotierenden Bezugssystem des Rotors beschrieben werden. Ein Beispiel hierfür ist *x_velocity_in_rotating_reference_system*. In diesem Fall definiert die Standardnamentabelle das rotierende Koordinatensystem eindeutig durch Orientierung und Ursprung.

... assuming condition

Der Zusatz *assuming_condition* qualifiziert den Standardnamen hinsichtlich seiner Gültigkeit. Der Wert der annotierten Variable gilt ausschließlich unter den angegebenen Annahmen. Typische Beispiele sind *assuming_swirl_free_condition* oder *assuming_incompressibility*. Diese Zusätze stellen sicher, dass die jeweilige Größe eindeutig im Kontext ihrer Annahmen interpretiert wird.

Die im Rahmen des Anwendungsfalls definierten Qualifikationen sind in Tabelle 5.3 zusammengefasst.

Qualifikation	**Werte**
surface	wall, inlet, outlet
component	x, y, z
location	fan_inlet, fan_outlet
medium	air
reference frame	rotating_reference_system
condition	swirl_free_condition, incompressibility

Tabelle 5.3: Definierte Werte der Qualifikationen für die Standardnamentabelle im Anwendungsfall des Radialventilatorprüfstands.

Konstruktionsregeln der Transformationen

Eine zweite Möglichkeit zur Konstruktion neuer Standardnamen stellt die Transformation dar. Darunter ist die Ableitung eines neuen Begriffs durch Anwendung einer mathematischen Operation auf einen bestehenden Standardnamen zu verstehen. Transformationen werden explizit in der Standardnamentabelle definiert (vgl. Unterabschnitt 4.6.2) und können insbesondere mit einer Änderung der physikalischen Einheit einhergehen.

Im Anwendungsfall betreffen diese vor allem Mittelwertbildungen, Standardabweichungen und Differenzbildungen. Eine Übersicht der eingesetzten Transformationen bietet Tabelle 5.4. Die am häufigsten verwendete Operation ist die arithmetische Mittelwertbildung, beispielsweise in Form von *arithmetic_average_of_X*. Ebenso relevant ist die Berechnung von Differenzen zwischen Messstellen, etwa in Form von:

difference_of_wall_static_pressure_between_point_c_and_point_a

Dieser Standardname bezeichnet die Differenz des statischen Wanddrucks zwischen Punkt C und Punkt A gemäß der schematischen Darstellung in Abbildung 5.3. In der Standardnamentabelle ist die zugehörige Transformation derart definiert, dass der Wert an der zweiten Position (Punkt C) minus dem an der ersten Position (Punkt A) gebildet wird. Dies entspricht der mathematischen Schreibweise $\Delta p_{A-C} = p_C - p_A$. Durch die explizite Festlegung der Operandenreihenfolge in der Standardnamentabelle wird ausgeschlossen, dass identisch benannte Differenzen unterschiedlich interpretiert oder berechnet werden.

Ein Sonderfall ist die Transformation *difference_of_X_across_device*, die sich auf Druckverluste über technische Komponenten bezieht, beispielsweise über eine Messblende. In solchen Fällen sind die exakten Messstellen entweder nicht spezifiziert oder nicht von Bedeutung, da normierte Positionen herangezogen werden. Die Standardnamentabelle trägt dem Rechnung, indem sie eine spezifische Liste domänenspezifischer Konzepte bereitstellt (Ontologieklasse *ssno:DomainConceptSet*), die im Kontext der Ventilatoruntersuchung durch die Kategorie *devices* abgebildet wird.

Transformation	**Einheit**	**Beschreibung**
arithmetic average of **X**	[X]	Mittelwerts eines Standardnamens *X*
standard deviation of **X**	[X]	Standardabweichung eines Standardnamens *X*
difference of **X** *and* **Y** *between* [location] *and* [location]	[X]	Differenz zwischen zwei Größen mit Standardnamen *X* und *Y*, die an zwei charakteristischen Stellen *location* vorliegen.
difference of **X** *between* [location] *and* [location]	[X]	Differenz eines Standardnamens *X* zwischen zwei charakteristischen Stellen *location*.
difference of **X** *across* [device]	[X]	Differenz einer Größe über eine ausgezeichnete Komponente. Die genaue Messstelle kann nicht spezifiziert werden oder ist nicht relevant.
ratio of **X** *to* **Y**	[X]/[Y]	Verhältnis eines Standardnamens *X* zu einem Standardnamen *Y*.
derivative of **X** *wrt* **Y**	[X]/[Y]	Ableitung der Größe *X* nach der Größe *Y*.

Tabelle 5.4: Im Anwendungsfall definierte Transformationen zur Konstruktion abgeleiteter Standardnamen. **X** und **Y** stehen für Standardnamen; Angaben in eckigen Klammern kennzeichnen optionale Qualifikatoren.

Konkrete Standardnamentabellen

Für die Radialventilatordatenbank, die sich aus primären experimentellen Messdaten sowie sekundären numerischen Simulationsdaten speist, lassen sich mehrere thematische und disziplinäre Modellierungsbereiche unterscheiden, die jeweils spezifische semantische Anforderungen an die Beschreibung physikalischer Größen stellen. Aus diesem Grund ist eine getrennte Definition von Standardnamentabellen erforderlich.
Ein zentraler Aspekt hierbei ist, dass identische physikalische Größen, z. B. Geschwindigkeiten oder Drücke, in unterschiedlichen Disziplinen nicht zwangsläufig dieselbe inhaltliche Bedeutung besitzen. So beschreibt eine mittels PIV ermittelte Geschwindigkeit eine gemessene, räumlich und zeitlich gemittelte sowie methodenabhängig approximierte Größe, während eine CFD Simulation vorliegende Geschwindigkeit eine modellbasierte, diskretisierte Zustandsgröße darstellt. Obwohl beide Größen über gemeinsame Einheiten verfügen und sich mithilfe etablierter Ontologien wie QUDT als Geschwindigkeitswerte klassifizieren lassen, unterscheiden sie sich grundlegend hinsichtlich ihrer Entstehung, Aussagekraft und Interpretierbarkeit. Weitere Beispiele stellen Drücke dar. Hier ist die präzise Angabe, um welche Drücke es sich handelt (statisch, dynamisch, total) sowie bei Differenzdrücken die Angabe der Messstellen entscheidend. In Simulationen ist es des Weiteren möglich, energetisch gemittelte Größen zu berechnen. Im Experiment ist hingegen typischerweise die indirekte Bestimmung aus Volumenstrom und statischem Druck an einer Wandlochbohrung gegeben. Fehlt hierbei der Kontext, werden ggf. unbeabsichtigt die falschen Größen miteinander verglichen.

Um diese disziplinären Unterschiede explizit abzubilden und semantische Mehrdeutigkeiten zu vermeiden, werden für die betrachteten Datenquellen jeweils eigene, kontextabhängige Standardnamen und damit entsprechende Standardnamentabellen definiert. Für den vorliegenden Anwendungsfall ergeben sich vier Modellierungsbereiche. Auf zwei repräsentative konkrete Tabellen wird nachfolgend eingegangen, die zur Beschreibung eines Radialventilators (Abschnitt 5.4) und auf die zur Erfassung von Betriebspunktmessungen (Abschnitt 5.5).

Die Trennung der Standardnamentabellen nach Modellierungs- bzw. Anwendungsbereichen ermöglicht es, disziplinspezifische Begriffe, Annahmen und Auswerteverfahren explizit zu berücksichtigen, während gleichzeitig durch die einheitlichen Konstruktions- und Transformationsregeln eine konsistente semantische Einordnung über alle Datenquellen hinweg gewährleistet bleibt. Die definierten Standardnamen sind dabei nicht auf die vorliegende Ventilatoruntersuchung beschränkt, sondern als generische Konzepte angelegt, die von der Fachgemeinschaft aufgegriffen, referenziert und erweitert werden können. Dies entspricht einer aus den FAIR-Prinzipien zur Interoperabilität (I1–I3) und Wiederverwendbarkeit (R1, R1.3) abgeleiteten Anforderung, nach der semantische Beschreibungen projektübergreifend konsistent nutzbar sein müssen.

5.4 Beschreibung der Ventilatorauslegung

Die Definition geometrischer und strömungsmechanischer Kenngrößen eines Radialventilators spielt sowohl im Auslegungsprozess als auch bei der Dokumentation experimenteller Messun-

gen oder numerischer Simulationen eine zentrale Rolle. In Lehrbüchern und wissenschaftlichen Publikationen finden sich jedoch unterschiedliche Schreibweisen und Benennungen für identische Größen, was die korrekte Interpretation erschwert und das Risiko von Fehlinterpretationen erhöht. Eine konsistente Standardisierung ist daher für die eindeutige Kommunikation dieser Größen von besonderer Bedeutung.
Die Auflistung von Parametern mit Symbolen, Werten und Einheiten, wie sie beispielsweise in Tabelle 5.1 dargestellt ist, ist in Kombination mit einer Zeichnung für textuelle Veröffentlichungen in der Regel ausreichend. Für die automatisierte Verarbeitung und maschinelle Interpretation digitaler Forschungsdaten ist diese Form der Beschreibung jedoch nicht hinreichend. Ziel der folgenden Ausführungen ist daher nicht die Einführung neuer geometrischer Definitionen, sondern die formale, maschinenlesbare und eindeutig referenzierbare Beschreibung etablierter Kenngrößen mittels standardisierter Metadaten.

Für den vorliegenden Anwendungsfall wird eine Standardnamentabelle spezifisch für den betrachteten generischen Radialventilator entwickelt. Aufgrund der gewählten Abstraktionsebene ist diese Tabelle jedoch nicht auf die konkrete Untersuchung beschränkt, sondern kann auch in anderen Arbeiten zu Radialventilatoren dieser Bauart verwendet werden.
Eine Übertragung des Ansatzes auf andere Ventilatortypen, wie Axialventilatoren, Mischformen oder weitere hydraulische Strömungsmaschinen, erfordert in der Regel Anpassungen oder Erweiterungen der Standardnamentabellen. Insbesondere bei Mischformen und speziellen Schaufelausführungen ist mit zusätzlichen Parametern zu rechnen, die eine differenzierte Modellierung notwendig machen. Die hier vorgenommene Eingrenzung dient daher der exemplarischen Demonstration der Methodik und der kontrollierten Reduktion der Komplexität, ohne den generischen Charakter des Ansatzes einzuschränken.

Abbildung 5.2 zeigt den Meridianschnitt (links) und die Draufsicht (rechts) des generischen Radialventilators, die stellvertretend für Radialventilatoren dieser Bauart gelten, sowie die charakteristischen Größen mit ihren Symbolen. Die Zuordnung der Symbole zu den Standardnamen ist in Tabelle 5.5 aufgeführt. Die vollständige Standardnamentabelle ist auf Zenodo veröffentlicht (Probst und Pritz, 2025e).

Die vollständige Beschreibung des Ventilators ist mithilfe der Standardnamentabelle als eigenständige Metadatendatei abgelegt (vgl. Probst und Pritz (2025a)). Für die Schaufelanzahl ist dies exemplarisch in Listing 5.1 dargestellt. Die Beschreibung erfolgt im Turtle-Format, wobei der Ventilator als *m4i:Tool* gemäß der Ontologie *M4I* modelliert wird. Die Schaufelanzahl wird hierbei als *m4i:NumericalVariable* mit dem Standardnamen *blade_number* beschrieben, der über *ssno:standardNameTable* auf die zugehörige Standardnamentabelle verweist.

Um den untersuchten Ventilator global eindeutig identifizieren und über einzelne Datensätze hinweg referenzieren zu können, wurde ein entsprechender Eintrag in Wikidata (Q131549102) angelegt. Die Metadatendatei ist gemeinsam mit der CAD-Datei des Ventilators auf Zenodo veröffentlicht (Probst und Pritz, 2025a) und verweist explizit auf die Wikidata-Ressource. Diese Verknüpfung unterstützt insbesondere die Interoperabilität und Wiederverwendbarkeit der beschriebenen Forschungsdaten im Sinne der FAIR-Prinzipien, da geometrische Parameter eindeutig referenzierbar und projektübergreifend konsistent nutzbar sind. Darüber hinaus trägt die Anbindung an Wikidata zur verbesserten Auffindbarkeit und zur nachhaltigen Einbettung des

Symbol	Einheit	Standardname
b_1	m	blade_inlet_width
b_2	m	blade_outlet_width
$b_{2'}$	m	impeller_outlet_width
D_1	m	impeller_inlet_diameter
$D_{2'}$	m	blade_outlet_diameter
D_2	m	impeller_outlet_diameter
β_{1b}	rad	blade_angle_at_blade_inlet
β_1	rad	flow_angle_at_blade_inlet
β_{2b}	rad	blade_angle_at_blade_outlet
β_2	rad	flow_angle_at_blade_outlet
L	m	volute_width
B	m	volute_depth
s	m	blade_thickness
s_b	m	blade_gap_depth
s_{ax}	m	impeller_axial_gap
n_b	-	blade_number

Tabelle 5.5: Zuordnung der charakteristischen Geometrieparameter eines Radialventilators zu SI-Einheiten und Standardnamen. Die Symbole und Ableitung der Standardnamen basiert im Wesentlichen auf (Gülich, 2020; Pelz et al., 2022).

beschriebenen Ventilators in bestehende Wissensgraphen der wissenschaftlichen Gemeinschaft bei, ohne eine Abhängigkeit von spezifischen Projekt- oder Softwaresystemen zu erzeugen.

Listing 5.1: Auszug aus der Beschreibung des betrachteten Laufrads im Turtle-Format.

```
@prefix m4i: <http://w3id.org/nfdi4ing/metadata4ing#> .
@prefix dcat: <http://www.w3.org/ns/dcat#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix ssno: <https://matthiasprobst.github.io/ssno#> .
@prefix wd: <http://www.wikidata.org/entity/> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix unit: <http://qudt.org/vocab/unit/> .
@prefix quantitykind: <https://qudt.org/vocab/quantitykind/> .

[] a m4i:Tool ;
    owl:sameAs wd:Q131549102 ;
    m4i:hasParameter [ a m4i:NumericalVariable ;
        rdfs:label "Schaufelzahl"@de ;
        m4i:hasNumericalValue 9 ;
        m4i:hasUnit unit:UNITLESS ;
```

```
        m4i:hasKindOfQuantity quantitykind:Count ;
        ssno:hasStandardName [ a ssno:StandardName ;
            ssno:standardName "blade_number" ;
            ssno:standardNameTable
    <https://doi.org/10.5281/zenodo.14055811> ;
            ssno:unit unit:UNITLESS
        ]
    ] .
```

5.5 Beschreibung von Betriebspunktmessungen

Betriebspunktmessungen beschreiben den (quasi-)stationären Betriebszustand eines Ventilators unter wohldefinierten Randbedingungen. Sie umfassen die Erfassung integraler strömungsmechanischer Größen wie Druckdifferenzen, Temperaturen, Drehmoment und Drehzahl sowie weiterer Messgrößen, die im Messaufbau entweder direkt erfasst oder aus primären Messgrößen abgeleitet werden, beispielsweise der Volumenstrom. Ergänzend zu den unmittelbar gemessenen Größen werden abgeleitete Kenngrößen bestimmt, etwa Wirkungsgrade, Totaldruckdifferenzen oder dimensionslose Kennzahlen.

Für eine fachlich belastbare Interpretation und eine nachhaltige Nachnutzung der Messdaten ist es erforderlich, nicht nur die numerischen Ergebnisse, sondern auch den gesamten messtechnischen Kontext explizit zu dokumentieren. Dazu zählen insbesondere die eingesetzte Sensorik, die Rand- und Umgebungsbedingungen, die angewendeten Auswerte- und Aggregationsverfahren sowie die zugrunde gelegten Messunsicherheiten.

Die Modellierung von Betriebspunktdaten unterscheidet sich grundlegend von der zuvor dargelegten Beschreibung geometrischer oder konstruktiver Parameter. Während letztere zeitlich invariant sind, entstehen Betriebspunktdaten im Rahmen eines Mess- und Auswerteprozesses und besitzen einen expliziten zeitlichen Bezug. Im vorliegenden Anwendungsfall werden die primären Messgrößen daher als Zeitreihen erfasst. Gemeinsam mit den daraus abgeleiteten Kenngrößen sowie mit Metadaten zum untersuchten Ventilator und zum Messaufbau werden sie in einer gemeinsamen Datei veröffentlicht. Dieses Vorgehen stellt sicher, dass die vollständige Datengrundlage für alternative Auswerteverfahren, erneuten Analysen und Vergleichsstudien erhalten bleibt und nicht auf ausschließlich aggregierte Ergebniswerte reduziert wird.

Als zentrales Datenformat wird HDF5 verwendet, wie es das in Abschnitt 4.3 beschriebene Datenmanagementkonzept vorsieht. Die initialen Ausgabeformate der eingesetzten Messsysteme sind demgegenüber häufig einfacher strukturiert. Im vorliegenden Messaufbau kommt ein Datenerfassungssystem von National Instruments zum Einsatz, das die Messdaten zunächst in Form von CSV-Dateien speichert. In einem nachgelagerten Verarbeitungsschritt werden diese Rohdaten in eine HDF5-Datei überführt und dort um eine hierarchische Struktur aus Gruppen, Datensätzen und Attributen ergänzt. Diese Struktur enthält sowohl technische Metadaten zum Messaufbau als auch organisatorische und manuell zu erfassende Kontextinformationen.

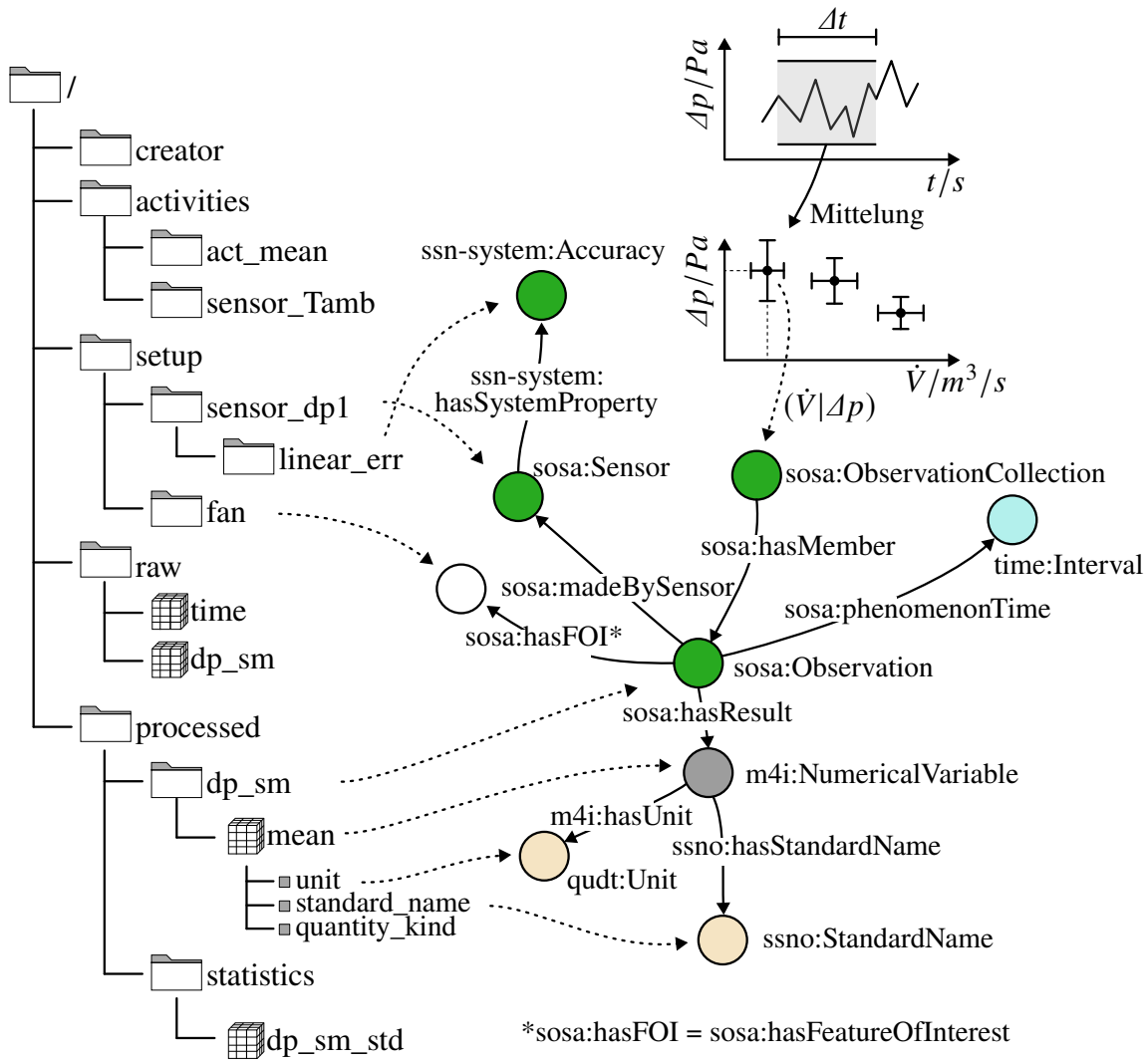


Abbildung 5.4: Schematische Darstellung der inneren Struktur einer HDF5-Datei für eine Betriebspunktmessung (links) und der zentralen semantischen Modellierung (rechts). Die gestrichelten Pfeile weisen auf die Zusammenhänge zwischen Struktur und Kontext hin. Die Darstellung ist nicht vollständig und dient der Illustration wesentlicher Modellierungsentscheidungen.

Abbildung 5.4 stellt die innere Struktur einer HDF5-Datei für eine Betriebspunktmessung (linke Bildhälfte) der zugehörigen semantischen Modellierung (rechte Bildhälfte) gegenüber. Die gestrichelten Pfeile verdeutlichen die Abbildung der strukturellen Elemente der Datei auf ihre semantischen Entsprechungen. Die Darstellung ist bewusst nicht vollständig, sondern hebt zentrale Modellierungsentscheidungen und deren Zusammenspiel hervor, die im folgenden detaillierter ausgeführt werden.

Informationen zur Messdurchführung, zu beteiligten Personen, zur Sensorik sowie zum Messaufbau werden in der HDF5-Datei durch Gruppen und Attribute repräsentiert und über RDF-Tripel

semantisch angereichert. Eine Sensorgruppe wie „/setup/sensor_dp1“ wird beispielsweise als Instanz von *sosa:Sensor* modelliert. Die zugehörigen Primärdaten werden als eindimensionale HDF5-Datensätze gespeichert (vgl. Gruppe „raw“).

Die Erzeugung der Zeitreihen wird semantisch als *prov:Activity* modelliert (in Abbildung 5.4 exemplarisch durch Gruppe „activities“ dargestellt), die mit dem jeweiligen Sensor verknüpft ist. Die Zeitreihen selbst tragen keine explizite numerische Messunsicherheit, sondern verweisen implizit auf den verwendeten Sensor, dessen messtechnische Spezifikation und Unsicherheitsangaben bekannt und separat dokumentiert sind.

Neben den zeitaufgelösten Primärdaten werden aggregierte Kennwerte wie Mittelwerte und Standardabweichungen der Signale in der HDF5-Datei abgelegt. Da skalare HDF5-Datensätze bei der Extraktion der Metadaten erhalten bleiben, sind diese aggregierten Werte Bestandteil der resultierenden Metadatendatei. Dies ermöglicht eine explorative Nutzung der Messergebnisse, ohne dass die potenziell große HDF5-Datei vollständig heruntergeladen werden muss. Mittelwerte von Messgrößen werden dabei unabhängig davon, ob sie direkt gemessen oder aus anderen Größen abgeleitet wurden, einheitlich als *sosa:Observation* modelliert.

Während Größen wie die Druckdifferenz im Messaufbau unmittelbar durch einen Sensor erfasst werden, nimmt der Volumenstrom eine Sonderrolle ein. Er wird im vorliegenden Messaufbau nicht durch einen einzelnen Sensor direkt gemessen, sondern mittels eines Messsystems (*sosa:System*) aus Messblende und Drucksensor bestimmt. Die Ermittlung des Volumenstroms erfolgt auf Grundlage eines definierten Berechnungsverfahrens, das die gemessene Druckdifferenz sowie geometrische und randbedingte Parameter berücksichtigt. Diese Berechnung wird explizit als eigenständige Auswerteaktivität modelliert und innerhalb der Gruppe „activities“ dokumentiert.

Der resultierende Volumenstrom stellt trotz seiner berechnungsbasierten Bestimmung eine beobachtete physikalische Eigenschaft des Ventilators dar. Er ist eine etablierte Messgröße der Strömungsmechanik mit eindeutigem physikalischem Bedeutungsgehalt, definierter Einheit und eigener Messunsicherheit. Entsprechend wird auch der Volumenstrom als *sosa:Observation* beschrieben. Die zugrunde liegenden Berechnungsannahmen, verwendeten Eingangsgrößen sowie die Unsicherheitsfortpflanzung werden dabei explizit dokumentiert.

Die numerischen Ergebnisse der Beobachtungen werden jeweils als *m4i:NumericalVariable* modelliert. Statistische Kenngrößen wie Standardabweichungen werden hingegen nicht als Beobachtungen interpretiert, sondern separat als numerische Variablen beschrieben, da sie das Ergebnis einer statistischen Auswertung darstellen und keine eigenständigen physikalischen Eigenschaften des Ventilators repräsentieren. Sowohl die Mittelwertbildung als auch die Berechnung statistischer Kenngrößen werden explizit über *prov:Activity*-Instanzen dokumentiert.

Für aggregierte Ergebnisgrößen, die einer Unsicherheitsfortpflanzung unterliegen, wird zusätzlich eine explizite Messunsicherheit angegeben. Dies betrifft insbesondere abgeleitete Größen wie den Volumenstrom. Die resultierenden Messunsicherheiten werden nicht auf Ebene der Zeitreihen modelliert, sondern als erweiterte Messunsicherheit über die SIS-Ontologie gemäß *Guide to the Expression of Uncertainty in Measurement* (GUM) an den jeweiligen aggregierten Ergebnisgrößen angegeben und über geeignete Unsicherheitsdeklarationen semantisch beschrie-

ben (Macilenti et al., 2025). Der betrachtete Messzeitraum wird explizit als *time:Interval*[3] erfasst, um eine Interpretation der Betriebspunktgrößen als zeitlich kontextfreie Skalare zu vermeiden.

Alle numerischen Datensätze in der HDF5-Datei werden neben Einheit und Art der physikalischen Größe (*quantity kind*) zusätzlich mit einem Standardnamen annotiert (vgl. HDF Datensatz „dp_sm“ in Abbildung 5.4). Abgeleitete Größen, etwa Mittelwerte oder Differenzen, werden konsistent als Transformationen im Sinne der in Abschnitt 5.3 eingeführten Regeln beschrieben. Dadurch bleiben auch sekundäre Auswerteprodukte eindeutig identifizierbar und maschinenlesbar interpretierbar.

Ventilatorkennlinien werden als *sosa:ObservationCollection* modelliert. Eine Kennlinie umfasst eine Menge von Betriebspunkten, die ihrerseits jeweils als *sosa:ObservationCollection* beschrieben sind. Sowohl die Kennlinie als auch die zugehörigen Betriebspunkte referenzieren über *sosa:featureOfInterest* denselben untersuchten Ventilator, sodass eindeutig festgelegt ist, auf welches physische System sich sämtliche enthaltenen Beobachtungen beziehen.

Die Modellierung einer Ventilatorkennlinie erfolgt erst im Rahmen der Auswertung. Eine einzelne HDF5-Datei einschließlich der daraus extrahierten RDF-Metadaten beschreibt jeweils genau eine Betriebspunktmessung. Die Beschreibung einer Ventilatorkennlinie ist daher nicht Bestandteil einer einzelnen HDF5- oder RDF-Datei. Stattdessen entsteht die Kennlinie durch das gezielte Zusammenführen mehrerer Betriebspunktmessungen über die Datenbankschnittstelle.

In diesem Kontext kann die Datenbank zusätzliches RDF bereitstellen, das die Ventilatorkennlinie als eigenständige *sosa:ObservationCollection* beschreibt. Dieses kennlinienbezogene RDF stellt ein abgeleitetes, auswertungsabhängiges Artefakt dar und ist klar von den dateibasierten Beschreibungen einzelner Betriebspunktmessungen getrennt.

Abschließend sei darauf hingewiesen, dass mit Abbildung 5.4 zwar exemplarisch eine konkrete HDF5-Organisationsstruktur dargestellt wird, die semantische Bedeutung der enthaltenen Daten jedoch nicht an diese physische Ablagestruktur gebunden ist. Die konkrete Ausgestaltung der HDF5-Hierarchie kann sowohl zwischen unterschiedlichen Experimenten als auch zwischen verschiedenen Instituten oder Messkampagnen variieren und unterliegt damit technischen, organisatorischen und historischen Randbedingungen. Eine solche Struktur eignet sich daher nicht als alleinige Trägerin der inhaltlichen Bedeutung der Daten. Es ist die Annotation der RDF-basierten Metadaten, die sicherstellt, dass die fachliche Interpretation der Daten unabhängig von Dateipfaden, Gruppennamen oder hierarchischer Organisation eindeutig und maschinenlesbar erfolgt. Der semantische Kontext bleibt damit stabil definiert, auch wenn sich die physische Organisationsstruktur der HDF5-Datei ändert. Ein konsistenter und sprechender Aufbau der HDF5-Datei bleibt dennoch ein wesentliches Element guter Datenorganisation und unterstützt die menschliche Lesbarkeit sowie die Wartbarkeit der Daten. Die formale, nachnutzbare und institutionsübergreifend konsistente Bedeutungszuweisung wird jedoch ausschließlich durch die RDF-Modellierung gewährleistet.

[3] Vgl. https://www.w3.org/TR/owl-time/#time:Interval

Standardisierte Beschreibung der Betriebspunktgrößen

Die numerischen Betriebspunktgrößen der Ventilatordatenbank werden durchgängig über eine projektspezifische Standardnamentabelle beschrieben, die gemäß den in Abschnitt 5.3 eingeführten Konstruktionsregeln erstellt wurde. Jeder numerische Datensatz wird dabei eindeutig einem Standardnamen zugeordnet. Dies gilt gleichermaßen für primäre Messgrößen, zeitlich aggregierte Kenngrößen sowie für abgeleitete Auswerteprodukte.

Für die Beschreibung von Betriebspunktmessungen ist diese explizite Zuordnung zwingend erforderlich, da zahlreiche physikalische Größen nur im Zusammenspiel aus Messstelle, Bezugsniveau, Druckart und Aggregation eindeutig interpretierbar sind. In der Ventilatordatenbank werden diese Kontexte nicht implizit vorausgesetzt, sondern über die jeweiligen Standardnamen explizit festgelegt. Dadurch wird sichergestellt, dass Betriebspunktgrößen unabhängig von Messkampagne, Prüfstand oder Auswerteverfahren konsistent interpretierbar bleiben.

An dieser Stelle ist hervorzuheben, dass die Kombination aus Einheit und Art (*quantity kind*) allein für die eindeutige Beschreibung vieler Betriebspunktgrößen nicht ausreicht. Zwar erlaubt etwa QUDT die Klassifikation einer Größe als statischer oder totaler Druck, jedoch wird dadurch nicht festgelegt, zwischen welchen Messstellen eine Druckdifferenz gebildet wurde. Die für Betriebspunktmessungen zentrale Information über räumliche Bezugspunkte, wie Ein- und Auslass des Ventilators, ist damit nicht Bestandteil der physikalischen Klassifikation. Eine explizite ontologische Modellierung dieser Zusammenhänge ist prinzipiell möglich, erfordert jedoch einen erheblichen zusätzlichen Modellierungsaufwand und ist für die schnelle Identifikation und Vergleichbarkeit von Betriebspunktgrößen in der praktischen Datenbanknutzung nur eingeschränkt geeignet.

Tabelle 5.6 zeigt einen exemplarischen Auszug der in der Ventilatordatenbank verwendeten Standardnamen für Betriebspunktmessungen. Die zugehörige Standardnamentabelle ist Probst und Pritz (2025d) zu entnehmen.

Abgeleitete Größen wie der Volumenstrom oder der Gesamtwirkungsgrad werden dabei gleichrangig zu direkt gemessenen Größen behandelt. Maßgeblich ist allein, dass der physikalische Bedeutungsgehalt der Größe eindeutig definiert ist. Die konkrete Ermittlung, etwa durch Berechnungen auf Basis mehrerer Eingangsgrößen, wird nicht im Standardnamen selbst kodiert, sondern separat über entsprechende *prov:Activity*-Instanzen dokumentiert. Standardname und Berechnung bleiben damit konzeptionell getrennt, aber eindeutig miteinander verknüpft.

Differenzgrößen, insbesondere Druckdifferenzen zwischen Ein- und Auslass des Ventilators, werden in der Ventilatordatenbank konsistent über zusammengesetzte Standardnamen beschrieben. Für die Totaldruckdifferenz wird beispielsweise der Standardname

difference_of_total_pressure_between_fan_outlet_and_fan_inlet

verwendet. Die zugehörige Transformation ist in der Standardnamentabelle eindeutig definiert, einschließlich der Reihenfolge der Operanden. Dadurch ist festgelegt, welche Messstellen verglichen werden und welches Vorzeichen die Differenz besitzt. Implizite Konventionen, wie sie in klassischen Kennliniendarstellungen häufig anzutreffen sind, werden damit vermieden.

Standardname	Einheit	Beschreibung
temperature	K	Temperatur an einer definierten Messstelle
total_fan_efficiency	–	Gesamtwirkungsgrad des Ventilators, berechnet aus der Totaldruckdifferenz
fan_torque	$N\,m$	Auf den Ventilator übertragenes Drehmoment
fan_volume_flow_rate	m^3/s	Vom Ventilator geförderter Volumenstrom
static_pressure	Pa	Statischer Druck
dynamic_pressure	Pa	Dynamischer Druck
total_pressure	Pa	Totaldruck
air_density	kg/m^3	Dichte der Luft

Tabelle 5.6: Auszug zentraler Standardnamen zur Beschreibung von Betriebspunktmessungen. Im Wesentlichen handelt es sich hierbei um Basisgrößen. Für die Darstellung und Analyse sind Größen wie die Totaldruckdifferenz entscheidend, deren Standardnamen aus den Konstruktionsregeln der Tabelle systematisch abgeleitet werden.

Auch zeitliche Aggregationen werden explizit als Transformationen behandelt und erhalten eigene Standardnamen, etwa *arithmetic_mean_of_volume_flow_rate*. Dadurch bleibt nachvollziehbar, ob ein numerischer Wert einen momentanen Messwert, einen zeitlichen Mittelwert oder ein weiteres Aggregat darstellt. Eine semantische Vermischung von Zeitreihen und aggregierten Betriebspunktgrößen wird auf diese Weise ausgeschlossen.

Die Standardnamentabelle übernimmt damit in der Ventilatordatenbank eine zentrale Rolle bei der formalen Beschreibung der Betriebspunktgrößen. Sie stellt eine effektive Lösung dar, um Forscher eine praktikable Möglichkeit zu bieten, auch komplexe Mess- und Simulationsergebnisse maschineninterpretierbar bereitzustellen, ohne eine hohe semantische Modellierungskomplexität vorauszusetzen. Physikalische Größe, räumliche Referenzen sowie gegebenenfalls Auswerte- oder Aggregationsverfahren werden in einer kompakten, formal definierten Bezeichnung zusammengeführt. Auf diese Weise wird eine konsistente, vergleichbare und automatisiert auswertbare Beschreibung von Betriebspunktmessungen erreicht, ohne dass hierfür domänenspezifische Ontologieerweiterungen erforderlich sind.

5.6 Datenbankentwurf

Das erarbeitete Datenmanagementkonzept aus Kapitel 4 führt im Wesentlichen eine Methodik und Lösung zur Trennung von Primärdaten und semantischen Metadaten ein. Dieses Kapitel beschreibt die konkrete Umsetzung hinsichtlich der Verteilung bzw. Bereitstellung der Daten, die sich das HDF-basierte Konzept zunutze macht.

Statt auf klassische, institutsspezifische Lösungen mit textbasierten Datenbanken und einfachen Webzugängen zurückzugreifen, setzt diese Arbeit auf eine moderne, vernetzte Dateninfrastruk-

tur. Ziel ist es, den steigenden Anforderungen an Nachvollziehbarkeit, Wiederverwendbarkeit und Interoperabilität wissenschaftlicher Daten gerecht zu werden, wie sie insbesondere durch die FAIR-Prinzipien formuliert sind. Durch die konsequente Nutzung semantischer Technologien wird eine flexible und anschlussfähige Struktur geschaffen, die den komplexen Anforderungen der Strömungsmechanik sowie der simulations- und messdatenbasierten Validierung deutlich besser gerecht wird als klassische Ansätze.

Der Begriff *Datenbank* wird in dieser Arbeit bewusst in einem erweiterten Sinne verwendet. Er bezeichnet keine monolithische, zentral betriebene Speicherlösung, sondern eine logische, durchsuchbare Aggregation verteilter Daten- und Metadatenressourcen. Die zugrunde liegenden Daten liegen nicht in einer einheitlichen Datenbankinstanz vor, sondern sind über persistente Repositorien verteilt. Die Zusammenführung zu einer nutzbaren Datenbasis erfolgt erst dynamisch über standardisierte, semantische Beschreibungen und Abfragemechanismen.

Die Mehrheit bestehender Datenbanken, insbesondere im Bereich der Strömungsmechanik, folgt weiterhin zentralisierten, monolithischen Entwürfen (vgl. Abschnitt 2.3). Wie Waagmeester et al. (2020) darlegen, stellt dieses Modell das eine Ende des Spektrums im Daten- und Wissensmanagement dar. In einer solchen Lösung werden Daten und Wissen in einer einzigen, einheitlichen Datenbank zusammengeführt, die einem festen Datenmodell folgt. Dieser Ansatz bietet Vorteile wie hohe Datenkonsistenz und die Möglichkeit präziser und effizienter Abfragen. Gleichzeitig erfordert die Integration neuer Daten einen erheblichen manuellen Aufwand, da diese an bestehende Schemata angepasst werden müssen. Zudem sind zentrale Datenbanken häufig durch institutionelle Grenzen, begrenzte Skalierbarkeit und eingeschränkte Interoperabilität limitiert.

Demgegenüber steht der verteilte Ansatz (Sima et al., 2019; Waagmeester et al., 2020), bei dem Daten und Wissen über verschiedene, voneinander unabhängige Ressourcen verteilt sind. Neue Daten können mit deutlich geringeren Eintrittsbarrieren integriert werden, sofern sie sich an übergeordnete Community-Standards halten. Dieser Ansatz fördert Offenheit und Skalierbarkeit, stellt jedoch hohe Anforderungen an Harmonisierung, Integration und maschinelle Durchsuchbarkeit der Ressourcen.

Die in dieser Arbeit entwickelte Datenbanklösung versteht den Begriff *Datenbank* bewusst als ganzheitliches, jedoch verteiltes System. Sie verfolgt einen hybriden Ansatz, der die Vorteile zentraler und verteilter Modelle kombiniert, ohne deren jeweilige Nachteile vollständig zu übernehmen. Die physische Speicherung der Forschungsdaten erfolgt verteilt über etablierte, öffentliche Repositorien, insbesondere Zenodo, während die logische Integration und Durchsuchbarkeit über semantische Metadaten realisiert wird.

Zentral im Sinne der Datenbank ist dabei nicht die physische Ablage der Daten, sondern ihre semantische Erschließung. Die Kohärenz der verteilten Ressourcen wird nicht durch ein monolithisches Datenbankschema erzwungen, sondern durch die konsequente Nutzung gemeinsamer Ontologien, standardisierter Vokabulare sowie formaler Validierungsmechanismen hergestellt. Die Datenbank entsteht somit erst durch die semantische Verknüpfung und Abfragbarkeit der verteilten Ressourcen und nicht durch deren physische Zusammenführung in einer einzelnen Datenbankinstanz.

Die konzeptionellen Zusammenhänge des in dieser Arbeit entwickelten Datenbankdesigns sind

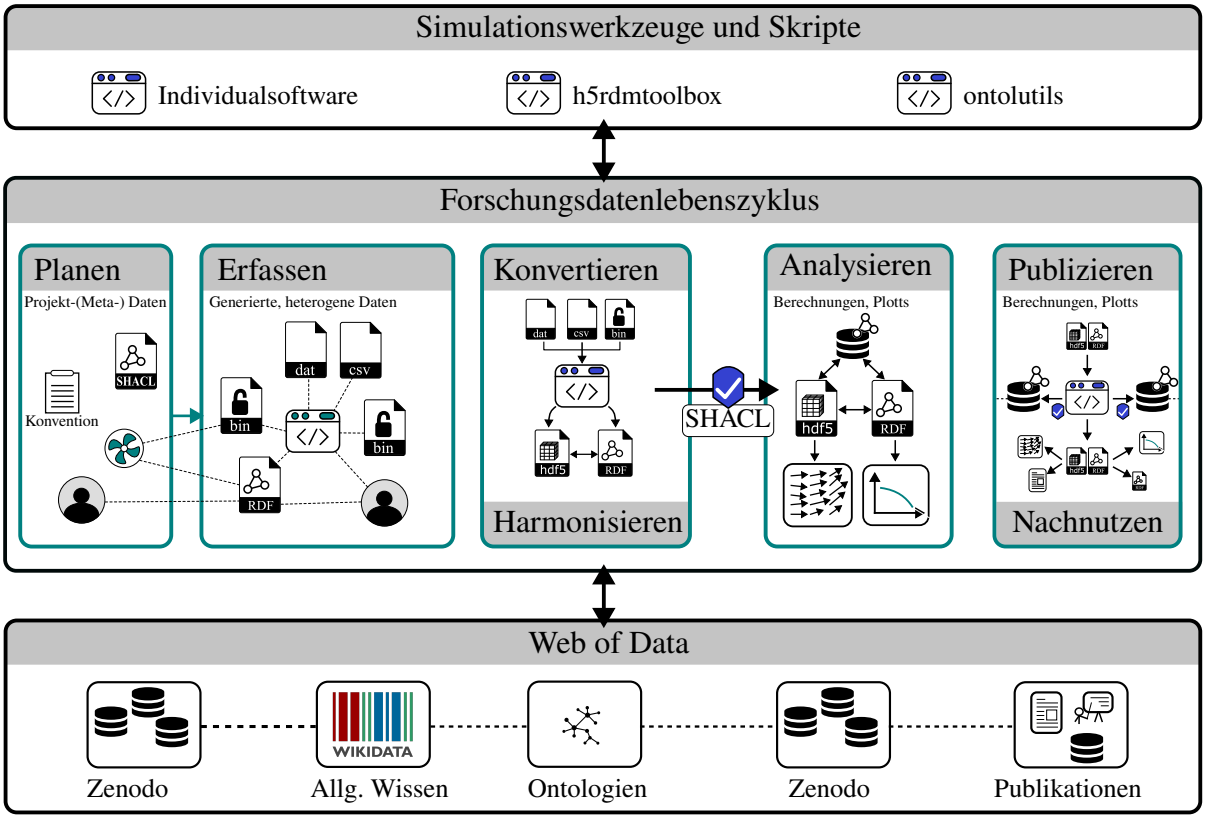


Abbildung 5.5: Forschungsdatenlebenszyklus des FAIRen Forschungsdatenmanagements für ingenieurwissenschaftliche Anwendungen basierend auf HDF5 und RDF. Heterogene Rohdaten aus Messungen und Simulationen werden nach einer projektbezogenen Planungsphase harmonisiert, in das strukturierte HDF5-Format überführt und mit formalen Metadaten (RDF) angereichert. Die Analyse erfolgt nach Validierung mit SHACL auf Basis der HDF5-Daten, während die Publikation kuratierter Datenobjekte mit persistenten Identifikatoren die Anbindung an das Web of Data und eine spätere Nachnutzung ermöglicht. Begleitet wird der Zyklus mit entsprechender Software.

schematisch in Abbildung 5.5 dargestellt. Die Abbildung ordnet das vorgestellte Forschungsdatenmanagement explizit in einen durchgängigen Forschungsdatenlebenszyklus ein und verdeutlicht, dass die Datenbank nicht als isolierte Komponente, sondern als integraler Bestandteil dieses Zyklus zu verstehen ist.

Ausgehend von einer projektbezogenen Planungsphase werden heterogene Rohdaten aus Messungen und Simulationen erfasst und zunächst harmonisiert. Die Harmonisierung umfasst sowohl die strukturelle Vereinheitlichung der Daten als auch deren Überführung in das HDF5-Format, das in der vorliegenden Arbeit als zentrales Trägerformat für die Primärdaten dient. Parallel dazu werden formale Metadaten erzeugt und in RDF serialisiert, wodurch eine semantische Beschreibung der Daten unabhängig von ihrer physischen Speicherung ermöglicht wird.

Die Abbildung verdeutlicht zudem die funktionale Trennung zwischen numerischen Primärda-

ten und semantischer Erschließung. Während Berechnungen, Auswertungen und Visualisierungen direkt auf den HDF5-Daten basieren, erfolgt die Validierung der Metadaten mithilfe von SHACL. Erst auf dieser validierten semantischen Grundlage werden die Daten publiziert und über persistente Identifikatoren in das Web of Data eingebunden.

Die eigentliche Datenbank entsteht dabei nicht durch eine zentrale physische Speicherung der Daten, sondern durch die semantische Verknüpfung und Durchsuchbarkeit der verteilten Ressourcen. Wie in Abbildung 5.5 dargestellt, sind die HDF5-Datenobjekte über Repositorien in Zenodo verteilt abgelegt, während Ontologien, kontrollierte Vokabulare und externe Wissensbasen wie Wikidata zur inhaltlichen Kontextualisierung beitragen. Die Datenbank ist somit als logische, verteilte Infrastruktur zu verstehen, die erst durch standardisierte Metadaten, Validierung und Abfragemechanismen kohärent wird.

Abbildung 5.5 macht darüber hinaus deutlich, dass der vorgestellte Ansatz nicht auf die einmalige Publikation von Daten abzielt, sondern explizit auf deren langfristige Nachnutzung. Die Bereitstellung strukturierter Datenobjekte, formaler Metadaten und begleitender Software schafft die Voraussetzung dafür, dass Betriebs- und Simulationsdaten reproduzierbar analysiert, verglichen und in neue Forschungskontexte integriert werden können.

Ein zentrales Gestaltungsprinzip ist die klare Trennung zwischen primären Forschungsdaten und abgeleiteten semantischen Metadaten. Die wissenschaftlichen Daten selbst werden ausschließlich in HDF5-Dateien gespeichert, die als autoritative Quelle gelten. Sämtliche RDF-Daten werden automatisiert aus diesen HDF5-Dateien extrahiert und stellen keine unabhängige Wissensquelle dar, sondern sind vollständig aus den HDF5-Dateien ableitbar. Die RDF-Repräsentationen dienen vielmehr als semantische Projektion der in den HDF5-Dateien enthaltenen strukturellen und kontextuellen Informationen. Dadurch wird es möglich, den Inhalt großer, binärer Datensätze effizient zu explorieren, ohne diese vollständig herunterladen oder lokal vorhalten zu müssen.

Die eigentliche Nutzung der Datenbank erfolgt über SPARQL-Abfragen auf die aggregierten RDF-Daten. Diese Abfragen können sowohl lokal als auch über föderierte Endpunkte ausgeführt werden[4]. Die Abfrageergebnisse erlauben es, relevante Datensätze zu identifizieren, zu vergleichen und gezielt auszuwählen. Erst im Anschluss werden die zugehörigen HDF5-Dateien heruntergeladen und für weiterführende Analysen verwendet. Die RDF-Schicht fungiert somit als Indexierungs-, Such- und Explorationsschicht über einem verteilten Datenraum. Hierzu führt Abschnitt 5.7 weiter aus.

Für die Nutzer der Validierungsdatenbank sind alle notwendigen Informationen über öffentlich zugängliche Zenodo-Repositorien verfügbar. Innerhalb von Forschungsgruppen können darüber hinaus zusätzliche, nicht öffentliche Datenquellen eingebunden werden, ohne das Gesamtsystem zu verändern. Die Datenbank ist damit nicht als statische Sammlung zu verstehen, sondern als dynamisch konfigurierbarer, erweiterbarer Datenraum.

Eine Gegenüberstellung klassischer Datenbanken und des hier entwickelten Ansatzes zeigt Tabelle 5.7.

[4] SPARQL-Abfragen über verteilte Endpunkte, die sowohl lokal als auch webbasiert sein können; engl. *federated SPARQL queries*.

Kriterium	Klassische Datenbanken	Vorgestellte Lösung
Nachhaltigkeit	Stark von institutioneller Infrastruktur abhängig	Langfristige Verfügbarkeit durch etablierte, öffentliche Infrastrukturen (Zenodo, Wikidata)
Umsetzung FAIR	Oft unvollständig	Systematisch integriert
Kosten	Hoher Wartungs- und Betriebsaufwand	Gering, Nutzung öffentlicher Infrastrukturen
Flexibilität	Starres Schema	Modular, konfigurierbar und erweiterbar
Vernetzung	Lokal und eingeschränkt	Global, interoperabel und verlinkt

Tabelle 5.7: Vergleich klassischer Datenbanken mit dem in dieser Arbeit entwickelten hybriden Ansatz.

Durch die Wahl dieses hybriden Ansatzes wird nicht nur die Interoperabilität zwischen unterschiedlichen Datenquellen maximiert, sondern auch die Integration neuer Daten ohne signifikanten manuellen Aufwand ermöglicht. Die resultierende Infrastruktur ist damit skalierbar, robust und langfristig nutzbar.

Die Entscheidung, die Daten über Zenodo zu veröffentlichen, wurde unter Berücksichtigung mehrerer zentraler Kriterien getroffen. An erster Stelle steht die Wiederverwendbarkeit. Zenodo ermöglicht eine langfristige Bereitstellung[5] der Daten bei gleichzeitig transparenter, zitierfähiger Beschreibung. Die Vergabe persistenter Identifikatoren stellt sicher, dass die Daten eindeutig referenzierbar und dauerhaft auffindbar bleiben.

Ein weiterer zentraler Aspekt ist die Auffindbarkeit. Durch die DOI-Vergabe und die Einbindung in übergeordnete Suchinfrastrukturen erreichen die Daten eine hohe Sichtbarkeit innerhalb der wissenschaftlichen Gemeinschaft. Ebenso wichtig sind die offene Verfügbarkeit und Versionierung. Zenodo erlaubt es, neue Datensätze und Erweiterungen systematisch zu versionieren und ältere Versionen weiterhin zugänglich zu halten.

Im Vergleich dazu erweisen sich klassische Journalpublikationen als ungeeignet für die langfristige Bereitstellung umfangreicher Forschungsdaten. Sie sind häufig mit Zugangsbeschränkungen, Volumenlimits und fehlender Änderbarkeit verbunden. Die hier vorgestellte Datenbanklösung verfolgt bewusst den umgekehrten Ansatz: Die Daten bilden die primäre Referenz, auf die zahlreiche Publikationen aufbauen können.

Auch institutionelle Repositorien sind mit Nachteilen verbunden, insbesondere durch erhöhten administrativen Aufwand und eingeschränkte Sichtbarkeit. Für eine nachhaltige Zitierbarkeit ist die Vergabe persistenter Identifikatoren essenziell, da rein webbasierte Veröffentlichungen keine langfristige Verfügbarkeit garantieren (Klump et al., 2006; Paskin, 2005).

Zenodo erfüllt diese Anforderungen in besonderem Maße. Durch die Integration in die Open-

[5]Zenodo sichert eine Speicherung für die Dauer des Betriebs der Einrichtung CERN zu, was zum Zeitpunkt der Abfassung dieser Arbeit einem Zeitraum von mindestens 20 Jahren entspricht

Access-Initiative, die breite Akzeptanz in verschiedenen Disziplinen und die Bereitstellung maschinenlesbarer Schnittstellen stellt es eine zentrale Säule der hier vorgestellten Infrastruktur dar. Insgesamt erweist sich Zenodo damit nicht nur als geeignete Plattform für die Veröffentlichung der in dieser Arbeit generierten Daten, sondern auch als Blaupause für nachhaltiges, FAIR-konformes Forschungsdatenmanagement in den Ingenieurwissenschaften.

5.6.1 Semantische Vernetzung im Forschungsdatenraum

Die Nützlichkeit wissenschaftlicher Datensätze hängt maßgeblich von ihrer Auffindbarkeit und Zugänglichkeit ab (vgl. die FAIR-Prinzipien F1 *Findable* und A1 *Accessible* in Tabelle 2.1). Die Veröffentlichung von Forschungsdaten in etablierten Repositorien mit DOI-Vergabe stellt hierfür eine notwendige Grundlage dar, da sie Persistenz, Zitierfähigkeit und langfristige Verfügbarkeit gewährleistet.

Das volle Potenzial dieser Daten entfaltet sich jedoch erst, wenn ihre Metadaten zusätzlich in offenen, semantischen Wissensgraphen verankert werden, wie dies beispielsweise mittels Wikidata erfolgt (Vrandečić und Krötzsch, 2014; Waagmeester et al., 2020). Die Einbindung von Ressourcen aus Wikidata in RDF-basierte Metadaten ermöglicht es, Datensätze in einen umfassenden Wissenskontext einzubetten und sie mit externen Entitäten (z. B. Experimenten, Methoden, Personen, Ontologien oder Softwareartefakten) zu verknüpfen. Auf diese Weise werden sowohl die maschinenlesbare Interoperabilität als auch die interdisziplinäre Nachnutzbarkeit wesentlich gestärkt. Handelt es sich bei einem Datensatz beispielsweise um experimentelle Ergebnisse, kann auf generische Wikidata-Ressourcen wie Q101965 verwiesen werden, wodurch die semantische Einordnung der Daten für alle Nutzer unmittelbar nachvollziehbar wird.

Die in Abbildung 5.6 schematisch dargestellte Struktur eines Wissensgraphen verdeutlicht dieses Konzept eines verteilten, FAIR-orientierten Forschungsdatenraums. Die Abbildung zeigt, wie heterogene Ressourcen, darunter Forschungsdaten, Modelle, Personen, Organisationen, Software und verwendete Methoden, als eigenständige, aber über persistente Identifikatoren und semantische Beschreibungen verknüpfte Knoten in einem dezentralen Netzwerk repräsentiert werden. Die Datenbank ist in diesem Sinne keine physische Einheit, sondern eine logisch aggregierte Sicht auf verteilte Ressourcen, die Beziehungen explizit macht, die in der Praxis oft implizit sind.

Im Kontext der vorliegenden Arbeit werden Datensätze daher nicht als isolierte, monolithische Einheiten verstanden, sondern als Teil eines semantisch strukturierten Forschungsdatenraums. Jede Ressource ist eindeutig adressierbar und kann über RDF-basierte Beschreibungen in Beziehung zu anderen Ressourcen gesetzt werden. Die Veröffentlichung in Wikidata stellt dabei keine Duplizierung der eigentlichen Daten dar, sondern eine semantische Referenzierung der Datensätze und ihrer zentralen Konzepte.

Die Integration von Metadaten in Wikidata adressiert alle Aspekte der FAIR-Prinzipien in komplementärer Weise:

- **Findable**: Die Registrierung von Metadaten in Wikidata macht Datensätze global auffindbar. Wikidata fungiert als zentraler Knotenpunkt innerhalb des Linked-Data-Ökosystems

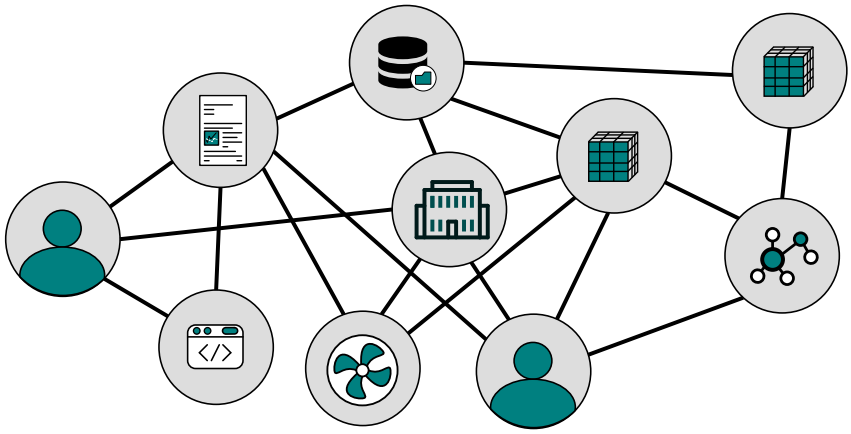


Abbildung 5.6: Konzept eines verteilten Forschungsdatenraums: Sämtliche Ressourcen (Daten, Modelle, Personen, Organisationen, Software und Methoden) sind dezentral verfügbar und über persistente Identifikatoren sowie semantische Beschreibungen miteinander verknüpft.

und ist in viele Forschungstools und semantische Suchsysteme integriert. Dies führt zu einer erheblichen Steigerung der Reichweite und Sichtbarkeit.

- **Accessible**: Zenodo gewährleistet durch DOI-Vergabe die persistente Zugänglichkeit von Datensätzen. Diese wird durch die zusätzliche Verknüpfung mit Wikidata gestärkt, indem die Metadaten in einem offenen, frei zugänglichen Wissensnetzwerk veröffentlicht werden.
- **Interoperable**: Wikidata nutzt standardisierte Identifikatoren und maschinenlesbare Strukturen, die eine nahtlose Integration in wissenschaftliche Infrastrukturen ermöglichen. Dies unterstützt die semantische Vernetzung von Datensätzen über disziplinäre und institutionelle Grenzen hinweg.
- **Reusable**: Die strukturierte und semantisch angereicherte Beschreibung in Wikidata erleichtert das Verständnis und die Nachnutzung der Datensätze. Dadurch wird nicht nur ihre Wiederverwendbarkeit gefördert, sondern auch ihre nachhaltige Einbindung in zukünftige Forschungsprozesse ermöglicht.

Ein weiterer wesentlicher Vorteil liegt in der Nachhaltigkeit: Wikidata ist als offene, durch die wissenschaftliche Gemeinschaft gepflegte Plattform konzipiert, wodurch die Sichtbarkeit und Aktualität der Einträge über die Laufzeit einzelner Projekte hinaus erhalten bleibt. Zudem werden Wikidata-Einträge von gängigen Suchmaschinen indexiert, was die Auffindbarkeit zusätzlich erhöht. Die dort hinterlegten Metadaten erlauben die Verknüpfung mit relevanten Publikationen, Personen und weiteren Entitäten, sodass Datensätze nicht isoliert, sondern innerhalb eines semantischen Netzwerks dargestellt werden. Dies eröffnet neue Möglichkeiten der Kontextualisierung und interdisziplinären Nutzung.

Obwohl Datensätze auf Repositorien wie Zenodo bereits über Suchdienste wie *DataCite Commons* (DataCite, 2025) auffindbar sind, erweitert die Veröffentlichung in Wikidata deren Sicht-

barkeit weit über statische Metadatenschemata hinaus. Während Zenodo-Einträge in ihrer Struktur vergleichsweise statisch bleiben, erlaubt Wikidata die *dynamische*, maschinenlesbare Verknüpfung mit weiteren Wissensressourcen. Diese Ergänzung stellt keine Redundanz dar, sondern eine wertvolle Erweiterung im Sinne der FAIR-Prinzipien, da sie Auffindbarkeit, Interoperabilität und Wiederverwendbarkeit auch langfristig verbessert.

Tabelle 5.8 listet die im Zusammenhang mit der Radialventilatordatenbank angelegten Wikidata-Einträge auf. Ein Großteil dieser Einträge, darunter der untersuchte Ventilator (Q131549102), wurde im Rahmen dieser Arbeit neu erstellt und ermöglicht es nun auch anderen Forschern, auf die konkreten Entitäten der Untersuchung präzise zu referenzieren.

Name	**Wikidata-Eintrag**
openCeFaDB (Datenbank)*	Q137561830
Untersuchter Ventilator*	Q131549102
Standardnamentabelle für Ventilatoreigenschaften*	Q131401744
SSNO*	Q131417267
PIVmeta*	Q131449486
Ventilatorkennlinie*	Q137525063
Ventilatorbetriebspunkt*	Q137525225
Affinitätsgesetze	Q632736

Tabelle 5.8: Ausgewählte, in Wikidata registrierte Ressourcen, die im Rahmen der Radialventilatordatenbank entstanden sind (mit * markiert) und in den Datensätzen referenziert wurden.

Zusammenfassend erweitert die Anreicherung und Integration der Metadaten in Wikidata die reine Veröffentlichung auf Zenodo erheblich: Zenodo stellt die Persistenz und Zitierbarkeit sicher und Wikidata gewährleistet Sichtbarkeit, semantische Vernetzung und maschinenlesbare Integration in das globale Linked-Data-Ökosystem.

5.7 Zugriff und Verwendung

Dieser Abschnitt beschreibt den praktischen Ablauf für Forscher, die eigene Simulationen des generischen Radialventilators anhand der bereitgestellten Referenzmessungen validieren möchten. Der Arbeitsablauf umfasst das Identifizieren relevanter Datensätze, die Durchführung eigener Simulationen sowie den Vergleich mit den bereitgestellten Referenzmessungen.

Die Nutzung der in dieser Arbeit entwickelten Datenbanklösung folgt einem klar strukturierten, mehrstufigen Workflow, der sowohl explorative als auch automatisierte Anwendungsfälle unterstützt. Ziel ist es, den Zugriff auf verteilte Forschungsdaten so zu gestalten, dass Nutzer relevante Datensätze effizient identifizieren, bewerten und gezielt weiterverarbeiten können, ohne große Datenmengen vorab herunterladen zu müssen.

Ausgangspunkt ist eine deklarative Konfigurationsdatei im Turtle-Format („catalog.ttl“), die die logische Struktur der Datenbank beschreibt. In dieser Datei werden *dcat:Dataset*-Instanzen definiert, die jeweils auf eine oder mehrere *dcat:Distribution*-Ressourcen verweisen. Diese Distributionen repräsentieren RDF-Metadaten, die gemeinsam mit den zugehörigen HDF5-Dateien über persistente Repositorien veröffentlicht wurden. Die Konfigurationsdatei definiert somit nicht die Daten selbst, sondern legt fest, welche verteilten Ressourcen zu einer konkreten Datenbankinstanz gehören (vgl. Abbildung 5.7).

Ein zentrales Merkmal des Systems ist die Trennung zwischen Konfiguration, Speicherung und Nutzung. Die Konfigurationsdatei kann unabhängig von der eigentlichen Datenhaltung versioniert, geteilt und angepasst werden. Auf diese Weise lassen sich unterschiedliche Sichten auf denselben Datenbestand oder projektspezifische Teilmengen definieren, ohne Daten zu duplizieren.

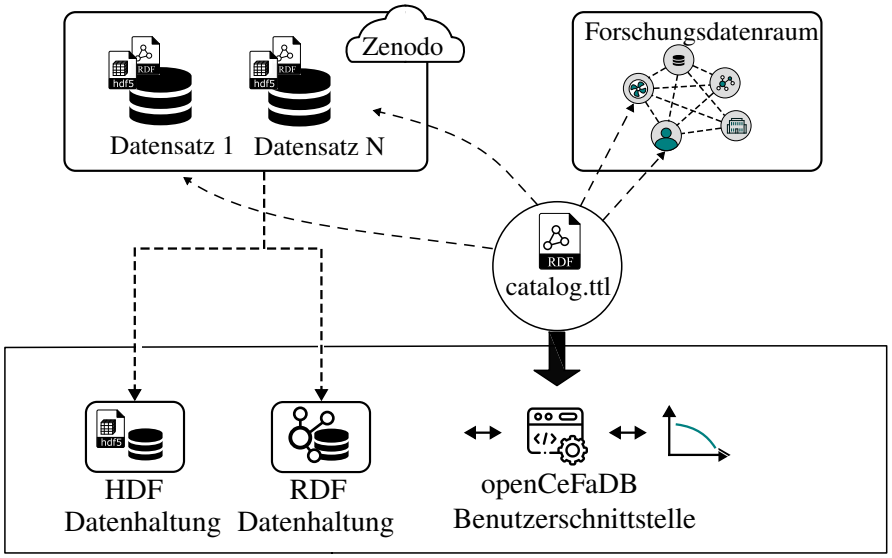


Abbildung 5.7: Die *openCeFaDB* stellt die Schnittstelle zu den verteilten Daten (Zenodo und Wikidata) zur Verfügung. Durch die zentrale Datei *catalog.ttl* werden alle der Datenbank zugehörigen Datensätze und Entitäten definiert und referenziert. Die Benutzerschnittstelle organisiert die Zugriffe auf die Datenhaltungen und unterstützt die Datenanalyse sowie Ergebnisvisualisierung.

Die eigentliche Nutzung erfolgt über das in dieser Arbeit entwickelte Python-Paket *opencefadb* (Probst, 2026b). Dieses übernimmt die automatisierte Verarbeitung der Konfigurationsdatei, das Herunterladen der referenzierten RDF-Distributionen sowie deren Import in einen RDF-Graphen. Dabei ist die konkrete Wahl des Graphdatenbanksystems abstrahiert. Unterstützt werden sowohl lokale *In-Memory*-Graphen (z. B. auf Basis von *rdflib*) als auch persistente oder externe (remote) *Triple Stores*. Für die Nutzer ergibt sich daraus eine einheitliche Schnittstelle, unabhängig von der zugrunde liegenden Infrastruktur.

Nach dem Laden der RDF-Daten steht die Datenbank vollständig für semantische Abfragen zur

Verfügung, wie sie exemplarisch in Abbildung 5.8 in vereinfachter Form angedeutet sind. Die Exploration erfolgt über SPARQL, wobei Abfragen entweder programmgesteuert erzeugt oder aus vordefinierten Vorlagen (Templates) ausgewählt werden können. Diese Vorlagen kapseln domänenspezifisches Wissen und ermöglichen es auch weniger erfahrenen Nutzer, komplexe Abfragen auszuführen. Gleichzeitig bleibt die volle Ausdrucksstärke von SPARQL erhalten, sodass auch individuelle, ad hoc formulierte Abfragen möglich sind.

Die Ergebnisse der SPARQL-Abfragen dienen primär der Exploration und Selektion. Typische Fragestellungen betreffen etwa die Identifikation bestimmter Betriebszustände, Versuchsparameter oder Simulationskonfigurationen. Da die RDF-Metadaten direkt aus den HDF5-Dateien extrahiert wurden, besteht eine eindeutige Rückverfolgbarkeit zwischen semantischen Entitäten und den zugehörigen binären Datensätzen. Auf Basis der Abfrageergebnisse können die relevanten HDF5-Dateien gezielt identifiziert und erst im Anschluss heruntergeladen werden.

Dieser zweistufige Zugriff aus semantischer Exploration vor physischem Datenzugriff reduziert nicht nur den Datenverkehr, sondern erhöht auch die Transparenz und Effizienz der Datennutzung. Insbesondere bei großen, hochaufgelösten Mess- oder Simulationsdaten stellt dies einen entscheidenden Vorteil dar. Häufig sind erste Analysen auf Basis der Metadaten und skalaren Ergebniswerte ausreichend, bevor detailliertere, mehrdimensionale Daten herangezogen werden.

Der grundlegende Ablauf des bis hierhin beschriebenen semantikbasierten Zugriffs, von der Abfrage verteilter Metadaten bis zur Darstellung der Betriebspunktkennlinien, ist exemplarisch in Abbildung 5.8 dargestellt. Die Abbildung visualisiert den Prozess von der SPARQL-Metadatenabfrage über die Verarbeitung des Abfrageergebnisses als RDF-Graph bis hin zur grafischen Darstellung der resultierenden Kennlinien. Dargestellt sind drei Ventilatorkennlinien, die auf realen Messdaten der Datenbank basieren. Die hierfür notwendigen programmgesteuerten Verarbeitungsschritte sind im Rahmen der in Probst (2026b) bereitgestellten Software implementiert.

Abbildung 5.8 stellt die Beziehung zwischen semantisch beschriebenen Metadaten im RDF-Graphen und den daraus abgeleiteten numerischen Daten im Plot bewusst nur schematisch und anhand ausgewählter Aspekte dar. Die Darstellung ist damit nicht vollständig, dient jedoch der Verdeutlichung des methodischen Prinzips und einer verbesserten Lesbarkeit.

Die Nutzung der Datenbank ist nicht auf interaktive Szenarien beschränkt. Durch die vollständige Automatisierbarkeit des Workflows eignet sich die Infrastruktur ebenso für reproduzierbare Auswertungen, Benchmarking-Studien oder kontinuierliche Validierungsprozesse. Die Datenbank fungiert damit nicht nur als Archiv, sondern als aktiver Bestandteil wissenschaftlicher Arbeitsabläufe.

Datenzugang

Die Forschungsdaten der Betriebspunktmessungen sind in mehreren Datensätzen auf Zenodo veröffentlicht. Für jede Messkampagne, die in der Regel aus mehreren Betriebspunkten bei einer festen Drehzahl besteht, wurde ein eigenständiger Datensatz angelegt. Jeder Datensatz enthält für jeden Betriebspunkt ein Tupel aus einer HDF5-Datei mit den Primärdaten (Zeitreihen und

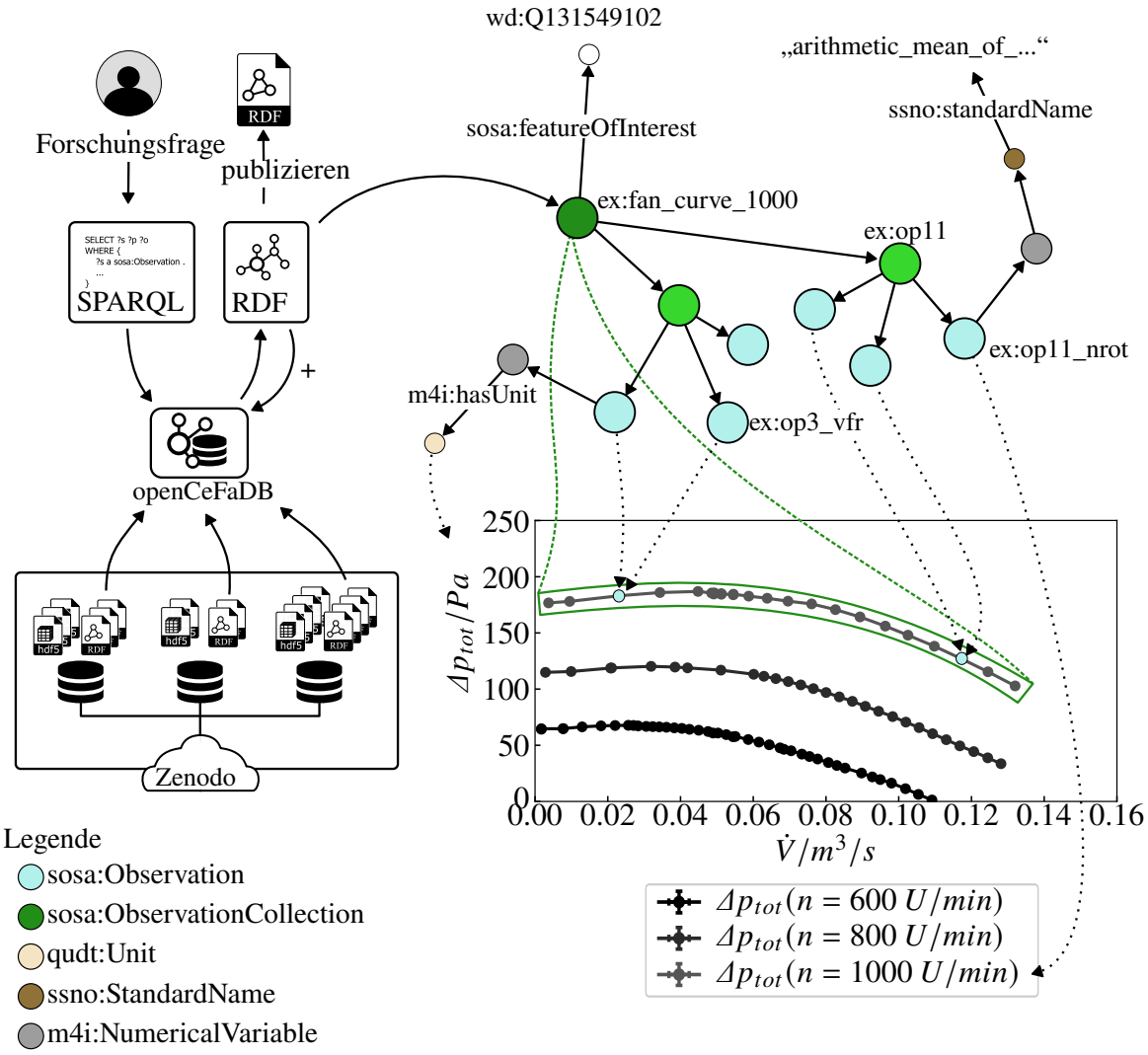


Abbildung 5.8: Visualisierung des semantikbasierten Zugriffs auf verteilte Forschungsdaten zur Darstellung von Ventilatorkennlinien. Ausgehend von einer SPARQL-Abfrage auf die Datenbankschnittstelle *openCeFaDB* werden Betriebspunktdaten aus verteilten Repositorien (Zenodo) über RDF-Metadaten selektiert und zur Darstellung der gemessenen Kennlinien verarbeitet. Die dargestellten Kennlinien für die Drehzahlen 600, 800 und 1000 U/min basieren vollständig auf den im RDF-Graphen beschriebenen Metadaten, einschließlich Standardnamen, Einheiten und Beobachtungsstrukturen. Die Abbildung zeigt bewusst eine vereinfachte Auswahl semantischer Beziehungen.

abgeleitete Größen) sowie einer zugehörigen Turtle-Datei mit den semantischen Metadaten. Durch die DOI-Vergabe sind alle Datensätze dauerhaft referenzierbar und sowohl manuell als auch maschinell eindeutig zitier- und abrufbar.

Die veröffentlichten Daten sind über verschiedene Zugangswege auffindbar:

- über die Suchfunktion des European Open Science Cloud Portals[6],
- direkt über Zenodo,
- über Verlinkungen im semantischen Datenökosystem Wikidata,
- über aggregierende Metadatenportale wie DataCite Commons (DataCite, 2025),
- sowie über die bereitgestellte Softwarelösung zur direkten Interaktion mit den Daten (*opencefadb*).

Nutzer haben zwei grundlegende Möglichkeiten, mit den Daten zu interagieren:

- durch manuelles Herunterladen der Datensätze und individuelle Auswertung oder
- durch die Nutzung der bereitgestellten Python-Software *opencefadb* für einen automatisierten Zugriff.

Die Kuratierung und Veröffentlichung neuer Datensätze erfolgt durch die Daten erhebenden Personen am Lehrstuhl. Hierfür wird das komplementäre, nicht öffentliche Repositorium *opencefadb-admin* verwendet, das die Qualitätssicherung und Konsistenz der veröffentlichten Daten unterstützt (Probst, 2026b).

Der beschriebene Publikations- und Zugangsweg adressiert insbesondere die FAIR-Prinzipien der Auffindbarkeit und Zugänglichkeit (Findable, Accessible). Persistente Identifikatoren (DOI), standardisierte Metadatendienste sowie mehrere unabhängige Zugriffskanäle stellen sicher, dass die Daten langfristig auffindbar und sowohl manuell als auch maschinell zugänglich sind.

Name	**DOI**
openCeFaDB Catalog	10.5281/zenodo.18349358
Dataset 1 - n600	10.5281/zenodo.18349039
Dataset 2 - n600	10.5281/zenodo.18349130
Dataset 1 - n800	10.5281/zenodo.18349134
Dataset 1 - n1000	10.5281/zenodo.18349196
Dataset 1 - n1200	10.5281/zenodo.18349210
Dataset 2 - n1200	10.5281/zenodo.18349214
CAD Modell	10.5281/zenodo.17871736
Standardnamentabelle	10.5281/zenodo.17572275
Zusätzliche Größenarten (Erweiterung der QUDT-Ontologie)	10.5281/zenodo.18297457

Tabelle 5.9: Übersicht der im Rahmen der OpenCeFaDB veröffentlichten Daten- und Metadatenartefakte mit ihren persistenten Identifikatoren.

[6] https://open-science-cloud.ec.europa.eu/

Softwareunterstützung

Neben den beschriebenen Datenformaten wurden zwei Softwarepakete entwickelt, die eine unmittelbare und reproduzierbare Interaktion mit den veröffentlichten Forschungsdaten ermöglichen und die in dieser Arbeit vorgestellte Methodik praktisch umsetzen. Hierzu stehen zwei komplementäre GitHub-Repositorien zur Verfügung:

- *opencefadb-admin* (nicht öffentlich): dient der internen Verwaltung, Pflege und Veröffentlichung der Datensätze und richtet sich ausschließlich an die Datenkuratierung.
- *opencefadb* (öffentlich): richtet sich an Nutzer der Datenbank und stellt Werkzeuge für den automatisierten Zugriff, die Validierung eigener Simulationsdaten sowie die Integration der Daten in bestehende Auswerteworkflows bereit.

Die Python-Bibliothek *opencefadb* bildet die zentrale Schnittstelle für die Nutzung der Datenbank. Sie ist gezielt so konzipiert, dass der Zugriff auf die Daten nicht über projektspezifische Dateipfade oder Kenntnis des Aufbaus einer HDF erfolgt, sondern ausschließlich über öffentliche, semantisch beschriebene Metadaten. Hierzu werden SPARQL-Abfragen auf den RDF-Beschreibungen der Betriebspunkte ausgeführt, anhand derer die relevanten Größen identifiziert, interpretiert und anschließend mit den zugehörigen HDF5-Daten verknüpft werden.

Für typische Anwendungsfälle stellt die Bibliothek vordefinierte SPARQL-Abfragen bereit, die in Form benutzerfreundlicher Funktionen und Klassen gekapselt sind. Zentrale Auswertungen, wie etwa die Extraktion von Betriebspunktdaten oder die Darstellung von Kennlinien, können dadurch ohne vertiefte Kenntnisse semantischer Abfragesprachen durchgeführt werden. Die Einstiegshürde für die Nutzung der Datenbank bleibt damit bewusst niedrig. Gleichzeitig bleibt der vollständige semantische Datenraum über SPARQL direkt zugänglich, sodass Nutzer mit entsprechender Expertise eigene, komplexere Abfragen formulieren können. Für beide Nutzungsszenarien stellt diese Arbeit mit dem Softwarepaket *opencefadb* eine Sammlung von Skripten und Beispielabfragen bereit, die sowohl als unmittelbar nutzbare Werkzeuge als auch als Ausgangspunkt für weiterführende Auswertungen dienen.

Ein wesentliches Merkmal dieses Ansatzes besteht darin, dass Auswerteskripte keine impliziten Annahmen über interne Datenstrukturen treffen müssen. Stattdessen werden die benötigten Informationen explizit über ihre semantische Beschreibung adressiert, insbesondere über Standardnamen, Einheiten, Größenarten sowie definierte Transformationen. Die fachliche Auswertelogik ist damit konsequent von technischen Speicherdetails entkoppelt, was einen zentralen Unterschied zu klassischen, dateibasierten Auswerteansätzen darstellt.

Die praktische Umsetzung dieses semantikbasierten Zugriffs ist exemplarisch in Abbildung 5.8 dargestellt. Die gezeigten Kennlinien für mehrere gemessene Drehzahlen wurden vollständig automatisiert aus der Datenbank erzeugt. Sowohl die Auswahl der Betriebspunkte als auch die Zuordnung der Achsen, Einheiten und Beschriftungen erfolgte ausschließlich auf Basis von SPARQL-Abfragen auf den RDF-Metadaten mithilfe von *opencefadb*. Ein direkter Zugriff auf spezifische Felder oder Namen der Datasets in den HDF5-Dateien ist hierfür nicht erforderlich.

Die semantikbasierte Nutzung der Daten adressiert insbesondere die FAIR-Prinzipien der Interoperabilität und Nachnutzbarkeit (*Interoperable*, *Reusable*). Durch die explizite Beschreibung

der Bedeutung der Daten mittels Standardnamen, Größenarten und Ontologien können die Daten unabhängig von ihrer internen Speicherstruktur interpretiert, kombiniert und in neuen Kontexten wiederverwendet werden. Zusätzlich überprüft *opencefadb* die Konformität der Daten gegenüber den definierten Metadaten- und Strukturvorgaben. Hierzu werden SHACL-Shapes eingesetzt, die sowohl die semantische Vollständigkeit als auch die strukturelle Konsistenz der veröffentlichten Datensätze prüfen und so eine qualitätsgesicherte, reproduzierbare Nutzung der Daten unterstützen. Dies ermöglicht schließlich die automatisierte Verarbeitung der Daten und die Erstellung von Diagrammen, wie in Abbildung 5.8 dargestellt. Die Darstellung der Ventilatorkennlinien erfolgt hierbei vollständig auf Basis der Metadatenanalyse und damit unabhängig von den in den HDF5-Dateien tatsächlich verwendeten Variablennamen. Der entsprechende Code ist dem Repository von *opencefadb* sowie dessen Veröffentlichung auf Zenodo zu entnehmen (Probst, 2026b).

6 Bewertung des Managementansatzes

Ziel dieses Kapitels ist die systematische Bewertung des in dieser Arbeit entwickelten Forschungsdatenmanagementansatzes im Hinblick auf seine Konformität mit den FAIR-Prinzipien und damit der Überprüfung des genuinen Ziels. Die Bewertung dient dabei nicht der formalen Zertifizierung oder der normativen Klassifikation des Ansatzes als „FAIR-konform“, sondern der analytischen Einordnung seiner architektonischen und methodischen Eigenschaften entlang etablierter FAIR-Kriterien. Im Zentrum steht die Frage, in welchem Umfang und mit welcher methodischen Tiefe die FAIR-Prinzipien durch den gewählten Ansatz technisch umgesetzt und langfristig abgesichert werden.

6.1 Evaluationsschema und methodisches Vorgehen

Als Bewertungsrahmen wird das von der Research Data Alliance (RDA) entwickelte FAIR Data Maturity Model herangezogen (Bahim et al., 2020; FAIR Data Maturity Model Working Group, 2020). Dieses Modell beschreibt FAIR nicht als binäre Eigenschaft, sondern als graduell erreichbaren Reifegrad, der anhand definierter Indikatoren und Reifestufen (Level 0 bis Level 4) eingeordnet wird. Die Reifestufen erlauben keine Aussage über absolute Abstände, sondern dienen ausschließlich der vergleichenden Einordnung der Umsetzungstiefe einzelner FAIR-Prinzipien.

Die Wahl dieses Modells begründet sich darin, dass es sowohl qualitative als auch strukturelle Aspekte der FAIR-Umsetzung berücksichtigt und explizit auf maschinenverarbeitbare, standardkonforme und dokumentierte Lösungen abzielt. Damit ist es insbesondere für technisch geprägte Forschungsdatenmanagementansätze in den Ingenieurwissenschaften geeignet, für die bislang nur wenige disziplinspezifische Bewertungsstandards existieren.

Die Bewertung erfolgt in einer kombinierten Vorgehensweise: Zunächst wird für jedes FAIR-Prinzip eine binäre Einschätzung („erfüllt“ bzw. „nicht erfüllt“) vorgenommen, um grundlegende Zielerreichungen festzustellen. Darauf aufbauend erfolgt eine qualitative Einordnung der Umsetzungstiefe entlang der sog. Reifestufen des Modells. Diese Kombination erlaubt sowohl eine klare Orientierung als auch eine differenzierte Bewertung der methodischen Ausgestaltung.

Es sei darauf hingewiesen, dass die Einordnung nicht auf einer automatisierten FAIR-Prüfung basiert, sondern auf einer qualitativen Analyse der konzeptionellen, strukturellen und semantischen Eigenschaften des entwickelten Ansatzes. Dabei werden sowohl die technische Umsetzung als auch bewusst in Kauf genommene Einschränkungen berücksichtigt.

6.2 Bewertung der FAIR-Dimensionen

Auffindbarkeit (Findable)

Die Kriterien zur Auffindbarkeit werden durch den vorgestellten Ansatz in weiten Teilen erfüllt. Die Veröffentlichung der Datensätze mit persistenten Identifikatoren (DOI) adressiert zentrale

Anforderungen an die dauerhafte Referenzierbarkeit sowohl für menschliche als auch für maschinelle Nutzer. Die Metadaten werden strukturiert, formalisiert und maschinenlesbar mittels RDF modelliert und sind eindeutig mit den jeweiligen Datenobjekten innerhalb der HDF5-Struktur verknüpft.

Durch die Verwendung kontrollierter Vokabulare, standardisierter Ontologien sowie eindeutig identifizierbarer Ressourcen wird eine semantisch präzise Beschreibung der Daten ermöglicht. Die explizite Referenzierung einzelner Datenobjekte innerhalb der HDF5-Dateien stellt sicher, dass Metadaten nicht nur auf Dateiebene, sondern auf Ebene fachlich relevanter Objekte zugeordnet sind.

Einschränkungen bestehen derzeit hinsichtlich der externen Auffindbarkeit neu eingeführter, projektspezifischer Ontologien und Standardnamensdefinitionen. Zwar werden diese offen im RDF/Turtle-Format bereitgestellt, die verwendeten Identifikatoren sind jedoch zum Zeitpunkt der Erstellung dieser Arbeit noch nicht über dauerhaft dereferenzierbare Namensräume publiziert oder in etablierten Ontologieverzeichnissen registriert. Diese Einschränkung ist nicht konzeptionell bedingt, sondern resultiert aus dem Entwicklungsstatus der Ontologien im Rahmen der Arbeit. Für eine vollständige Erreichung hoher FAIR-Reifegrade ist eine zukünftige Veröffentlichung über persistente Namensräume vorgesehen.

Insgesamt erreicht der Ansatz hinsichtlich der Auffindbarkeit einen hohen, jedoch noch nicht maximalen Reifegrad.

Zugänglichkeit (Accessible)

Die Zugänglichkeit der Daten und Metadaten wird durch den Einsatz standardisierter Zugriffsmechanismen gewährleistet. Der Zugriff auf Metadaten erfolgt über etablierte Protokolle, insbesondere über SPARQL-Endpunkte sowie über eine programmgesteuerte Python-API. Die Daten selbst werden im Open-Access-Modus veröffentlicht, sodass keine technischen oder rechtlichen Zugangshürden bestehen.

Ein wesentliches Merkmal des Ansatzes ist die Trennung zwischen Primärdaten und Metadaten bei gleichzeitiger formaler Kopplung beider Ebenen. Dadurch bleibt die Zugänglichkeit der Metadaten auch dann erhalten, wenn Primärdaten aus technischen oder rechtlichen Gründen nicht mehr verfügbar sind. Die langfristige Verfügbarkeit der Metadaten wird durch die Archivierung in institutionell getragenen Repositorien sichergestellt.

Zusätzlich erhöht die Hinterlegung ausgewählter Metadaten in externen Wissensbasen die Sichtbarkeit und Redundanz der Informationen. Damit werden zentrale Anforderungen an nachhaltige Zugänglichkeit erfüllt, auch über den ursprünglichen Projektkontext hinaus.

Interoperabilität (Interoperable)

Die Interoperabilität stellt eine der zentralen Stärken des entwickelten Ansatzes dar. Sie ergibt sich nicht allein aus der Nutzung offener Standards wie HDF5, RDF und OWL, sondern aus deren

konsequenter Anwendung bis auf Ebene einzelner Datenobjekte. Die semantische Modellierung erfolgt explizit, formalisiert und unter Nutzung sowohl generischer als auch domänenspezifischer Vokabulare.

Eigene Ontologien werden modular eingeführt und an bestehende Modelle angebunden, wodurch eine Erweiterbarkeit ohne Bruch bestehender Strukturen ermöglicht wird. Die Bedeutung der Daten wird nicht implizit vorausgesetzt, sondern explizit modelliert, wodurch eine maschinelle Interpretation und Weiterverarbeitung über System- und Disziplingrenzen hinweg unterstützt wird.

Durch die qualifizierte Modellierung von Relationen, Einheiten, Größen und Provenienzinformationen wird ein sehr hoher Interoperabilitätsgrad erreicht, der deutlich über konventionelle daten- oder dateizentrierte Ansätze hinausgeht.

Nachnutzbarkeit (Reusable)

Die Nachnutzbarkeit der Daten wird durch eine konsistente Kombination aus strukturierter Datenhaltung, formaler Metadatenmodellierung und Validierungsmechanismen unterstützt. Die explizite Beschreibung von Kontext, Parametern, Annahmen und Provenienzinformationen schafft die Voraussetzungen für eine langfristige, projektunabhängige Wiederverwendung der Daten.

Darüber hinaus wird die Nachnutzbarkeit der begleitenden Software als integraler Bestandteil des wissenschaftlichen Beitrags berücksichtigt. In Anlehnung an die Empfehlungen von Lamprecht et al. (2020) wird die entwickelte Software versioniert, offen zugänglich veröffentlicht, mit persistenten Identifikatoren versehen und mittels *CodeMeta* semantisch beschrieben. Dadurch wird die Software zitierfähig, eindeutig referenzierbar und reproduzierbar nutzbar.

Die Software ist nicht als bloßes Hilfsmittel zu verstehen, sondern als methodisches Artefakt, das die konzeptionellen Entwurfsprinzipien technisch umsetzt und überprüfbar macht.

6.3 Vergleich mit konventioneller Praxis

Zur Einordnung des entwickelten Ansatzes wird dieser einer konventionellen Praxis der Daten- und Metadatenorganisation gegenübergestellt, wie sie in vielen ingenieurwissenschaftlichen Forschungsprojekten anzutreffen ist. Unter konventioneller Praxis wird dabei eine überwiegend datei- und ordnerzentrierte Organisation verstanden, bei der Metadaten in separaten, meist textbasierten Artefakten (z. B. README-Dateien, Tabellen oder Dateinamenkonventionen) abgelegt werden.

In solchen Ansätzen sind semantische Bedeutungen und Zusammenhänge häufig implizit und stark vom projektspezifischen Kontextwissen der beteiligten Personen abhängig. Eine formale Kopplung zwischen Daten und Metadaten sowie systematische Validierungsmechanismen fehlen in der Regel.
Demgegenüber basiert der in dieser Arbeit vorgestellte Ansatz auf einer objektzentrierten Organisation der Daten, einer expliziten semantischen Modellierung und einer formalisierten Kopplung

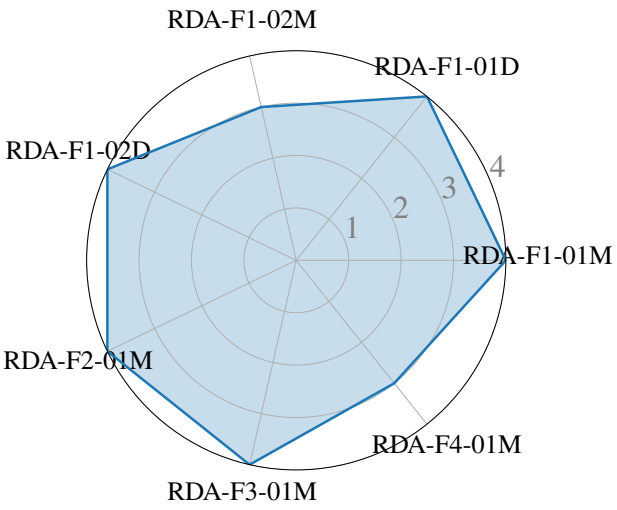
zwischen Daten und Metadaten. Die in Tabelle A.2 dargestellte Gegenüberstellung macht diese Unterschiede systematisch sichtbar und dient als analytische Grundlage für die Einordnung der FAIR-Reifegrade.

6.4 Gesamtbewertung

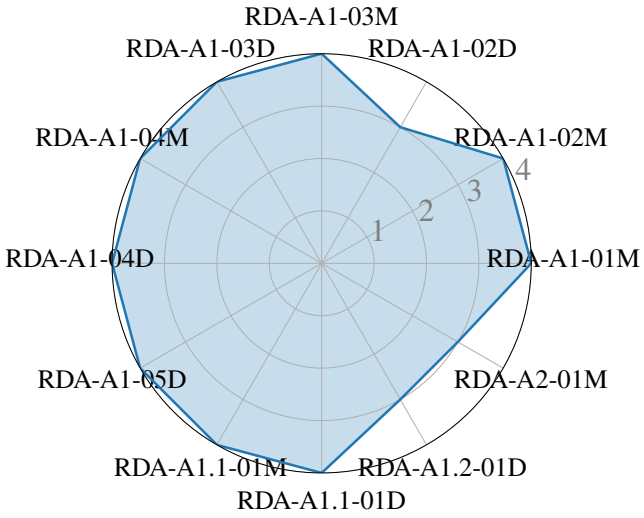
Die Anwendung des FAIR Data Maturity Model zeigt, dass der entwickelte Forschungsdatenmanagementansatz die FAIR-Prinzipien nicht nur weitgehend erfüllt, sondern deren Umsetzung systematisch durch die zugrunde liegende Architektur absichert. Verbleibende Einschränkungen betreffen überwiegend Aspekte der externen Sichtbarkeit und formalen Etablierung projektspezifischer Ontologien und sind ohne grundlegende Änderungen des Konzepts adressierbar.

Der wissenschaftliche Beitrag der Arbeit liegt in der integrativen Verbindung von konzeptionellen Designprinzipien, technischer Umsetzung und softwaregestützter Operationalisierung. Der Ansatz zeigt, dass FAIR nicht als nachgelagerte Eigenschaft verstanden werden muss, sondern als gestaltendes Entwurfsprinzip, das bereits während der Datenentstehung wirksam wird.

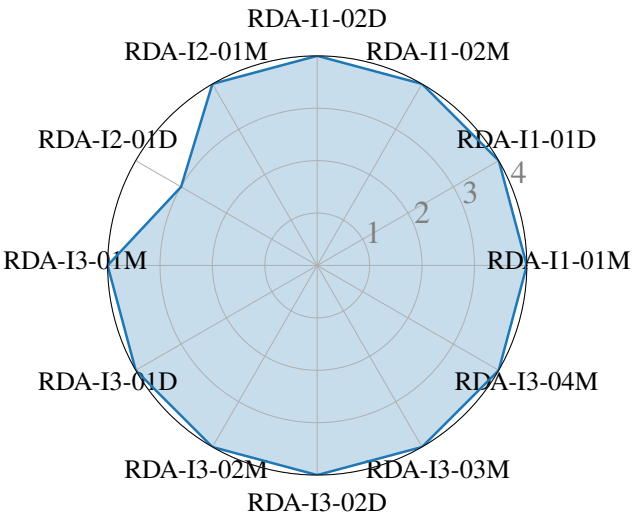
Damit leistet die Arbeit einen methodisch fundierten Beitrag zum nachhaltigen Forschungsdatenmanagement in den Ingenieurwissenschaften, der über bestehende Praktiken hinausgeht und zugleich realistisch in bestehende Forschungsprozesse integrierbar ist.



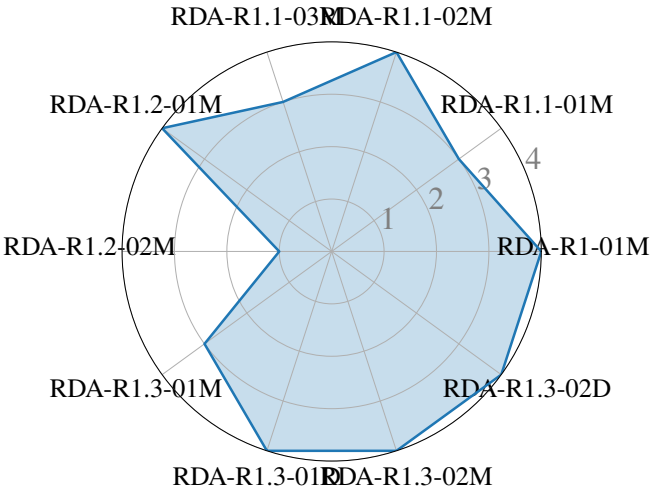
(a) Auffindbarkeit



(b) Zugänglichkeit



(c) Interoperabilität



(d) Nachnutzbarkeit

Abbildung 6.1: Auswertung des FAIR Data Maturity Models, angewandt auf das Datenmanagementkonzept dieser Arbeit. Die Identifikatoren sind Abschnitt A.1 bzw. (FAIR Data Maturity Model Working Group, 2020) zu entnehmen.

7 Zusammenfassung und Ausblick

Die nachhaltige Archivierung und Bereitstellung von Forschungsdaten gewinnt angesichts stetig wachsender Datenmengen, zunehmender interdisziplinärer Zusammenarbeit und neuer datengetriebener Methoden kontinuierlich an Bedeutung (Neuroth, Strathmann et al., 2012). Insbesondere in den Ingenieurwissenschaften stellen heterogene Systemlandschaften, komplexe Datenstrukturen und lange Lebenszyklen von Forschungsdaten hohe Anforderungen an ein belastbares Forschungsdatenmanagement (FDM). Vor diesem Hintergrund fördern wissenschaftspolitische Initiativen und Forschungsförderer zunehmend die Umsetzung der FAIR-Prinzipien (Findable, Accessible, Interoperable, Reusable) als Leitlinien für nachhaltige Dateninfrastrukturen.

Die vorliegende Arbeit zeigt, dass ein konsequent umgesetztes FDM in den Ingenieurwissenschaften eine zentrale Voraussetzung darstellt, um mit großen Datenvolumina, domänenspezifischer Heterogenität und interdisziplinären Nutzungsszenarien umzugehen. Am Beispiel numerischer und experimenteller Untersuchungen an einem Radialventilator wird deutlich, dass fehlende Standardisierung und unzureichend strukturierte Metadaten zu Informationsverlust, eingeschränkter Nachnutzbarkeit und erheblichem Mehraufwand führen können. Diese Problematik ist nicht auf die Strömungsmechanik beschränkt, sondern betrifft zahlreiche natur- und ingenieurwissenschaftliche Disziplinen (Crystal-Ornelas et al., 2022; Michener et al., 2006).

Vor diesem Hintergrund wurde ein Forschungsdatenmanagementkonzept entwickelt, das auf zwei komplementären Säulen beruht: der strukturierten Ablage wissenschaftlicher Primärdaten im HDF5-Format sowie der semantischen Beschreibung von Metadaten mittels RDF und Ontologien. Die Kombination beider Technologien ermöglicht nicht nur eine robuste Speicherung und effiziente Verarbeitung der Daten, sondern schafft zugleich die Grundlage für eine maschinenlesbare, eindeutig interpretierbare und vernetzte Beschreibung wissenschaftlicher Inhalte im Sinne der FAIR-Prinzipien.

Die praktische Umsetzung des Konzepts erfolgte anhand einer Validierungsdatenbank für numerische Strömungssimulationen eines generischen Radialventilators und umfasst insbesondere eine HDF5-basierte Dateikonvention mit klar definiertem Layout, die Einführung standardisierter Namen für physikalische Größen, die Entwicklung einer generischen Ontologie (SSNO) sowie die Bereitstellung einer Python-basierten Softwaretoolbox (*h5rdmtoolbox*, *ontolutils*, *opencefadb*) zur Erstellung, Validierung und Nutzung semantisch angereicherter Daten.

Die Veröffentlichung der Daten über ein disziplinübergreifendes Repositorium gewährleistet Persistenz, Zitierfähigkeit und langfristige Verfügbarkeit. Die zusätzliche Einbindung ausgewählter Metadaten in externe Wissensbasen erhöht die Sichtbarkeit und Anschlussfähigkeit der Daten über institutionelle Grenzen hinweg. Durch föderierte SPARQL-Abfragen können lokale und externe Ressourcen kombiniert und flexibel genutzt werden, womit der Ansatz der Philosophie des Web of Data folgt und die Einschränkungen klassischer, monolithischer Datenbanksysteme überwindet.

Die Bewertung des entwickelten Konzepts anhand des FAIR Data Maturity Models zeigt, dass insbesondere die Interoperabilität und Nachnutzbarkeit auf hohem Reifegrad realisiert werden konnten. Identifizierte Verbesserungspotenziale betreffen vor allem die externe Auffindbarkeit neu entwickelter Ontologiebegriffe, etwa durch die Veröffentlichung über dereferenzierbare

Namensräume, sowie die weitere Formalisierung einzelner Metadatenaspekte. Diese Punkte sind implementierungs- bzw. publikationsbedingt und erfordern keine grundlegenden Anpassungen der zugrunde liegenden Architektur.

Ein zentraler Mehrwert der Arbeit liegt zudem in der Rolle der begleitenden Software als wissenschaftliches Artefakt. Durch Versionierung, persistente Identifikatoren und eine formale Beschreibung mittels CodeMeta wird die Software selbst FAIR-orientiert bereitgestellt und als reproduzierbares, zitierfähiges Forschungsergebnis nutzbar gemacht. Damit fungiert die Software nicht lediglich als Hilfsmittel, sondern als integraler Bestandteil der entwickelten Methodik und als Träger der zugrunde liegenden Designprinzipien.

Ein Ausblick ergibt sich aus aktuellen Entwicklungen im Bereich der Künstlichen Intelligenz. Während Ontologien formale Konsistenz, Transparenz und semantische Eindeutigkeit gewährleisten, bieten Large Language Models neue Möglichkeiten für flexible, interaktive Zugänge zu komplexen Datenbeständen. Kombinationen beider Ansätze, etwa im Rahmen von Retrieval-Augmented Generation oder Prompt-to-Query-Verfahren (DeBellis et al., 2024), eröffnen Perspektiven für eine niederschwellige Nutzung semantisch strukturierter Forschungsdaten. Voraussetzung hierfür bleibt eine saubere formale Modellierung, insbesondere im Hinblick auf Provenienz, Validierbarkeit und Nachvollziehbarkeit.

Langfristig sollten Ontologien als dynamische Instrumente wissenschaftlicher Kommunikation verstanden werden, deren Qualität sich nicht allein an formaler Konsistenz, sondern vor allem an ihrer Akzeptanz, Pflege durch die Community und ihrem praktischen Nutzen bemisst. Die vorliegende Arbeit liefert hierfür einen methodischen und technischen Rahmen, der auf andere ingenieurwissenschaftliche Anwendungsfälle übertragbar ist.

Zusammenfassend adressiert diese Arbeit die formulierte Forschungslücke zwischen normativen FAIR-Forderungen und deren praktischer Umsetzung für komplexe ingenieurwissenschaftliche Forschungsdaten. Durch die Kombination aus strukturierter Primärdatenhaltung, semantischer Metadatenmodellierung und softwaregestützter Validierung wird ein methodischer Rahmen bereitgestellt, der FAIR-Prinzipien nicht nur fordert, sondern technisch überprüfbar und reproduzierbar umsetzt. Damit wird gezeigt, wie Daten und Software gleichermaßen als nachhaltige, nachnutzbare wissenschaftliche Artefakte etabliert werden können, die über einzelne Anwendungsfälle hinaus anschlussfähig sind.

Literatur

Aggour, K. S. et al. „Semantics-Enabled Data Federation: Bringing Materials Scientists Closer to FAIR Data“. In: *Integrating Materials and Manufacturing Innovation*, Band 13, Heft 2 (2024), S. 420–434.

Allianz der Wissenschaftsorganisationen. „Grundsätze zum Umgang mit Forschungsdaten“. In: *SSRN Electronic Journal* (2010). DOI: 10.2312/ALLIANZOA.019.

Allotrope Foundation. *The HDF5 data description ontology. (REC/2024/12)*. Allotrope Foundation. 2024. URL: https://purl.allotrope.org/voc/adf/REC/2024/12/hdf.ttl (abgerufen am 04. 03. 2025).

Almeida, A. V. d., Borges, M. M. und Roque, L. „The European open science cloud: a new challenge for Europe“. In: *Proceedings of the 5th international conference on technological ecosystems for enhancing multiculturality*. 2017, S. 1–4.

Anzt, H. et al. „An environment for sustainable research software in Germany and beyond: current state, open challenges, and call for action“. In: *F1000Research*, Band 9 (2021), S. 295.

Aranda, B. et al. *SPARQL 1.1 Overview*. 2013. URL: https://www.w3.org/TR/sparql11-overview/ (abgerufen am 28. 06. 2024).

Arndt, S. et al. *Metadata4Ing: An ontology for describing the generation of research data within a scientific activity.* 2023. DOI: 10.5281/zenodo.8382665.

Ashburner, M. et al. „Gene ontology: tool for the unification of biology. The Gene Ontology Consortium“. In: *Nat Genet*, Band 25, Heft 1 (2000), S. 25–29. DOI: 10.1038/75556.

Auer, S. et al. „DBpedia: A Nucleus for a Web of Open Data“. In: *The Semantic Web*. Springer, Berlin, Heidelberg. Springer Berlin Heidelberg, 2007, S. 722–735. DOI: 10.1007/978-3-540-76298-0_52.

Bahim, C. et al. „The FAIR Data Maturity Model: An Approach to Harmonise FAIR Assessments.“ In: *Data Science Journal*, Band 19 (2020). DOI: 10.5334/dsj-2020-041.

Beckett, D. und Berners-Lee, T. *RDF 1.1 Turtle: Terse RDF Triple Language*. W3C Recommendation. W3C Recommendation, 25.02.2014. World Wide Web Consortium (W3C), 2014. URL: https://www.w3.org/TR/turtle/.

Berman, F. und Crosas, M. „The research data alliance: Benefits and challenges of building a community organization“. In: *Harvard Data Science Review*, Band 2, Heft 1 (2020), S. 1–11.

Berners-Lee, T., Hendler, J. und Lassila, O. „The Semantic Web“. In: *Scientific American*, Band 284, Heft 5 (2001), S. 34–43. URL: http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21.

Brandt, N. et al. „Kadi4Mat: A research data infrastructure for materials science“. In: *Data Science Journal*, Band 20 (2021), S. 8–8.

Budroni, P., Claude-Burgelman, J. und Schouppe, M. „Architectures of Knowledge: The European Open Science Cloud“. In: *ABI Technik*, Band 39, Heft 2 (2019), S. 130–141. DOI: 10.1515/abitech-2019-2006.

Carolus, T. *Ventilatoren :* 3., überarb. u. erw. Aufl. 2013. SpringerLink. Wiesbaden: Vieweg+Teubner Verlag, 2013. DOI: 10.1007/978-3-8348-2472-1.

Crystal-Ornelas, R. et al. „Enabling FAIR data in Earth and environmental science with community-centric (meta) data reporting formats“. In: *Scientific data*, Band 9, Heft 1 (2022), S. 700. DOI: 10.1038/s41597-022-01606-w.

DataCite. *DataCite Commons*. URL: https://commons.datacite.org/ (abgerufen am 19. 12. 2025).

De Carlo, F. et al. „Scientific data exchange: a schema for HDF5-based storage of raw and analyzed data“. In: *Journal of synchrotron radiation*, Band 21, Heft 6 (2014), S. 1224–1230.

DeBellis, M. et al. „Integrating Ontologies and Large Language Models to Implement Retrieval Augmented Generation (RAG)“. In: *Applied Ontology*, Band 1 (2024), S. 1–5.

Degtyarenko, K. et al. „ChEBI: a database and ontology for chemical entities of biological interest“. In: *Nucleic Acids Research*, Band 36, Heft suppl_1 (2007), S. D344–D350. ISSN: 0305-1048. DOI: 10.1093/nar/gkm791.

Deutsche Forschungsgemeinschaft. *DFG: Leitlinien zum Umgang mit Forschungsdaten.* 2015. URL: http://www.dfg.de/download/pdf/foerderung/antragstellung/forschungsdaten/richtlinien_forschungsdaten.pdf (abgerufen am 09. 04. 2025).

Deutsche Forschungsgemeinschaft. *Empfehlungen zur gesicherten Aufbewahrung und Bereitstellung digitaler Forschungsprimärdaten*. 2009. URL: https://www.dfg.de/resource/blob/169298/51f011ec1a047637243ea95a994a49e6/ua-inf-empfehlungen-200901-data.pdf (abgerufen am 12. 05. 2025).

Deutsche Forschungsgemeinschaft. *Guidelines for Safeguarding Good Research Practice. Code of Conduct*. 2025. DOI: 10.5281/zenodo.14281892.

Dublin Core Metadata Initiative. *Dublin Core Metadata Initiative*. 2007. URL: http://dublincore.org (abgerufen am 17. 10. 2023).

Dürst, M. und Suignard, M. *Internationalized Resource Identifiers (IRIs)*. Techn. Ber. Network Working Group, 2005.

Ehrlinger, L. und Wöß, W. „Towards a definition of knowledge graphs.“ In: *SEMANTiCS (Posters, Demos, SuCCESS)*, Band 48, Heft 1-4 (2016), S. 2.

Europäische Kommission. *Richtlinie 2009/125/EG*. 2009.

European Commission. *Cost of not having FAIR research data. Cost-Benefit analysis for FAIR research data*. Techn. Ber. European Commission, 2018. DOI: 10.2777/02999.

European Organization For Nuclear Research und OpenAIRE. *Zenodo*. en. 2013. DOI: 10.25495/7GXK-RD71. URL: https://www.zenodo.org/.

FAIR Data Maturity Model Working Group. „FAIR Data Maturity Model: specification and guidelines“. In: *Research Data Alliance* (2020). DOI: 10.15497/RDA00050.

figshare. 2023. DOI: 10.17616/R3PK5R. (Abgerufen am 01. 03. 2025).

Floridi, L. „Is semantic information meaningful data?“ In: *Philosophy and phenomenological research*, Band 70, Heft 2 (2005), S. 351–370.

Garijo, D. „WIDOCO: a wizard for documenting ontologies“. In: *International Semantic Web Conference*. Springer, Cham. 2017, S. 94–102. DOI: 10.1007/978-3-319-68204-4_9.

Gemeinsame Wissenschaftskonferenz. *Bund-Länder-Vereinbarung zu Aufbau und Förderung einer Nationalen Forschungsdateninfrastruktur (NFDI)*. 2018. URL: https://www.gwk-bonn.de/fileadmin/Redaktion/Dokumente/Papers/NFDI.pdf (abgerufen am 29. 09. 2025).

Ghiringhelli, L. M. et al. „Towards efficient data exchange and sharing for big-data driven materials science: metadata and data formats“. In: *npj computational materials*, Band 3, Heft 1 (2017), S. 46. DOI: 10.1038/s41524-017-0048-5.

Gosink, L. et al. „HDF5-FastQuery: Accelerating complex queries on HDF datasets using fast bitmap indices“. In: *18th International Conference on Scientific and Statistical Database Management (SSDBM'06)*. IEEE. 2006, S. 149–158.

Gray, J. et al. „Scientific data management in the coming decade“. In: *Acm Sigmod Record*, Band 34, Heft 4 (2005), S. 34–41.

Gülich, J. F. *Centrifugal Pumps*. 4. Aufl. Springer Cham, 2020. ISBN: 978-3-030-14788-4. DOI: 10.1007/978-3-030-14788-4.

Haller, A. et al. *Semantic Sensor Network Ontology*. W3C Recommendation. W3C Recommendation, 19.10.2017. World Wide Web Consortium (W3C), 2017. URL: https://www.w3.org/TR/vocab-ssn/.

Harris, C. R. et al. „Array programming with NumPy“. In: *Nature*, Band 585, Heft 7825 (2020), S. 357–362. DOI: 10.1038/s41586-020-2649-2.

Harris, M. *CF Conventions Home Page*. 2025. URL: https://cfconventions.org/ (abgerufen am 10. 04. 2025).

Harris, M. *Software that "Understands" CF Data*. 2025. URL: https://cfconventions.org/software.html (abgerufen am 16. 04. 2025).

Hartl, N., Wössner, E. und Sure-Vetter, Y. „Nationale Forschungsdateninfrastruktur (NFDI)“. In: *Informatik Spektrum*, Band 44, Heft 5 (2021), S. 370–373.

Haslhofer, B. und Isaac, A. „data.europeana.eu: The Europeana Linked Open Data Pilot“. In: *International Conference on Dublin Core and Metadata Applications*, Band 2011 (2011). DOI: 10.23106/dcmi.952135673.

Hassell, D. et al. „A data model of the Climate and Forecast metadata conventions (CF-1.6) with a software implementation (cf-python v2. 1)“. In: *Geoscientific Model Development*, Band 10, Heft 12 (2017), S. 4619–4646.

HDFql - The easy way to manage HDF5 data. HDFql Project. URL: https://www.hdfql.com/ (abgerufen am 25. 06. 2024).

Heber, G. und Folk, M. „An OWL description of HDF5“. In: *Proceedings of the IADIS International Conference WWW/Internet 2013*. 2013, S. 251–258. ISBN: 978-989-8533-16-6.

Heinrichs, B. et al. „Automatic General Metadata Extraction and Mapping in an HDF5 Usecase.“ In: *KDIR*. 2021, S. 172–179.

Hitzler, P. et al. *Semantic Web: Grundlagen*. Bd. 1. Springer, 2008.

Horsch, M. T., Chiacchiera, S., Bami, Y. et al. *Reliable and interoperable computational molecular engineering: 2. Semantic interoperability based on the European Materials and Modelling Ontology*. 2020. DOI: 10.48550/arXiv.2001.04175.

Horsch, M. T., Chiacchiera, S., Cavalcanti, W. L. et al. *Data Technology in Materials Modelling*. Springer Cham, 2021. DOI: 10.1007/978-3-030-68597-3.

Horsch, M. T., Morgado, J. F. et al. „Domain-specific metadata standardization in materials modelling“. In: *Proceedings of DORIC-MM* (2021), S. 12–27.

Hoyer, S. und Hamman, J. „xarray: N-D labeled arrays and datasets in Python“. In: *Journal of Open Research Software*, Band 5, Heft 1 (2017). DOI: 10.5334/jors.148.

Iglezakis, D. et al. „Modelling Scientific Processes With the M4I Ontology“. In: *Proceedings of the Conference on Research Data Infrastructure*. Bd. 1. 2023. DOI: 10.52825/cordi.v1i.271.

Jones, M. B. et al. *CodeMeta: an exchange schema for software metadata. Version 3.0*. 2023. URL: https://w3id.org/codemeta/3.0 (abgerufen am 07. 07. 2024).

Junkes, H. et al. „FAIRmat-a consortium of the German research-data infrastructure (NFDI)“. In: *18th International Conference on Accelerator and Large Experimental Physics Control Systems (ICALEPCS'21), Shanghai, China, 14-22 October 2021*. JACoW Publishing, Geneva, Switzerland. 2022, S. 558–563.

Kailus, A. *Handreichung für ein FAIRes Management kulturwissenschaftlicher Forschungsdaten*. 2023. DOI: 10.5281/zenodo.7716941.

Kettinger, W. J. und Li, Y. „The infological equation extended: towards conceptual clarity in the relationship between data, information and knowledge“. In: *European Journal of Information Systems*, Band 19 (2010), S. 409–421.

Kinkade, D. und Shepherd, A. „Geoscience data publication: Practices and perspectives on enabling the FAIR guiding principles“. In: *Geoscience Data Journal*, Band 9, Heft 1 (2022), S. 177–186.

Klosowski, P. et al. „NeXus: A common format for the exchange of neutron and synchroton data“. In: *Physica B: Condensed Matter*, Band 241 (1997), S. 151–153.

Klump, J. et al. „Data publication in the open access initiative“. In: *Data Science Journal*, Band 5 (2006), S. 79–83.

Knublauch, H. und Kontokostas, D. *Shapes Constraint Language (SHACL)*. W3C Recommendation. W3C Recommendation, 20.07.2017. World Wide Web Consortium (W3C), 2017. URL: https://www.w3.org/TR/shacl/.

Kompenhans, J. „Test and comparison of various methods of analysis and post-processing on a Database of PIV records“. In: *Particle Image Velocimetry: Progress towards Industrial Application*. Springer, 2000, S. 37–89.

Kraft, A., Engel, F. und Klinger, A. „Terminologies in RDM for Engineering–a Service Approach: NFDI4Ing Terminology Service“. In: *Proceedings of the Conference on Research Data Infrastructure*. Bd. 1. 2023.

Kröger, J. und Wedlich-Zachodin, K. „Das Beteiligungsmodell von forschungsdaten.info: Ein kleines ABC der Nachhaltigkeit“. In: *Bausteine Forschungsdatenmanagement*, Heft 1 (2020), S. 86–95. DOI: 10.17192/bfdm.2020.1.8160.

Lamprecht, A.-L. et al. „Towards FAIR principles for research software“. In: *Data Science*, Band 3, Heft 1 (2020), S. 37–59. DOI: 10.3233/DS-190026.

Lanquillon, C. und Schacht, S. *Knowledge Science–Grundlagen*. Springer, 2023.

Legensky, S. et al. „CFD general notation system (CGNS)-status and future directions“. In: *40th AIAA Aerospace Sciences Meeting & Exhibit*. 2002, S. 752.

Liao, X. et al. „FAIR Data Cube, a FAIR data infrastructure for integrated multi-omics data analysis“. In: *Journal of Biomedical Semantics*, Band 15, Heft 1 (2024), S. 20.

Macilenti, G., Stellato, A. und Fiorelli, M. „SIS: Leveraging Semantically-Informed Similarity of Text Embeddings for Enhanced Ontology Alignment“. In: *Procedia Computer Science*, Band 270 (2025), S. 505–514.

Malyshev, S. et al. „Getting the Most Out of Wikidata: Semantic Technology Usage in Wikipedia's Knowledge Graph“. In: *The Semantic Web – ISWC 2018*. Hrsg. von D. Vrandečić et al. Cham: Springer International Publishing, 2018, S. 376–394. ISBN: 978-3-030-00668-6.

Manola, F. und Miller, E. *Resource Description Framework (RDF). Primer. W3C Recommendation 10 February 2004*. 2004. URL: http://www.w3.org/TR/rdf-primer/ (abgerufen am 17. 06. 2024).

Masters, J., Hodgson, R. und Keller, P. J. *FAIRsharing.org: QUDT; Quantities, Units, Dimensions and Types*. DOI: 10.25504/FAIRsharing.d3pqw7.

Michener, W. K. et al. „Meta-information concepts for ecological data management“. In: *Ecological informatics*, Band 1, Heft 1 (2006), S. 3–7.

Millecam, T. et al. „Coming of age of Allotrope: Proceedings from the Fall 2020 Allotrope Connect“. In: *Drug Discovery Today*, Band 26, Heft 8 (2021), S. 1922–1928. ISSN: 1359-6446. DOI: 10.1016/j.drudis.2021.03.028.

Musen, M. A. „The protégé project: a look back and a look forward“. In: *AI Matters*, Band 1, Heft 4 (2015), S. 4–12. DOI: 10.1145/2757001.2757003.

Neuroth, H., Putnings, M. und Neumann, J. *Praxishandbuch Forschungsdatenmanagement*. De Gruyter, 2021.

Neuroth, H., Strathmann, S. et al., Hrsg. *Langzeitarchivierung von Forschungsdaten: Eine Bestandsaufnahme*. Boizenburg: Verlag Werner Hülsbusch in Kooperation mit Universitätsverlag Göttingen, 2012, S. 378. ISBN: 978-3-86488-008-7.

Nihar, A. et al. „Toward findable, accessible, interoperable and reusable (fair) photovoltaic system time series data“. In: *2021 IEEE 48th Photovoltaic Specialists Conference (PVSC)*. IEEE. 2021, S. 1701–1706.

Otte, J. N., Beverley, J. und Ruttenberg, A. „BFO: Basic formal ontology“. In: *Applied ontology*, Band 17, Heft 1 (2022), S. 17–43.

Pareti, P. und Konstantinidis, G. „A Review of SHACL: From Data Validation to Schema Reasoning for RDF Graphs“. In: *Reasoning Web. Declarative Artificial Intelligence : 17th International Summer School 2021, Leuven, Belgium, September 8–15, 2021, Tutorial Lectures*. Hrsg. von M. Šimkus und I. Varzinczak. Cham: Springer International Publishing, 2022, S. 115–144. DOI: 10.1007/978-3-030-95481-9_6.

Paskin, N. „Digital object identifiers for scientific data“. In: *Data science journal*, Band 4 (2005), S. 12–20.

Payne, K. und Verhey, C. „Schema.org for research data managers: a primer“. In: *International Journal of Big Data Management*, Band 2, Heft 2 (2022), S. 95–116.

Pelz, P., Saul, S. und Brötz, J. „Efficiency scaling: influence of Reynolds and Mach numbers on fan performance“. In: *Journal of Turbomachinery*, Band 144, Heft 6 (2022), S. 061001.

Pezoa, F. et al. „Foundations of JSON schema“. In: *Proceedings of the 25th international conference on World Wide Web*. 2016, S. 263–273.

Pimenta, I. S. et al. „A FAIR Future for Engineering Sciences: Linking an RDM Community Through a Scientific Journal“. In: *Proceedings of the Conference on Research Data Infrastructure*. Bd. 1. 2023.

Preston-Werner, T. *Semantic Versioning 2.0.0*. URL: https://semver.org/lang/de/ (abgerufen am 10. 04. 2025).

Preuß, N. und Pelz, P. „Integrated management of experimental research-and meta-data for fan test rigs“. In: *International Conference on Fan Noise, Aerodynamics*. 2018.

Preuss, N. et al. „Methods and technologies for Research-and Metadata Management in Collaborative Experimental Research“. In: *Applied Mechanics and Materials*. Bd. 885. Trans Tech Publ. 2018, S. 170–183.

Probst, M. *Ontolutils - Object-oriented „Things“*. Version 0.27.6. 2026. DOI: 10.5281/zenodo.18450246.

Probst, M. *opencefadb*. Version 1.0.0. 2026. DOI: 10.5281/zenodo.18412779.

Probst, M. *ssnolib*. Version 2.2.0.3. 2025. DOI: 10.5281/zenodo.18070322.

Probst, M. und Pritz, B. *Generic Centrifugal Fan CAD File*. Version 1.0.1. Zenodo, 2025. DOI: 10.5281/zenodo.17871736.

Probst, M. und Pritz, B. „h5RDMtoolbox - A Python Toolbox for FAIR Data Management around HDF5“. In: *ing.grid*, Band 2 (1 2024). DOI: 10.48694/inggrid.4028.

Probst, M. und Pritz, B. *h5RDMtoolbox - Supporting a FAIR Research Data lifecycle using Python and HDF5*. Version 2.7.3. 2026. DOI: 10.5281/zenodo.18494608.

Probst, M. und Pritz, B. *PIVMeta: An ontology draft for describing Particle Image Velocimetry data*. Version 3.1.0. 2025. DOI: 10.5281/zenodo.17560453.

Probst, M. und Pritz, B. *SSNO: A simple Standard Name Ontology*. Version 2.2.0. 2025. DOI: 10.5281/zenodo.17604194.

Probst, M. und Pritz, B. *Standard Name Table for the openCeFaDB*. Version 2.0.1. 2025. DOI: 10.5281/zenodo.17572275.

Probst, M. und Pritz, B. *Standard Name Table for the Property Descriptions of Centrifugal Fans*. Version 1.2.0. 2025. DOI: 10.5281/zenodo.17271932.

Queralt-Rosinach, N. et al. „Applying the FAIR principles to data in a hospital: challenges and opportunities in a pandemic“. In: *Journal of biomedical semantics*, Band 13, Heft 1 (2022), S. 12.

Robertson-von Trotha, C. Y. und Schneider, R. H., Hrsg. *Digitales Kulturerbe : Bewahrung und Zugänglichkeit in der wissenschaftlichen Praxis*. Bd. 2. Kulturelle Überlieferung - Digital. KIT Scientific Publishing, 2015. 220 S. ISBN: 978-3-7315-0317-0. DOI: 10.5445/KSP/1000044869.

Rocca-Serra, P. et al. *FAIR Cookbook*. 2022. DOI: 10.5281/zenodo.6783564.

Rumsey, C. et al. „Recent updates to the CFD general notation system (CGNS)“. In: *50th AIAA Aerospace Sciences Meeting including the New Horizons Forum and Aerospace Exposition*. 2012, S. 1264.

Safi, S., Thiessen, T., Schmailzl, K. J. et al. „Acceptance and resistance of new digital technologies in medicine: qualitative study“. In: *JMIR research protocols*, Band 7, Heft 12 (2018), e11072.

Schembera, B. und Iglezakis, D. „EngMeta: metadata for computational engineering“. In: *International Journal of Metadata, Semantics and Ontologies*, Band 14, Heft 1 (2020), S. 26–38.

Schembera, B., Selent, B. et al. *Datenmanagement in Infrastrukturen, Prozessen und Lebenszyklen für die Ingenieurwissenschaften : Abschlussbericht des BMBF-Projektes Dipl-Ing*. Stuttgart, 2019. DOI: 10.2314/KXP:1693393980.

Schneider, M. et al. „Datenmanagement im SFB 1313“. In: *Bausteine Forschungsdatenmanagement*, Heft 1 (2020), S. 28–38.

Selent, B. et al. „Management of Research Data in Computational Fluid Dynamics and Thermodynamics“. In: *Tage 2019* (2019), S. 128.

Shannon, C. E. „A mathematical theory of communication“. In: *The Bell system technical journal*, Band 27, Heft 3 (1948), S. 379–423.

Sima, A. C. et al. „Semantic integration and enrichment of heterogeneous biological databases“. In: *Evolutionary genomics: statistical and computational methods* (2019), S. 655–690.

Singh, J. „FigShare“. In: *Journal of Pharmacology and Pharmacotherapeutics*, Band 2, Heft 2 (2011), S. 138–138.

Srivastava, D. J. et al. „Core Scientific Dataset Model: A lightweight and portable model and file format for multi-dimensional scientific data“. In: *Plos one*, Band 15, Heft 1 (2020), e0225953.

Unidata. *NetCDF: Interoperability with HDF5*. Official NetCDF documentation. 2024. URL: https://docs.unidata.ucar.edu/netcdf-c/4.9.2/interoperability_hdf5.html.

Voß, J. „Was sind eigentlich Daten?“ In: *LIBREAS. Library Ideas*, Heft 23 (2013), S. 4–11.

Vrandečić, D. und Krötzsch, M. „Wikidata: a free collaborative knowledgebase“. In: *Communications of the ACM*, Band 57, Heft 10 (2014), S. 78–85. DOI: 10.1145/2629489.

W3C OWL Working Group. *OWL 2 Web Ontology Language Document Overview (Second Edition)*. World Wide Web Consortium (W3C). 2012. URL: https://www.w3.org/TR/owl2-overview/ (abgerufen am 11. 04. 2025).

Waagmeester, A. et al. „Science Forum: Wikidata as a knowledge graph for the life sciences“. In: *eLife*, Band 9 (2020). Hrsg. von P. Rodgers und C. Mungall, e52614. ISSN: 2050-084X. DOI: 10.7554/eLife.52614.

Weibel, S. L. und Koch, T. „The Dublin core metadata initiative“. In: *D-lib magazine*, Band 6, Heft 12 (2000).

Wikidata: Statistiken. Wikidata Project. 2025. URL: https://www.wikidata.org/wiki/Wikidata:Statistics/de (abgerufen am 29. 09. 2025).

Wilkinson, M. D. et al. „The FAIR Guiding Principles for scientific data management and stewardship“. In: *Scientific data*, Band 3, Heft 1 (2016), S. 1–9.

Willert, C. „Proposal for NetCDF (re) implementation for use with Planar Velocimetry Data“. In: *Particle Image Velocimetry: Recent Improvements: Proceedings of the EUROPIV 2 Workshop held in Zaragoza, Spain, March 31–April 1, 2003*. Springer. 2004, S. 251–260.

Wissel, S. et al. *Die GO FAIR Initiative: Ein offenes und integratives Ökosystem für FAIR Data Pioniere*. Konferenzabstract. 2020.

Mitbetreute studentische Arbeiten

Büttner, L. „Einführung optimierten Datenmanagements zur experimentellen Untersuchung eines Radialventilators“. Masterarbeit. Karlsruher Institut für Technologie (KIT), 2023.

Dreisbach, M. „Particle detection by means of neural networks in defocusing particle tracking velocimetry on synthetic and experimental data“. Masterarbeit. Karlsruher Institut für Technologie (KIT), 2020.

Duran, C. „Experimental Investigation into optimal Particle Image Velocimetry Parameters for a Centrifugal Fan“. Masterarbeit. Karlsruher Institut für Technologie (KIT), 2021.

Eke, E. „Optimierungspotentiale an einem Radialventilator-Prüfstand“. Bachelorarbeit. Karlsruher Institut für Technologie (KIT), 2020.

Gawlitza, L. „Untersuchung des Optimierungspotentials turbulenter Strömungen in einer abrupten Querschnittserweiterung“. Bachelorarbeit. Karlsruher Institut für Technologie (KIT), 2019.

Neyhousser, M. „Untersuchung des Optimierungspotentials laminarer Strömungen in einer abrupten Querschnittserweiterung“. Bachelorarbeit. Karlsruher Institut für Technologie (KIT), 2019.

Nial, M. „Quantitative Comparison of PIV and CFD Data of a Centrifugal Fan“. Masterarbeit. Karlsruher Institut für Technologie (KIT), 2019.

Peters, H. „Simulation eines generischen Ventilatormodells mit sieben und neun Schaufeln mittels OpenFOAM“. Bachelorarbeit. Karlsruher Institut für Technologie (KIT), 2022.

Raus, D. „Experimentelle Untersuchung der Defocusing Particle Tracking Velocimetry-Methode“. Bachelorarbeit. Karlsruher Institut für Technologie (KIT), 2020.

Reuß, J.-H. von. „Optimierung der Durchführung und Auswertung von Strömungssimulationen am Beispiel eines Radialventilators“. Masterarbeit. Karlsruher Institut für Technologie (KIT), 2023.

Roth, H. „Entwicklung einer Traversensteuerung für einen PIV-Aufbau“. Bachelorarbeit. Karlsruher Institut für Technologie (KIT), 2019.

Saal, J. „Numerische Untersuchung der Einlassrandbedingung eines generischen Radialventilatormodells“. Bachelorarbeit. Karlsruher Institut für Technologie (KIT), 2020.

Szemskat, J. L. „Untersuchung der Netzqualität am Stator-Rotor-Interface für einen radialventilator“. Bachelorarbeit. Karlsruher Institut für Technologie (KIT), 2023.

Anhang

A.1 FAIR Data Maturity Model der Research Data Alliance

In dieser Arbeit wird das FAIR Data Maturity Model der Research Data Alliance als Referenzrahmen für die Einordnung der FAIR-Umsetzung verwendet. Das Modell beschreibt die FAIR-Prinzipien anhand definierter Indikatoren, die den einzelnen FAIR-Dimensionen (Findable, Accessible, Interoperable, Reusable) und deren Unterkriterien zugeordnet sind.

Tabelle A.1 enthält die in dieser Arbeit herangezogenen Indikatoren mit zugehörigem FAIR-Prinzip, Indikator-ID und Priorität gemäß den Originalquellen.

FAIR	RDA-ID	Indikator	Priorität
F1	RDA-F1-01M	Metadaten werden durch einen persistenten Identifikator identifiziert	Wesentlich
F1	RDA-F1-01D	Daten werden durch einen persistenten Identifikator identifiziert	Wesentlich
F1	RDA-F1-02M	Metadaten werden durch einen global eindeutigen Identifier identifiziert	Wesentlich
F1	RDA-F1-02D	Daten werden durch einen global eindeutigen Identifier identifiziert	Wesentlich
F2	RDA-F2-01M	Umfangreiche Metadaten werden bereitgestellt, um ein Auffinden zu ermöglichen	Wesentlich
F3	RDA-F3-01M	Metadaten beinhalten den Identifikator der Daten	Wesentlich
F4	RDA-F4-01M	Metadaten werden so angeboten, dass sie abgefragt und indexiert werden können	Wesentlich
A1	RDA-A1-01M	Metadaten enthalten Informationen, die es dem Nutzer ermöglichen, auf die Daten zuzugreifen	Wichtig
A1	RDA-A1-02M	Metadaten können manuell (d. h. mit menschlicher Beteiligung) abgerufen werden	Wesentlich
A1	RDA-A1-02D	Daten können manuell (d. h. mit menschlicher Beteiligung) abgerufen werden	Wesentlich
A1	RDA-A1-03M	Der Metadatenidentifikator führt zu einem Metadatensatz	Wesentlich
A1	RDA-A1-03D	Der Datenidentifikator führt zu einem digitalen Objekt	Wesentlich
A1	RDA-A1-04M	Der Zugriff auf Metadaten erfolgt über ein standardisiertes Protokoll	Wesentlich

FAIR	RDA-ID	Indikator	Priorität
A1	RDA-A1-04D	Der Zugriff auf Daten erfolgt über ein standardisiertes Protokoll	Wesentlich
A1	RDA-A1-05D	Der Zugriff auf Daten kann automatisch (d. h. durch ein Computerprogramm) erfolgen	Wichtig
A1.1	RDA-A1.1-01M	Der Zugriff auf Metadaten kann über ein freies Zugriffsprotokoll erfolgen	Wesentlich
A1.1	RDA-A1.1-01D	Der Zugriff auf Daten kann über ein freies Zugriffsprotokoll erfolgen	Wichtig
A1.2	RDA-A1.2-01D	Der Zugriff auf Daten kann über ein Zugriffsprotokoll erfolgen, das Authentifizierung und Autorisierung unterstützt	Nützlich
A2	RDA-A2-01M	Metadaten bleiben garantiert verfügbar, wenn die Daten nicht mehr verfügbar sind	Wesentlich
I1	RDA-I1-01M	Metadaten verwenden eine in einem standardisierten Format ausgedrückte Wissensdarstellung	Wichtig
I1	RDA-I1-01D	Daten verwenden eine in einem standardisierten Format ausgedrückte Wissensdarstellung	Wichtig
I1	RDA-I1-02M	Metadaten verwenden eine maschinenlesbare Wissensdarstellung	Wichtig
I1	RDA-I1-02D	Daten verwenden eine maschinenlesbare Wissensdarstellung	Wichtig
I2	RDA-I2-01M	Metadaten verwenden FAIR-konforme Vokabularien	Wichtig
I2	RDA-I2-01D	Daten verwenden FAIR-konforme Vokabularien	Nützlich
I3	RDA-I3-01M	Metadaten enthalten Verweise auf andere Metadaten	Wichtig
I3	RDA-I3-01D	Daten enthalten Verweise auf andere Daten	Nützlich
I3	RDA-I3-02M	Metadaten enthalten Verweise auf andere Daten	Nützlich
I3	RDA-I3-02D	Daten enthalten qualifizierte Verweise auf andere Daten	Nützlich
I3	RDA-I3-03M	Metadaten enthalten qualifizierte Verweise auf andere Metadaten	Wichtig
I3	RDA-I3-04M	Metadaten enthalten qualifizierte Verweise auf andere Daten	Nützlich

FAIR	RDA-ID	Indikator	Priorität
R1	RDA-R1-01M	Eine Mehrzahl genauer und relevanter Attribute wird bereitgestellt, um eine Nachnutzung zu ermöglichen	Wesentlich
R1.1	RDA-R1.1-01M	Metadaten enthalten Informationen über die Lizenz, unter der die Daten nachgenutzt werden können	Wesentlich
R1.1	RDA-R1.1-02M	Metadaten beziehen sich auf eine standardisierte Nachnutzungslizenz	Wichtig
R1.1	RDA-R1.1-03M	Metadaten beziehen sich auf eine maschinenverständliche Nachnutzungslizenz	Wichtig
R1.2	RDA-R1.2-01M	Metadaten enthalten Herkunftsinformationen gemäß den fachspezifischen Standards	Wichtig
R1.2	RDA-R1.2-02M	Metadaten enthalten Herkunftsinformationen gemäß einer fachübergreifenden Sprache	Nützlich
R1.3	RDA-R1.3-01M	Metadaten entsprechen einem Communitystandard	Wesentlich
R1.3	RDA-R1.3-01D	Daten entsprechen einem Communitystandard	Wesentlich
R1.3	RDA-R1.3-02M	Metadaten werden in Übereinstimmung mit einem maschinenlesbaren Communitystandard ausgedrückt	Wesentlich
R1.3	RDA-R1.3-02D	Daten werden in Übereinstimmung mit einem maschinenlesbaren Communitystandard ausgedrückt	Wichtig

Tabelle A.1: Indikatoren des FAIR Data Maturity Model (FAIR Data Maturity Model Working Group, 2020), (Bahim et al., 2020). Der Tabelleninhalt ist den Quellen entnommen.

A.1.1 Gegenüberstellung konventioneller Praxis und vorgestelltem Ansatz mittels HDF und RDF

Die Zuordnung der FAIR-Indikatoren zu typischen Merkmalen konventioneller Praxis sowie zu den im vorgestellten Ansatz umgesetzten organisatorischen und technischen Maßnahmen sind in Tabelle Tabelle A.2 dargestellt. Die Gegenüberstellung dient als kontextualisierende Referenz für die im Bewertungskapitel vorgenommene Einordnung und erhebt keinen Anspruch auf Vollständigkeit oder formale Vergleichbarkeit, da eine binäre Bewertung im Sinne „erfüllt“/„erfüllt nicht“ nicht möglich ist.

Tabelle A.2: Heuristische Gegenüberstellung der Umsetzung der FAIR-Kriterien in einer konventionellen ingenieurwissenschaftlichen Datenpraxis und im in dieser Arbeit entwickelten Ansatz. Die Darstellung dient der qualitativen Operationalisierung des Evaluationsschemas auf Basis des FAIR Prinzipien und erhebt keinen Anspruch auf Vollständigkeit.

FAIR	**Kriterium)**	**Konventionelle ingenieurwissenschaftliche Praxis**	**Vorgestellter Ansatz (HDF5 + RDF + SHACL)**
F1	(Meta-)Daten sind durch einen global eindeutigen und persistenten Identifikator referenzierbar.	Referenzierung häufig über lokale Dateinamen, Pfade oder projektinterne Kennzeichnungen; persistente Identifikatoren meist nur auf Publikationsebene, nicht für interne Datenobjekte.	Persistente Identifikatoren (DOI) für publizierte Datensätze; konsistente interne URIs zur eindeutigen Referenzierung von Entitäten und Datenobjekten innerhalb der Datenstruktur.
F2	Daten sind mit reichhaltigen Metadaten beschrieben.	Metadaten überwiegend in textbasierten Artefakten (README, Tabellen, Dateinamenkonventionen); Umfang, Struktur und Granularität projektabhängig und oft unvollständig.	Formalisierte, maschinenlesbare Metadaten in RDF zur Beschreibung von Kontext, Parametern, Einheiten, Provenienz und Struktur; projektweit definierte Konventionen.
F3	(Meta-)Daten enthalten explizite Referenzen aufeinander.	Bezüge zwischen Daten und Metadaten meist implizit (Ordnerstruktur, Benennung); formale, maschinenlesbare Verknüpfungen selten.	Explizite, bidirektionale Verknüpfung zwischen HDF5-Datenobjekten und RDF-Metadaten über eindeutige Referenzen; nachvollziehbare Zuordnung auf Objektebene.
F4	(Meta-)Daten sind in einer durchsuchbaren Ressource registriert oder indexiert.	Auffindbarkeit meist auf projektinterne Ablagen oder begleitende Publikationen beschränkt; systematische Indexierung selten.	Registrierung und Indexierung der Metadaten über Repositorien mit Suchfunktionalität; zusätzliche Abfrage- und Filtermöglichkeiten über RDF-basierte Indizes (z. B. SPARQL), sofern bereitgestellt.

Fortsetzung auf der nächsten Seite

FAIR	Kriterium	Konventionelle Praxis	Vorgestellter Ansatz
A1	(Meta-)Daten sind über ein standardisiertes Kommunikationsprotokoll abrufbar.	Zugriff primär über Dateisysteme, Netzlaufwerke oder manuelle Weitergabe (E-Mail, Archivdateien); keine standardisierte Zugriffsschnittstelle.	Abruf der Daten und Metadaten über HTTP(S) via Repository Website; Metadaten zusätzlich maschinenlesbar zugänglich.
A1.1	Das verwendete Protokoll ist offen, frei und universell implementierbar.	Abhängig von lokaler Infrastruktur; keine explizite Trennung zwischen Zugriffsmechanismus und Organisationsform.	Verwendung offener, weit verbreiteter Protokolle (HTTP(S)) sowie offener Standards zur Beschreibung der Daten und Metadaten (RDF).
A1.2	Das Protokoll unterstützt Authentifizierung und Autorisierung, falls erforderlich.	Zugriffsregelungen heterogen (Dateirechte, VPN); häufig implizit und unzureichend dokumentiert.	Repository-basierte Zugriffskontrollen (z. B. Embargo, Rollenmodelle) mit dokumentierten Zugriffsbedingungen.
A2	Metadaten bleiben zugänglich, auch wenn die Daten nicht mehr verfügbar sind.	Metadaten häufig gemeinsam mit den Daten verloren oder nur in projektinternen Dokumenten vorhanden.	Persistente Metadaten und Landing Pages im Repositorium; Auflösung der DOI auch bei eingeschränkter oder aufgehobener Datenverfügbarkeit.
I1	(Meta-)Daten verwenden eine formale, zugängliche, geteilte Sprache zur Wissensrepräsentation.	Überwiegend natürliche Sprache oder projektspezifische Tabellenformate; keine formale Semantik.	RDF als formale Repräsentationssprache; Nutzung standardisierter Modellierungsprinzipien für maschinelle Interpretation.
I2	(Meta-)Daten verwenden FAIR-konforme Vokabulare.	Begriffe häufig ad hoc definiert; inkonsistente Terminologie zwischen Projekten.	Verwendung etablierter, versionierter Vokabulare und Ontologien (z. B. für Einheiten, Provenienz); projektspezifische Erweiterungen modular angebunden.

Fortsetzung auf der nächsten Seite

FAIR	Kriterium	Konventionelle Praxis	Vorgestellter Ansatz
I3	(Meta-)Daten enthalten qualifizierte Referenzen zu anderen (Meta-)Daten.	Referenzen meist informell (z. B. Literaturangaben im Text); keine formalen Relationen.	Qualifizierte, maschinenlesbare Referenzen zu Publikationen, Ontologien, Datensätzen und Software über URIs.
R1	(Meta-)Daten enthalten ausreichend Informationen zur Wiederverwendung.	Erforderliches Kontextwissen häufig implizit; Parameterstände, Randbedingungen und Versionen nicht systematisch dokumentiert.	Explizite Beschreibung von Randbedingungen, Parametern, Datenstruktur und Verarbeitungsschritten durch RDF; Granularität abhängig von verfügbarer Information und Bereitstellung.
R1.1	(Meta-)Daten sind mit einer klaren und zugänglichen Nutzungslizenz versehen.	Lizenzangaben fehlen häufig oder sind uneindeutig.	Explizite Lizenzierung der Daten und Metadaten im Repositorium; optionale Wiederholung in den RDF-Metadaten.
R1.2	(Meta-)Daten sind mit detaillierter Provenienz verknüpft.	Provenienzinformationen meist informell (Labornotizen, Skriptnamen); selten maschinenlesbar.	Formale Modellierung von Provenienz (z. B. Erfassung, Verarbeitung, Analyse) mit Verknüpfung zu Software, Versionen und Parametern, soweit praktikabel.
R1.3	(Meta-)Daten erfüllen domänenrelevante Community-Standards.	Explizite Standardbindung uneinheitlich; Abweichungen selten dokumentiert.	Domänenspezifische Konventionen, formalisiert und überprüfbar modelliert; Validierung der Metadatenstruktur mittels SHACL sowie dokumentierte Abweichungen von Standards.

A.2 Softwarepublikationen

Die nachfolgenden Abschnitte dokumentieren die im Rahmen dieser Arbeit entwickelten und eingesetzten Softwareartefakte. Sie dienen der technischen Konkretisierung, Referenzierbarkeit und Reproduzierbarkeit der in dieser Arbeit beschriebenen Konzepte und Methoden. Um den Lesefluss der konzeptionellen Argumentation nicht zu beeinträchtigen, sind Implementierungsdetails, konkrete Werkzeuge sowie Referenzen auf veröffentlichte Code- und Ontologieartefakte bewusst in den Anhang ausgelagert und hier zusammengefasst. Darüber hinaus sei an dieser Stelle ebenfalls auf die technisch detaillierteren und anwenderorientierten Beispiele in den jeweiligen Onlinedokumentationen verwiesen.

Die FAIR-Prinzipien sind bewusst als allgemeine Leitlinien formuliert und lassen einen erheblichen Interpretations- und Gestaltungsspielraum zu (vgl. Unterabschnitt 2.1.2). Vor diesem Hintergrund ist die Vielfalt existierender Softwarelösungen im Bereich des Forschungsdatenmanagements nicht überraschend. Um die in dieser Arbeit entwickelte konzeptionelle Grundlage praktisch nutzbar zu machen, wurden gezielt komplementäre Softwareartefakte entwickelt, die eine konsistente Umsetzung des vorgeschlagenen Ansatzes in realen Forschungsworkflows unterstützen.
Zu diesem Zweck wurden mehrere Python-Bibliotheken implementiert, die zentrale Bausteine des entwickelten Forschungsdatenmanagementkonzepts adressieren. Dazu zählen insbesondere die strukturierte Organisation und Annotation von HDF5-basierten Primärdaten, die formale Modellierung semantischer Metadaten mithilfe von Ontologien sowie die Nutzung standardisierter Benennungsmechanismen. Die Programmierschnittstellen zur Arbeit mit Ontologien werden im folgenden Abschnitt behandelt, während die Integration in den Gesamtzusammenhang des Datenmanagementkonzepts in Unterabschnitt A.2.1 detailliert dargestellt ist.

Tabelle A.3 gibt eine Übersicht über die im Rahmen dieser Arbeit entwickelten Softwarepakete. Diese implementieren ausgewählte Teilaspekte des Forschungsdatenmanagementkonzepts, darunter die HDF5-basierte Datenhaltung (*h5rdmtoolbox*), generische Ontologie-Schnittstellen (*ontolutils*) sowie die Standardnamenontologie *SSNO*, deren unterstützende Softwarebibliothek *ssnolib* und letztlich die Schnittstelle zu den Daten des Praxisbeispiels der generischen Ventilatordtenbank *opencefadb*.

Alle Softwareartefakte wurden entlang etablierter FAIR-orientierter Praktiken publiziert. Die jeweiligen Quelltexte sind über öffentlich zugängliche GitHub-Repositorien verfügbar und enthalten standardisierte *codemeta.json*-Dateien zur maschinenlesbaren Beschreibung zentraler Softwaremetadaten. Zur Sicherstellung der langfristigen Zitierfähigkeit und Versionierbarkeit wurden die relevanten Softwareversionen zusätzlich über Zenodo archiviert und mit persistenten DOIs versehen.
Ergänzend wurden die Softwarepakete als eigenständige Entitäten in Wikidata angelegt und dort mit den zugehörigen Zenodo-Datensätzen sowie weiteren Referenzen verknüpft. Auf diese Weise sind die Softwareartefakte nicht nur zitierfähig, sondern auch semantisch in das Linked-Data-Ökosystem eingebunden, was ihre Auffindbarkeit, Referenzierbarkeit und Nachnutzbarkeit unterstützt.
Die gewählte Publikations- und Metadatenstrategie orientiert sich dabei an den in der Literatur

Name	Wikidata ID	DOI
h5rdmtoolbox	Q126946499	10.5281/zenodo.18494608
opencefadb	Q137885626	10.5281/zenodo.18368245
ontolutils	Q131448345	10.5281/zenodo.18450246
ssnolib	Q131448607	10.5281/zenodo.18070322
SSNO	Q131417267	10.5281/zenodo.10909129

Tabelle A.3: Eine Übersicht aller im Rahmen dieser Arbeit entwickelten und eingesetzten Softwarebibliotheken und softwarenahen Infrastrukturkomponenten mit den zugehörigen Wikidata-IDs und Zenodo-DOIs ist in Tabelle A.3 dargestellt.

vorgeschlagenen FAIR-Prinzipien für Forschungssoftware. Lamprecht et al. (2020) übertragen die FAIR-Grundideen explizit auf Softwareartefakte und benennen unter anderem maschinenlesbare Metadaten (z. B. CodeMeta (Jones et al., 2023)), persistente Identifikatoren sowie die Verknüpfung von Code-Repositorien mit archivierten Softwareversionen als zentrale Bausteine FAIR-orientierter Softwarebereitstellung. Die in dieser Arbeit umgesetzte Kombination aus öffentlich zugänglichen Repositorien, *codemeta.json*-Beschreibungen, DOI-basierter Archivierung über Zenodo und semantischer Verknüpfung in Wikidata folgt diesen Empfehlungen und unterstützt eine nachhaltige Auffindbarkeit, Referenzierbarkeit und Nachnutzung der entwickelten Softwareartefakte.

Die Python-Bibliothek *ssnolib* (Probst, 2025) implementiert eine High-Level-Schnittstelle zwischen der Ontologie *SSNO* (Probst und Pritz, 2025c) und den Anwendern, sodass das Anlegen von semantischen Metadaten erleichtert wird. Sie basiert auf *ontolutils* und nutzt dessen Mechanismen zur Modellierung formeller Ontologien in Python, um ontologische Konzepte als Pydantic-Klassen verfügbar zu machen. Auf diese wird nachfolgend detaillierter eingegangen.

Die Entwicklung der Werkzeuge folgt den FAIR-Prinzipien mit einem besonderen Fokus auf Wiederverwendbarkeit und Auffindbarkeit. Sämtliche Softwarepakete basieren auf etablierten Python-Bibliotheken und werden über die öffentliche Versionsverwaltungsplattform GitHub bereitgestellt. Zur Sicherstellung der langfristigen Zitierfähigkeit und Referenzierbarkeit sind die jeweiligen Versionen über Zenodo archiviert und mit persistenten DOIs versehen. Ergänzend wurden die Softwareartefakte als eigenständige Entitäten in Wikidata angelegt und mit den entsprechenden Zenodo-Datensätzen verknüpft, um eine semantische Einbindung in das Linked-Data-Ökosystem zu ermöglichen.

A.2.1 HDF5-Management Toolbox - *h5rdmtoolbox*

Die *h5rdmtoolbox* (im Folgenden auch *Toolbox*) ist das zentrale Softwareartefakt dieser Arbeit und stellt die technische Realisierung des entwickelten Forschungsdatenmanagementkonzepts auf Basis von HDF5 dar. Durch die Abstraktion komplexer, technisch anspruchsvoller Details ermöglicht sie einen niedrigschwelligen Zugang zu zentralen Funktionen des Daten- und Metada-

tenmanagements entlang des gesamten Forschungszyklus und unterstützt damit eine konsistente Anwendung des Konzepts in wissenschaftlichen Arbeitsprozessen.

Abstrakte Schnittstelle durch die Integration von *xarray* und HDF5

Ein zentraler Entwurfsgedanke der Toolbox ist die enge Kopplung von Daten und Metadaten in einer Form, die unmittelbar in typische Analyse- und Visualisierungsprozessen integrierbar ist. In Python werden HDF5-Datensätze als *NumPy*-Arrays (C. R. Harris et al., 2020) verarbeitet, wodurch zwar numerische Werte effizient verfügbar sind, jedoch Kontextinformationen (z. B. Dimensionen, Koordinaten, Attributmetadaten) leicht verloren gehen, nur implizit im Kontext des Arbeitsschrittes vorhanden oder separat gepflegt werden müssen.

Die Toolbox nutzt *xarray* (Hoyer und Hamman, 2017) als Übertragungsglied: Statt reiner *NumPy*-Arrays werden *xarray.DataArray*-Objekte zurückgegeben, in denen Dimensionen, Koordinaten und Attribute explizit repräsentiert sind. Dadurch können typische Operationen wie Selektion entlang von Koordinaten, Aggregation (z. B. gleitende Mittelwerte) oder Visualisierung direkt auf der kombinierten Daten–Metadaten-Struktur ausgeführt werden. Listing A.1 illustriert diesen Workflow exemplarisch; Abbildung A.1 verdeutlicht das zugrunde liegende Prinzip schematisch.

Listing A.1: Beispiel für die Datenextraktion mithilfe der Toolbox analog zur Darstellung in Abbildung A.1. Der zurückgegebene Wert ist ein *xarray.DataArray* mit umfassenden Metadaten aus dem zugrunde liegenden HDF-Datensatz. Dies erleichtert transparente Datenoperationen und reduziert Fehlerquellen. Darüber hinaus können viele Operationen auf eine einzige Codezeile reduziert werden, was Skripte übersichtlich und nachvollziehbar macht.

```python
import h5rdmtoolbox as h5tbx

with h5tbx.File(filename) as h5:
    # Lese Daten an spezifischen Koordinaten aus:
    arr = h5["velocity"].sel(x=4.3, y=0.2, method="nearest")

# Berechne den gleitenden, zeitlichen Mittelwert:
drm = arr.rolling(time=3).mean()

# Visualisiere das Ergebnis:
drm.plot()
```

Semantische Metadaten als Linked Data: Verknüpfung von HDF5-Objekten und RDF

Das HDF5-Format ermöglicht die syntaktische Metadatenbeschreibung über Attribute an Gruppen und Datensätzen. Die Benennung und Interpretation dieser Attribute bleibt jedoch frei, sodass ohne zusätzliche Maßnahmen weder eine disziplinübergreifend eindeutige Semantik noch

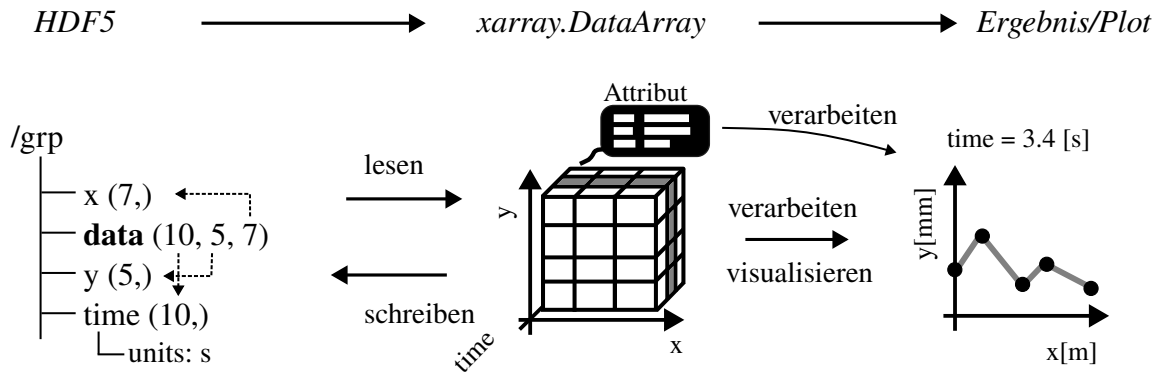


Abbildung A.1: Schematische Darstellung der Nutzung von *xarray* als Bindeglied zwischen HDF5 und darauf aufbauenden Verarbeitungsschritten. Anstelle einfacher *NumPy*-Arrays werden *xarray*-Objekte zurückgegeben, in denen Dimensionsreferenzen und Metadaten (Attribute) erhalten bleiben.

eine maschinelle Integration in Wissensgraphen gewährleistet ist. Zur Umsetzung der FAIR-Prinzipien, insbesondere im Hinblick auf Interoperabilität (I1/I2) und Wiederverwendbarkeit (R1), erweitert *h5rdmtoolbox* dieses Metadatenmodell um eine optionale semantische Schicht auf Basis von RDF.

Konzeptionell werden zwei Ebenen unterschieden:

- **Strukturelle Beschreibung:** Abbildung der internen HDF5-Struktur (Gruppen, Datensätze, Attribute und technische Eigenschaften) als RDF-Graph. Hierfür wird eine HDF-Ontologie genutzt, die zentrale HDF5-Konzepte formal beschreibt.

- **Kontextuelle Beschreibung:** Anwenderspezifische RDF-Tripel zur Beschreibung fachlicher Bedeutung (z. B. physikalische Größe, Einheit, Provenienz, beteiligte Personen) unter Nutzung domänenspezifischer Ontologien und Vokabulare.

Beide Ebenen sind konzeptionell orthogonal zueinander und können unabhängig voneinander genutzt, in gängigen RDF-Serialisierungen exportiert werden (z. B. Turtle, JSON-LD) und ausgewertet werden. Die Serialisierung erfolgt über einen expliziten Parameter, z. B. *structural=False* für ausschließlich kontextuelle Metadaten bzw. *structural=True* für die zusätzliche strukturelle Repräsentation.

Listing A.2 demonstriert die Annotation semantischer Metadaten während der Arbeit mit HDF5. Im Beispiel werden (i) ein Versionsattribut mit *schema:version* verknüpft, (ii) eine Kontaktgruppe als Instanz von *foaf:Person* modelliert und (iii) ein Datensatz mit einem fachlichen Typ sowie einer standardisierten Einheit aus QUDT versehen. Das resultierende Turtle ist in Listing A.3 angegeben.

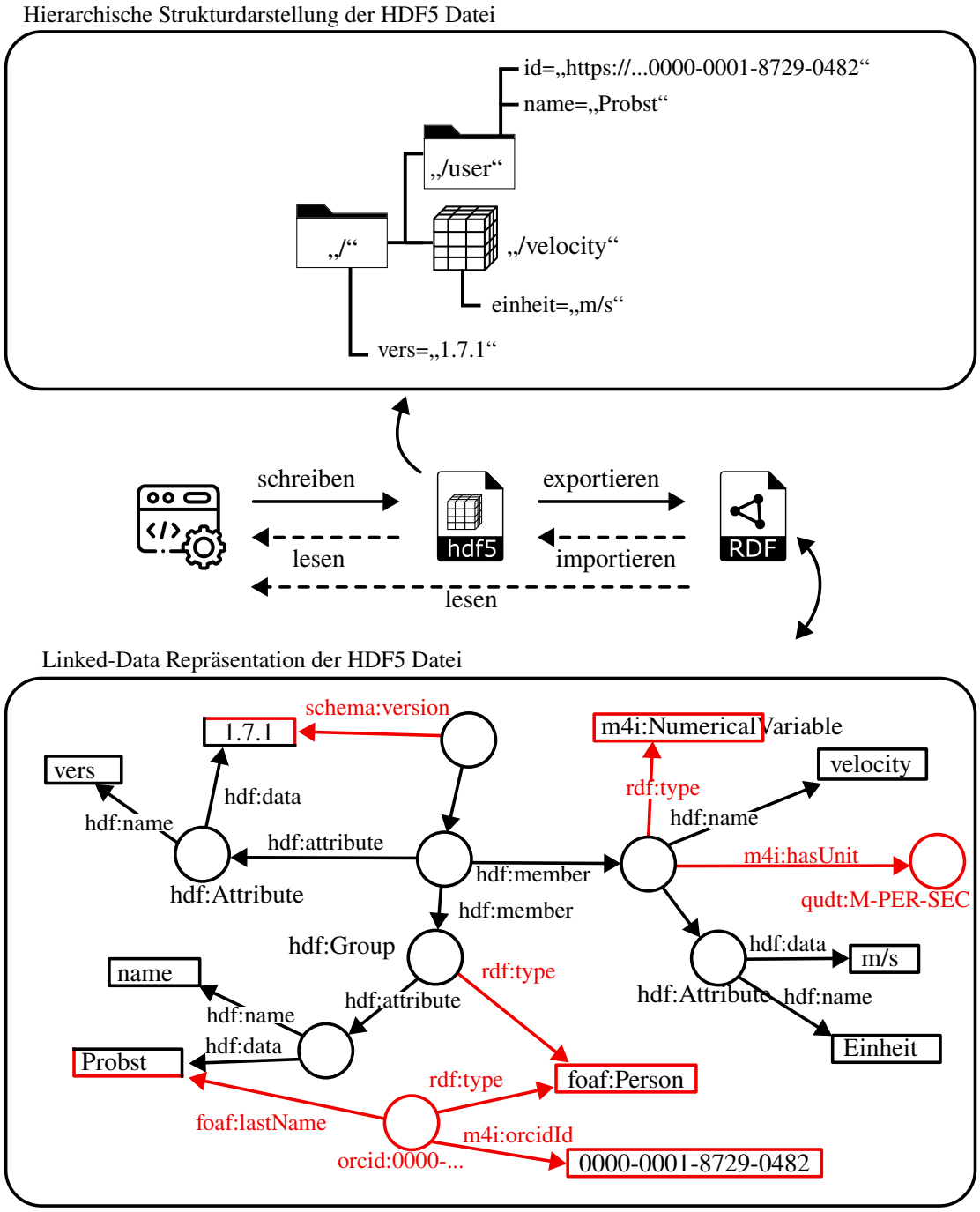


Abbildung A.2: Prinzipielle Ergänzung einer HDF5-Datei um semantische Informationen mittels *h5rdmtoolbox*. Die kontextuellen, benutzerdefinierten Metadaten sind hervorgehoben; die strukturellen Relationen sind aus Übersichtsgründen reduziert dargestellt.

Formalisierung der IRI-Struktur innerhalb semantisch angereicherter HDF5-Dateien

Damit HDF5-Objekte (Datei, Gruppen, Datensätze, Attribute sowie ausgewählte technische Eigenschaften) in RDF eindeutig referenzierbar sind, implementiert die Toolbox ein determinis-

Listing A.2: Beispielcode zur Annotation semantischer Metadaten für HDF5-Dateien. Das resultierende RDF als Ausgabe von *print(ttl)* ist in Listing A.3 dargestellt.

```python
import h5rdmtoolbox as h5tbx

from rdflib import FOAF
from ontolutils import M4I, QUDT_UNIT, SCHEMA

with h5tbx.File() as h5:
    h5.attrs["vers"] = "2.4.0"
    h5.frdf["vers"].predicate = SCHEMA.version

    g = h5.create_group(
        name="contact",
        rdf_type=FOAF.Person,
        rdf_subject="https://orcid.org/0000-0001-8729-0482"
    )
    g.attrs["name", FOAF.lastName] = "Probst"
    g.attrs["id", M4I.orcidId] = "0000-0001-8729-0482"

    ds = h5.create_dataset(
        name="velocity",
        data=5.6,
        rdf_type=M4I.NumericalVariable
    )
    ds.attrs["einheit", M4I.hasUnit] = "m/s"
    ds.rdf["einheit"].object = QUDT_UNIT.M_PER_SEC

    ttl = h5.serialize(structural=False, fmt="ttl")
print(ttl)
```

tisches IRI-Schema. Die IRIs werden aus drei Komponenten konstruiert:

1. einer benutzerdefinierten **Basis-IRI** (z. B. DOI-basierter Namespace),
2. dem **Dateinamen** der HDF5-Datei und
3. einem **Fragment**, das aus dem internen HDF5-Pfad sowie dem jeweiligen Objekttyp abgeleitet wird.

Die Basis-IRI sollte dabei idealerweise auf einem persistenten Identifikator (z.,B. DOI) oder einem institutionell stabilen Namensraum beruhen. Dateiname und Fragment sichern die lokale Eindeutigkeit innerhalb einer Sammlung gewährleisten. IRIs folgen dabei dem allgemeinen Muster *<schema>://<authority>/<path>#<fragment>* (vgl. (Dürst und Suignard, 2005)). Dieses Vorgehen unterstützt insbesondere die FAIR-Prinzipien F1 und I1, da eindeutige, maschinenlesbare Identifikatoren erzeugt werden, ohne zusätzliche infrastrukturelle Abhängigkeiten einzuführen.

Die Toolbox überführt die hierarchische HDF5-Pfadstruktur direkt in den Fragmentteil und er-

Listing A.3: Ergebnis der kontextuellen Metadaten im Turtle-Format, erzeugt durch den Python-Code in Listing A.2.

```
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix m4i: <http://w3id.org/nfdi4ing/metadata4ing#> .
@prefix schema: <https://schema.org/> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

<https://orcid.org/0000-0001-8729-0482> a foaf:Person ;
   m4i:orcidId "0000-0001-8729-0482" ;
   foaf:lastName "Probst" .

[] schema:version "2.4.0" .

[] a m4i:NumericalVariable ;
   m4i:hasUnit <http://qudt.org/vocab/unit/M-PER-SEC> .
```

weitert diese um Konventionen zur Referenzierung von Attributen und technischen Eigenschaften. Beispiele sind in Tabelle A.4 zusammengefasst. Alle Bestandteile werden URL-kodiert, um syntaktisch gültige IRIs auch bei nicht-trivialen Namen sicherzustellen.

HDF-Objekt	HDF-Fragment (Beispiel)
Group	*file.hdf/simulation/temperature*
Datensatz	*file.hdf/simulation/temperature/data*
Attribut	*file.hdf/simulation/temperature/data@units*
Eigenschaft	*file.hdf/simulation/temperature/data__dataspace*

Tabelle A.4: Beispiele von HDF-Fragmenten für verschiedene HDF5-Objekte, die durch *h5rdmtoolbox* automatisch konstruiert werden.

Serialisierung der HDF5-Struktur

Neben der kontextuellen Annotation fachlicher Inhalte ist für bestimmte Anwendungsfälle (z. B. Validierung, automatisierte Integration in Wissensgraphen oder Rekonstruktion von Datenstrukturen) eine explizite maschinenlesbare Repräsentation der internen HDF5-Struktur erforderlich. Die Toolbox stellt hierfür eine strukturelle Serialisierung bereit, bei der zentrale HDF5-Entitäten und Relationen (Gruppen, Datensätze, Attribute sowie ausgewählte technische Eigenschaften wie Dimensionen oder Datentypen) als RDF-Graph exportiert werden.

Die strukturelle Serialisierung basiert auf der hierarchischen Organisation von HDF5-Dateien, in der Gruppen und Datensätze über interne Pfade eindeutig adressierbar sind (z. B. */simulation/temperature*). Diese Pfadstruktur wird systematisch in den Fragmentteil der jeweils erzeugten IRIs überführt. Die resultierenden IRIs folgen damit unmittelbar der internen Dateistruktur

und erlauben eine eindeutige Identifikation einzelner HDF5-Objekte innerhalb einer publizierten Datei.

Die allgemeine Form der durch die Toolbox verwendeten IRIs lautet:

<Basis-IRI>#<Dateiname><HDF-Fragment> .

Die *Basis-IRI* verweist typischerweise auf einen DOI- oder institutionsbasierten Namensraum, unter dem die Datei als digitale Ressource veröffentlicht wird. Der Dateiname wird explizit in den Fragmentteil aufgenommen, um die Eindeutigkeit der Identifikatoren auch in Szenarien sicherzustellen, in denen mehrere HDF5-Dateien unter derselben Basis-IRI gemeinsam referenziert oder archiviert werden. Das anschließende Fragment wird direkt aus dem internen HDF5-Pfad des adressierten Objekts abgeleitet.

Die Verwendung des Fragmentbezeichners (#) stellt eine bewusste Entwurfsentscheidung dar. Da Fragmente clientseitig interpretiert werden, können Teilressourcen innerhalb einer Datei referenziert werden, ohne dass einzelne HDF5-Objekte als eigenständige Webressourcen aufgelöst oder publiziert werden müssen. Die Kodierung und Dekodierung der Fragmentbezeichner erfolgt automatisch durch die *h5rdmtoolbox*.

Speicherung von RDF-Informationen in HDF5 über reservierte Attribute

Da das HDF5-Format keine native Unterstützung für RDF-Strukturen vorsieht, speichert die Toolbox RDF-bezogene Informationen über zusätzliche, reservierte Attribute, ohne das Dateiformat selbst zu verändern. Damit bleiben die erzeugten Dateien vollständig kompatibel zu etablierten HDF5-Werkzeugen (z. B. *h5py*); die semantische Interpretation der reservierten Attribute erfolgt jedoch durch die Toolbox.

Die verwendeten Attributnamen folgen einer reservierten Namenskonvention (vollständig in Großbuchstaben), um Kollisionen mit benutzerdefinierten Attributen zu vermeiden und die automatische Erkennung zu erleichtern. Diese reservierten Attribute enthalten insbesondere die IRIs der jeweiligen RDF-Subjekt-, -Prädikat- und -Objektressourcen und stellen damit die Verbindung zwischen den HDF5-internen Strukturen und der externen semantischen Repräsentation her.

Die Entscheidung, die IRIs vollständig innerhalb der HDF5-Datei zu persistieren und nicht ausschließlich in externen RDF-Serialisierungen vorzuhalten, ist bewusst gewählt. Auf diese Weise bleibt die semantische Identität der internen Objekte auch dann erhalten, wenn eine Datei unabhängig von begleitenden Metadatendateien weitergegeben oder archiviert wird. Gleichzeitig können RDF-Graphen bei Bedarf vollständig aus der Datei rekonstruiert und in standardisierte Formate (z. B. Turtle oder JSON-LD) exportiert werden.

Tabelle A.5 gibt eine Übersicht der in der Toolbox verwendeten reservierten Attributnamen und deren Funktion.

Die Entscheidung für dieses Integrationsschema folgt zwei zentralen Anforderungen: (i) **Abwärtskompatibilität** durch ausschließliche Nutzung des etablierten HDF5-Attributmechanismus

Attributname	Beschreibung
RDF_OBJECT	Speichert objektbezogene RDF-Informationen zu einer Gruppe oder einem Datensatz.
RDF_PREDICATE_ATTR_NAME	Referenz auf die RDF-Prädikatressource (Prädikatszuordnung).
RDF_SUBJECT_ATTR_NAME	Enthält die Kennung der Subjektressource (typischerweise als IRI).
RDF_TYPE_ATTR_NAME	Platzhalter zur Speicherung von RDF-Typinformationen (z. B. Klassen) auf Objektebene.
RDF_FILE_PREDICATE_ATTR_NAME	Prädikatsbezogene RDF-Metadaten auf Dateiebene.
RDF_FILE_SUBJECT_ATTR_NAME	Subjektbezogene RDF-Metadaten auf Dateiebene.
RDF_FILE_OBJECT_ATTR_NAME	Objektbezogene RDF-Metadaten auf Dateiebene.
RDF_FILE_TYPE_ATTR_NAME	Platzhalter zur Speicherung des RDF-Typs auf Dateiebene.

Tabelle A.5: In *h5rdmtoolbox* verwendete reservierte Attributnamen zur Speicherung RDF-bezogener Informationen innerhalb von HDF5-Dateien.

und (ii) **algorithmische Auswertbarkeit** durch eindeutig erkennbare, erweiterbare Attributmuster. In Kombination mit der fragmentbasierten IRI-Konstruktion wird sichergestellt, dass semantisch angereicherte HDF5-Dateien sowohl in klassischen HDF5-Workflows nutzbar bleiben als auch konsistent in semantische Infrastrukturen und Wissensgraphen integriert werden können.

A.2.2 Generische Programmierschnittstellen zur Arbeit mit Ontologien

Der Beitrag dieser Arbeit beschränkt sich nicht auf die konzeptionelle Ausarbeitung eines FAIR-orientierten Forschungsdatenmanagements, sondern umfasst auch die Bereitstellung softwaretechnischer Schnittstellen zur praktischen Nutzung ontologischer Konzepte in datengetriebenen Arbeitsprozessen. Ontologien werden dabei nicht als externe Beschreibungsartefakte verstanden, sondern als integraler Bestandteil wissenschaftlicher Programmcodes.

Hierzu wurden Python-basierte Programmierschnittstellen entwickelt, die eine typsichere Modellierung, Validierung und Serialisierung ontologischer Konzepte erlauben. Die Bibliothek *ontolutils* stellt eine generische Schnittstelle zur Arbeit mit Ontologien bereit, auf der auch

weitere Werkzeuge dieser Arbeit aufbauen.

Die Bibliothek *ontolutils* ermöglicht die objektorientierte und typsichere Abbildung ontologischer Klassen in Python (Probst, 2026a). Obwohl Python Typannotationen unterstützt, erfolgt ohne zusätzliche Mechanismen keine Laufzeitvalidierung der modellierten Daten. Dies ist insbesondere bei semantischen Metadaten problematisch, da Inkonsistenzen häufig erst in nachgelagerten Verarbeitungsschritten sichtbar werden.

Zur Sicherstellung einer konsistenten und frühzeitigen Validierung integriert *ontolutils* die Bibliothek *pydantic*. Ontologieklassen werden dadurch als strikt validierte Datenmodelle implementiert, deren Instanzen automatisch auf Typkonsistenz und Vollständigkeit geprüft werden. Fehlerhafte oder unvollständige Angaben können somit frühzeitig erkannt werden, was die Robustheit und Verlässlichkeit der erzeugten Metadaten erhöht.

```python
from pydantic import EmailStr, Field

from ontolutils import Thing, urirefs, namespaces

@namespaces(
    prov="http://www.w3.org/ns/prov#",
    foaf="http://xmlns.com/foaf/0.1/"
)
@urirefs(
    Person='prov:Person',
    givenName='foaf:givenName',
    mbox='foaf:mbox'
)
class Person(Thing):
    givenName: str = Field(default=None, alias="given_name")
    mbox: EmailStr = Field(default=None, alias="email")

# Beschreibung einer Person:
person1 = Person(givenName="Erika", mbox="erikas@email.com")
# Folgendes wird einen Fehler wegen invalider E-Mail hervorrufen:
person2 = Person(given_name="Max", mbox="email.com")
```

Listing A.4: Beispiel einer Implementierung der Klasse Person aus der Ontologie *PROV* mittels *ontolutils*.

Die zentrale Basisklasse *Thing* repräsentiert die generische OWL-Klasse *owl:Thing* und bildet den gemeinsamen semantischen Ausgangspunkt aller modellierten Entitäten. Sie stellt grundlegende Eigenschaften zur semantischen Beschreibung bereit, darunter *rdfs:label*, *dcterms:relation* sowie Abbildungsrelationen wie *skos:closeMatch* und *skos:exactMatch*. Ontologische Konzepte können dadurch als Python-Objekte instanziiert, validiert und in standardisierte RDF-Formate überführt werden.
Die Zuordnung zwischen Python-Klassen bzw. -Attributen und ontologischen Konzepten erfolgt deklarativ über die Dekoratoren *@namespaces* und *@urirefs*. Diese verknüpfen Klassen- und

Attributnamen eindeutig mit IRIs aus definierten Namensräumen und stellen eine konsistente Verwaltung semantischer Referenzen sicher.

Listing A.4 zeigt exemplarisch die Implementierung der Klasse *Person* aus der Ontologie *PROV*. Die Attribute *givenName* und *mbox* sind dabei mit IRIs aus den Namensräumen *foaf* bzw. *prov* verknüpft. Durch die Verwendung des Typs *EmailStr* wird die syntaktische Korrektheit von E-Mail-Adressen automatisch validiert. Die zweite Instanziierung im Beispiel führt erwartungsgemäß zu einem Validierungsfehler.

Eine weitere zentrale Funktionalität von *ontolutils* ist die Serialisierung von Klasseninstanzen in RDF-kompatible Formate wie Turtle oder JSON-LD. Dadurch können modellierte Entitäten ohne zusätzliche Transformationsschritte direkt in semantische Infrastrukturen eingebunden werden.
Für das gezeigte Beispiel wird durch den Aufruf *person.serialize(format='json-ld', indent=2)* eine JSON-LD-Repräsentation erzeugt, die die ontologische Typisierung sowie alle zugeordneten Eigenschaften explizit enthält (vgl. Listing A.5).

Listing A.5: JSON-LD Serialisierung einer Beispielperson.

```
{
    "@context": {
        "owl": "http://www.w3.org/2002/07/owl#",
        "rdfs": "http://www.w3.org/2000/01/rdf-schema#",
        "prov": "http://www.w3.org/ns/prov#",
        "foaf": "http://xmlns.com/foaf/0.1/"
    },
    "@type": "prov:Person",
    "foaf:givenName": "John",
    "foaf:mbox": "john.doe@email.com",
    "@id": "_:N10ca0c5fba734dbcbd162836b1e9fbae"
}
```

Neben der Definition eigener Ontologieklassen stellt *ontolutils* im Submodul *ex* auch vordefinierte Klassenimplementierungen für ausgewählte, etablierte Ontologien und kontrollierte Vokabulare bereit. Dadurch wird der Einstieg in die Nutzung existierender Ontologien deutlich erleichtert und die Konsistenz mit externen Standards gefördert.
Ein zentrales Beispiel ist die Unterstützung der Ontologie *Metadata4Ing* (*M4I*), die im Kontext von *NFDI4Ing* als domänenübergreifendes Metadatenmodell für die Ingenieurwissenschaften etabliert ist. Die in *ontolutils* bereitgestellten Klassen orientieren sich eng an der formalen Ontologiedefinition und erlauben eine direkte Instanziierung entsprechender Konzepte auf Programmebene, wie das Beispiel in Listing A.6 zeigt. Hierfür wird die besonders praxisrelevante Klasse *NumericalVariable* verwendet, um den Wert eines Volumenstroms formal zu beschreiben. Das Beispiel beinhaltet ebenfalls das serialisierte Ergebnis in Turtle.

Ergänzend zu den generischen und vordefinierten Ontologieklassen stellt die Bibliothek *ssnolib* eine domänenspezifische Programmierschnittstelle für die Standardnamenontologie *SSNO*

Listing A.6: Instanziierung einer numerischen Variable gemäß der Ontologie *M4I* und Serialisierung in Turtle.

```python
from ontolutils.ex.m4i import NumericalVariable

var = NumericalVariable(
    id="http://example.org/variable/volume_flow_rate",
    label=["Volume Flow Rate@en", "Volumenstrom@de"],
    hasUnit="http://qudt.org/vocab/unit/M3-PER-SEC",
    hasNumericalValue=0.12,
    hasSymbol="VFR",
    hasVariableDescription="Volumetric flow rate at the fan inlet"
)

ttl = var.serialize(format="ttl")
print(ttl)

# @prefix m4i: <http://w3id.org/nfdi4ing/metadata4ing#> .
# @prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
# @prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
# @prefix ex: <http://example.org/> .
#
# ex:variable/volume_flow_rate a m4i:NumericalVariable ;
#     rdfs:label "Volumenstrom"@de,
#         "Volume Flow Rate"@en ;
#     m4i:hasNumericalValue 1.2e-01 ;
#     m4i:hasSymbol "VFR" ;
#     m4i:hasUnit <http://qudt.org/vocab/unit/M3-PER-SEC> ;
#     m4i:hasVariableDescription "Volumetric flow rate at fan inlet" .
```

bereit, die auf *ontolutils* aufbaut (Probst, 2025). Sie bildet die regelbasierte Struktur der Standardnamen explizit auf Programmebene ab und ermöglicht deren konsistente Erstellung und Serialisierung.

Ein minimales Beispiel zur Instanziierung eines Standardnamens ist in Listing A.7 dargestellt. Für weiterführende Funktionalitäten, insbesondere zur Arbeit mit vollständigen Standardnamentabellen und zur Regelprüfung, wird auf die Online-Dokumentation verwiesen.

Listing A.7: Anlegen eines Standardnamens mit *ssnolib*.

```python
import ssnolib
sn = ssnolib.StandardName(
    standardName="x_velocity",
    unit="m/s",
    description="The velocity in x-axis direction."
)
print(sn.serialize(format="ttl"))

```

```
# @prefix ssno: <https://matthiasprobst.github.io/ssno#> .
#
# [] a ssno:StandardName ;
#     ssno:description "The velocity in x-axis direction." ;
#     ssno:standardName "x_velocity" ;
#     ssno:unit "http://qudt.org/vocab/unit/M-PER-SEC" .
```

Insgesamt stellt *ontolutils* eine schlanke, generische Programmierschnittstelle zur Arbeit mit Ontologien in Python dar. Durch die Kombination aus objektorientierter Modellierung, Laufzeitvalidierung und standardkonformer RDF-Serialisierung wird eine robuste Grundlage geschaffen, auf der domänenspezifische Ontologieanwendungen sowie datengetriebene Workflows aufsetzen können.

A.3 Ontologien

A.3.1 HDF Ontologie

Die in dieser Arbeit verwendete HDF-Ontologie wird von der Allotrope Foundation definiert und verwaltet. An dieser Stelle soll lediglich ein Überblick über die relevanten Klassen und Eigenschaften gegeben werden, die in den Abbildungen und Codebeispielen dargestellt sowie in *h5rdmtoolbox* implementiert sind. Zur Vollständigkeit wird zudem auf die vollständige Definition der Ontologie verwiesen, die unter (Allotrope Foundation, 2024) zu finden ist.

Für HDF5 existiert bislang keine von der HDF5 Group offiziell veröffentlichte Ontologie. Zwar diskutieren Heber und Folk (2013) die grundsätzlichen Möglichkeiten einer OWL-basierten Repräsentation von HDF5-Strukturen und schlagen auch erste Klassen vor, eine ausgereifte Ontologie liegt jedoch nicht vor. Die in dieser Arbeit formulierte Zielsetzung, eine textbasierte RDF-Repräsentation von HDF5-Dateien zu erzeugen, knüpft an diese Ansätze an und entwickelt sie weiter.

Eine frei verfügbare Ontologie wird derzeit von der *Allotrope Foundation* bereitgestellt[1]. Sie ist unter einer Creative-Commons-Lizenz[2], sowie einer spezifischen HDF-Lizenz[3] veröffentlicht und wird im proprietären Allotrope Data Format (ADF) eingesetzt, das auf HDF5 aufbaut und insbesondere in der Pharmaindustrie etabliert ist. Um die Ontologie eindeutig von HDF5 selbst zu unterscheiden, wird im Folgenden die Bezeichnung **AF-HDF** (*Allotrope Foundation – HDF Ontology*) verwendet. Zentrale Klassen und Eigenschaften dieser Ontologie, wie *hdf:File*, *hdf:NamedObject* und *hdf:Attribute*, sind in Abbildung A.3 dargestellt.

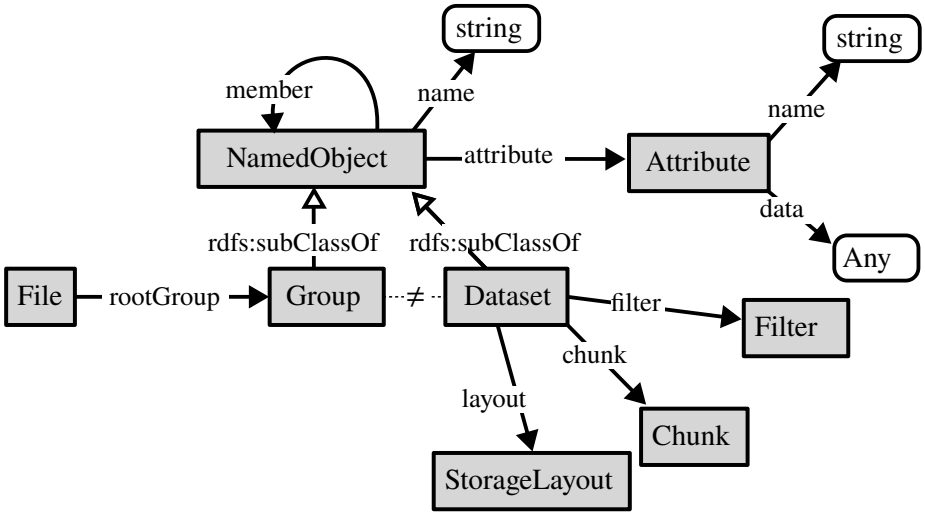


Abbildung A.3: Auswahl zentraler Klassen der AF-HDF Ontologie mit den ausgewählten Eigenschaften. Für eine vollständige Darstellung sei auf die Ontologiedefinition selbst verwiesen.

[1] https://purl.allotrope.org/voc/adf/REC/2024/12/hdf.ttl
[2] http://purl.allotrope.org/voc/creative-commons-attribution-license
[3] http://purl.allotrope.org/voc/hdf-license

Listing A.8: Beschreibung des Zenodo Datensatzes, der die CAD Datei des generischen Radialventilators beinhaltet

```
@prefix adms: <http://www.w3.org/ns/adms#> .
@prefix dcat: <http://www.w3.org/ns/dcat#> .
@prefix dcmitype: <http://purl.org/dc/dcmitype/> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix org: <http://www.w3.org/ns/org#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

<https://doi.org/10.5281/zenodo.14551648> dcterms:identifier
    "https://doi.org/10.5281/zenodo.14551648" .

<https://orcid.org/0000-0001-8729-0482> a foaf:Person ;
    org:memberOf <https://ror.org/https://ror.org/04t3en479> ;
    foaf:familyName "Probst" ;
    foaf:givenName "Matthias" ;
    foaf:name "Probst, Matthias" .

<https://orcid.org/0000-0001-9560-500X> a foaf:Person ;
    org:memberOf <https://ror.org/https://ror.org/04t3en479> ;
    foaf:familyName "Pritz" ;
    foaf:givenName "Balazs" ;
    foaf:name "Pritz, Balazs" .

<https://ror.org/https://ror.org/04t3en479> a foaf:Organization ;
    dcterms:identifier "https://ror.org/04t3en479"^^xsd:anyURI ;
    foaf:name "Karlsruhe Institute of Technology" .

<https://doi.org/10.5281/zenodo.14551649> a dcat:Dataset ;
    dcterms:creator <https://orcid.org/0000-0001-8729-0482>,
        <https://orcid.org/0000-0001-9560-500X> ;
    dcterms:description "The CAD file of the generic centrifugal ..." ;
    dcterms:identifier
    "https://doi.org/10.5281/zenodo.14551649"^^xsd:anyURI ;
    dcterms:isVersionOf <https://doi.org/10.5281/zenodo.14551648> ;
    dcterms:issued "2024-12-29"^^xsd:date,
        "2024"^^xsd:gYear ;
    dcterms:language
    <http://publications.europa.eu/resource/authority/language/ENG> ;
    dcterms:modified "2024-12-29"^^xsd:date ;
    dcterms:publisher [ a foaf:Agent ;
        foaf:name "Zenodo" ] ;
```

```
41      dcterms:title "Generic Centrifugal Fan CAD File" ;
42      dcterms:type dcmitype:Dataset ;
43      owl:sameAs <https://www.wikidata.org/wiki/Q131549102>,
44          <https://zenodo.org/records/14551649> ;
45      owl:versionInfo "1.0.0" ;
46      adms:identifier [ a adms:Identifier ;
47          skos:notation
      "https://www.wikidata.org/wiki/Q131549102"^^xsd:anyURI ;
48          adms:schemeAgency "URL" ],
49      [ a adms:Identifier ;
50          skos:notation "oai:zenodo.org:14551649" ;
51          adms:schemeAgency "oai" ],
52      [ a adms:Identifier ;
53          skos:notation "https://zenodo.org/records/14551649"^^xsd:anyURI ;
54          adms:schemeAgency "URL" ] ;
55          dcat:distribution [ a dcat:Distribution ;
56          dcat:accessURL <https://doi.org/10.5281/zenodo.14551649> ;
57          dcat:byteSize "15927" ;
58          dcat:downloadURL
      <https://zenodo.org/records/14551649/files/metadata.jsonld> ],
59      [ a dcat:Distribution ;
60          dcterms:license
      <https://creativecommons.org/licenses/by/4.0/legalcode> ;
61          dcat:accessURL <https://doi.org/10.5281/zenodo.14551649> ],
62      [ a dcat:Distribution ;
63          dcat:accessURL <https://doi.org/10.5281/zenodo.14551649> ;
64          dcat:byteSize "14278906" ;
65          dcat:downloadURL
      <https://zenodo.org/records/14551649/files/cefa_asm_v1.igs> ] ;
66      dcat:keyword "CAD",
67      "centrifugal fan",
68      "open centrifugal fan database",
69      "opencefadb" ;
70      foaf:page <https://doi.org/10.5281/zenodo.14551649> .
```

Ein Datensatz mit dem Namen *u* hat das Attribut *Einheit*. Somit ist im semantischen Sinne *u* eine numerische Variable, die eine Einheit hat. Dies kann einfach mit der *Metadata4Ing* Ontologie (Iglezakis et al., 2023) modelliert werden: Allerdings ist hier zu beachten, dass dies nur für einzelne Zahlenwerte und nicht für i. A. mehrdimensionale Arrays praktikabel ist. Die Beschreibung mittels m4i ist eher für Eigenschaften (z. B. Einstellung in Software oder Hardware). Die Stärke von HDF5 liegt aber eben im Speichern großer Datensätze. Es ist also zielführender, die Dimension abzuspeichern und geeignete Namen zu verwenden (und standard_name)

```
PREFIX m4i: <http://w3id.org/nfdi4ing/metadata4ing#>

<ex#NumVar>
    a m4i:NumericalVariable ;
    m4i:hasNumericalValue 4.5 ;
    m4i:hasUnit "m/s" ;
    m4i:hadRole <https://orcid.org/0000-0001-8729-0482> .
```

Listing A.9: Beispiel für die Beschreibung eines HDF5-Datensatzes mittels RDF.

A.3.2 CodeMeta

Zur Beschreibung der im Rahmen dieser Arbeit entwickelten Softwareartefakte wird das Schema CodeMeta in Form von JSON-LD verwendet (Jones et al., 2023). Die CodeMeta-Dateien enthalten zentrale softwarebezogene Metadaten, darunter persistente Identifikatoren, Lizenzinformationen, Quellcode-Repositorien, Versionierung, unterstützte Programmiersprachen und Betriebssysteme sowie Angaben zu Autoren und institutioneller Zugehörigkeit.

Listing A.10 zeigt exemplarisch die CodeMeta-Beschreibung des Pythonpakets *opencefadb*. Die maschinenlesbare Repräsentation ermöglicht eine eindeutige Identifikation, automatisierte Indexierung und strukturierte Zitierfähigkeit der Software und unterstützt damit deren nachhaltige Nachnutzung im Kontext des Forschungsdatenmanagements.

Listing A.10: Codemeta JSON-LD-Inhalt für das Pythonpaket opencefadb.

```
{
  "@id": "https://doi.org/10.5281/zenodo.18368245",
  "@context": "https://doi.org/10.5063/schema/codemeta-2.0",
  "@type": "SoftwareSourceCode",
  "license": "https://spdx.org/licenses/GPL-3.0-only",
  "codeRepository":
    "git+https://github.com/matthiasprobst/opencefadb.git",
  "name": "opencefadb",
  "version": "1.0.0",
  "description": "OpenCeFaDB - A FAIR Database for a Generic Centrifugal
    Fan",
  "applicationCategory": "Engineering",
  "programmingLanguage": ["Python 3", "Python 3.10", "Python 3.11",
    "Python 3.12", "Python 3.13"],
  "operatingSystem": ["Linux", "Windows", "macOS"],
  "author": [
    {
      "@type": "Person",
      "@id": "https://orcid.org/0000-0001-8729-0482",
      "givenName": "Matthias",
      "familyName": "Probst",
```

```
      "email": "matth.probst@gmail.com",
      "affiliation": {
        "@type": "Organization",
        "@id": "https://ror.org/04t3en479",
        "name": "Karlsruhe Institute of Technology, Institute of Thermal Turbomachinery"
      }
    }
  ]
}
```

A.3.3 Simple Standard Name Ontology

Die Simple Standard Name Ontology (SSNO) in der Version v2.2.0 ist auf Zenodo veröffentlicht (Probst und Pritz, 2025c) und wird in einem öffentlich zugänglichen Git-Repository gepflegt. Das Repository dient zugleich der kollaborativen Weiterentwicklung, der Versionierung sowie der Bereitstellung der Ontologie-Dokumentation und einer begleitenden Projektwebsite. Die Dokumentation wurde automatisiert mit dem Werkzeug *Widoco* (Garijo, 2017) erzeugt und wird über GitHub Pages gehostet.

Die Modellierung der SSNO folgt dem Prinzip der maximalen Wiederverwendung etablierter Ontologien und Vokabulare. Wo immer möglich, werden existierende Konzepte übernommen und lediglich domänenspezifisch ergänzt. Abweichungen von etablierten Standards erfolgen ausschließlich dort, wo dies aus funktionalen oder methodischen Gründen erforderlich ist und explizit begründet werden kann.

Ein zentrales Beispiel hierfür ist die menschenlesbare Beschreibung von Standardnamen. Diese wird in der SSNO bewusst als Daten-Property (*ssno:description*) modelliert und nicht als Annotations-Property wie etwa *skos:definition* oder *dcterms:description*. Hintergrund dieser Entscheidung ist, dass Annotations-Properties nicht Gegenstand ontologischer Schlussfolgerungen sind und somit weder Kardinalitätsrestriktionen noch Ableitungsregeln darauf angewendet werden können. Für die SSNO ist jedoch sichergestellt, dass jeder Standardname mindestens eine formale Beschreibung besitzt und dass diese Beschreibung bei der algorithmischen Ableitung neuer Standardnamen systematisch neu konstruiert werden kann. Diese Anforderungen lassen sich konsistent nur durch die Verwendung eines Daten-Properties abbilden.

Eine weitere bewusste Abweichung betrifft die Behandlung physikalischer Einheiten. In den originalen CF Conventions werden Einheiten über sogenannte *canonical units* beschrieben, die informell dokumentiert sind und primär auf einer textuellen Referenz basieren. Eine direkte ontologische Integration dieses Konzepts ist nur eingeschränkt möglich. Vor diesem Hintergrund verzichtet die SSNO auf vollständige Kompatibilität zur aktuellen CF-Standardnamensliste und bevorzugt stattdessen die Verwendung von SI-konformen Einheiten aus der QUDT-Ontologie. Dadurch wird eine formale, maschinenlesbare und konsistent referenzierbare Beschreibung physikalischer Größen ermöglicht, die insbesondere für automatische Validierung, Inferenz und Weiterverarbeitung von Vorteil ist.

Zur Beschreibung und Verwaltung von Standardnamentabellen nutzt die SSNO etablierte Metadatenvokabulare. Tabelle A.6 zeigt ausgewählte Eigenschaften aus SKOS, die zur semantischen Beschreibung, Auffindbarkeit und Pflege von Standardnamentabellen eingesetzt werden. Ergänzend werden Dublin-Core-Terme verwendet, um Versionierungs- und Lebenszyklusinformationen explizit abzubilden (vgl. Tabelle A.7).

Insgesamt operationalisiert die SSNO das ursprünglich tabellarische Konzept der CF Standardnamen erstmals als inferenzfähige Ontologie und schafft die methodische Grundlage für eine konsistente, maschinenlesbare und automatisierbare Ableitung standardisierter physikalischer Bezeichnungen.

SKOS-Eigenschaft	**Erläuterung**
skos:prefLabel	Bevorzugter, aussagekräftiger Name einer Standardnamentabelle.
skos:altLabel	Alternativer Name oder Synonym, erleichtert die Auffindbarkeit.
skos:scopeNote	Beschreibung des Anwendungsbereichs der Tabelle.
skos:example	Beispiel für eine konkrete Nutzung der Tabelle.
skos:related	Verweis auf verwandte Konzepte oder Tabellen.
skos:note	Freitext-Informationen, die nicht in andere Eigenschaften passen.
skos:changeNote	Dokumentation von Änderungen, z. B. Ergänzung neuer Standardnamen.
skos:editorialNote	Redaktionelle Hinweise zur Pflege und Verwaltung.

Tabelle A.6: Ausgewählte Eigenschaften aus SKOS zur Beschreibung und Verwaltung von Standardnamentabellen in SSNO.

Dublin-Core-Eigenschaft	**Nutzung für Standardnamentabellen**
dcterms:hasVersion	Angabe der Versionsnummer einer Tabelle.
dcterms:created	Erstellungsdatum der jeweiligen Version.
dcterms:modified	Änderungsdatum der jeweiligen Version.
dcterms:description	Textuelle Beschreibung der Tabelle.
dcterms:subject	Thematische Einordnung, z. B. über eine Wikidata-URI.

Tabelle A.7: Ausgewählte Eigenschaften aus dem Dublin-Core-Vokabular zur Versionierung und inhaltlichen Beschreibung von Standardnamentabellen im Kontext der SSNO.