HELMHOLTZ SPITZENFORSCHUNG FÜR GROSSE HERAUSFORDERUNGEN

HMC

Helmholtz Metadata Collaboration Platform

HELMHOLTZ SPITZENFORSCHUNG FÜR GROSSE HERAUSFORDERUNGEN

Helmholtz Metadata Collaboration (HMC) Plattform

15.04.2019

Konzept vorgelegt von

Deutsches Krebsforschungszentrum
Janko Ahlbrandt
Rumyana Proynova
Frank Ückert

Geomar Helmholtz-Zentrum für Ozeanforschung Kiel
Sören Lorenz

Helmholtz-Zentrum Berlin für Materialien und Energie
Ants Finke

Helmholtz-Zentrum München – Deutsches Forschungszentrum für Gesundheit und Umwelt
Wolfgang zu Castell

Helmholtz Zentrum Potsdam – Deutsches Geoforschungszentrum
Roland Bertelmann
Kirsten Elger
Jörn Lauterjung

Karlsruher Institut für Technologie
Nanette Rißler-Pipka
Rainer Stotzka

in Zusammenarbeit mit der
Arbeitsgruppe „Mehrwerte aus Forschungsdaten durch Metadaten“,
im Rahmen des Helmholtz-Inkubators Information & Data Science

Koordination der Überarbeitung im Auftrag des Präsidiums

Antje Boetius, Alfred-Wegener-Institut

Wiss. Vertreter für die HGF Forschungsbereiche

Forschungsbereich Energie
Veit Hagenmeyer, Karlsruher Institut für Technologie

Forschungsbereich Erde und Umwelt
Frank Oliver Glöckner, Alfred-Wegener-Institut

Forschungsbereich Gesundheit
Alice McHardy, Helmholtz-Zentrum für Infektionsforschung

Forschungsbereich Luftfahrt, Raumfahrt und Verkehr
Tobias Schlauch, Deutsches Zentrum für Luft- und Raumfahrt

Forschungsbereich Materie
Michael Bussmann, Helmholtz-Zentrum Dresden Rossendorf

Forschungsbereich Information
Michael Denker, Forschungszentrum Jülich

Inhaltsverzeichnis

Zusammenfassung

Die Helmholtz Metadata Collaboration Plattform ist eine dezentrale Einrichtung mit übergreifenden Diensten, Beratungsangeboten, Informationen und Werkzeugen zum effizienten Umgang mit Metadaten, das die Expertisen aus den Fachdomänen und Forschungsbereichen auf diesem Gebiet zusammenführt.

In einem interdisziplinären Forschungsumfeld wie der Helmholtz-Gemeinschaft und ihren internationalen Partnern müssen Forschungsdaten auf höchstem Niveau generiert, ausgewertet, ausgetauscht, annotiert, gespeichert und in neuen Kontexten wiederverwendet werden können. Ein leistungsstarkes und zukunftsfähiges Forschungsdatenmanagement stärkt die Effektivität und Effizienz der Forschung, sichert Forschungsergebnisse langfristig und erhöht das Vertrauen durch die Reproduzierbarkeit der Ergebnisse (Helmholtz-Gemeinschaft 2016). Metadaten sind essentielle Informationen über Forschungsdaten, die für deren Auffinden und Verstehen sowie für deren Vernetzung und Nachnutzung im Sinne der FAIR-Prinzipien erforderlich sind. Wie Metadaten erhoben und in welcher Form sie gespeichert werden, so dass andere sie in Zukunft verwenden können, und das Wissen um die Analysemöglichkeiten von Metadaten, stellen wichtige Schlüsselkompetenzen dar, die innerhalb der Helmholtz-Gemeinschaft durch HMC langfristig ausgebaut werden.

Die Helmholtz Metadata Collaboration Plattform (HMC) wird die qualitative Anreicherung von Forschungsdaten durch Metadaten vorantreiben sowie organisatorisch und technisch umsetzen. Sie führt die wissenschaftliche Expertise zum Thema Metadaten aus den Domänen in den Metadata Hubs der Forschungsbereiche zusammen, zeigt die Bedeutung von Metadaten innerhalb des Forschungsdatenmanagements auf, bietet Beratung und gewährleistet langfristig nachhaltige Infrastrukturdienste zum Speichern, Nachnutzen und zum internationalen Austausch der Metadaten. Vor dem Hintergrund dieser nachhaltig und langfristig ausgerichteten Angebote und der konkreten wissenschaftlichen Vorteile sind die Forschenden in den Domänen bereit, Ressourcen in die Anreicherung ihrer Forschungsdaten mit aussagekräftigen Metadaten zu investieren. Die Daten können durch Metadaten gefunden werden, sie sind zugreifbar, interoperabel und können disziplinär wie auch interdisziplinär wiederverwendet werden. Die Zitierung und Wiederverwendung steigert die internationale Sichtbarkeit und die hohe Qualität der Forschungsdaten (durch die Beschreibung mit Metadaten) erhöht die wissenschaftliche Reputation. Durch das Zusammenführen von Forschungs- und Metadaten aus verschiedenen Quellen entstehen wissensbasierte Informationssysteme, die die Analyse von übergeordneten Zusammenhängen und damit die Generierung von neuen Erkenntnissen erlauben.

Die HMC Plattform basiert auf einer Struktur aus dezentralen (Metadata Hubs) und zentralen (HMC Office, technische Komponenten), mit statischen (Metadata Hubs, HMC Office, technische Komponenten) und dynamischen Elementen (Inkubator-Projekte). Die Forschungsbereiche bringen Kompetenzen, Ideen und Anforderungen ihrer Fachdomänen über die Metadata Hubs ein. Zentral organisiert werden generisch nutzbare Prozesse, technische Lösungen sowie Bildungs-/Beratungsangebote entwickelt und den Metadata Hubs zur fachspezifischen Anpassung und Nutzbarmachung bereitgestellt.

Summary

The Helmholtz Metadata Collaboration Platform provides comprehensive services, consulting, information and tools for efficient metadata handling as a distributed shared facility. Community-Expertise for metadata of the six Helmholtz research areas (Energy; Earth and Environment; Health; Matter; Information; Aeronautics, Space and Transport) is jointly developed, shared and consolidated on the platform.

In an interdisciplinary research environment such as the Helmholtz Association of German Research Centers and its international partners, outstanding research data are generated, evaluated, exchanged, annotated, stored and reused in novel contexts. The effectiveness and efficiency of research is improved by powerful and future-oriented research data management. It preserves research results in the long term and increases confidence in the data through the reproducibility of the results (Helmholtz-Gemeinschaft 2016). Metadata are essential information for finding, understanding and reusing research data and for implementing the FAIR Principles. Within the Helmholtz Association, the Helmholtz Metadata Collaboration Platform establishes key competences for collecting, storing, analysing and re-using metadata for future purposes.

The Helmholtz Metadata Collaboration (HMC) Platform aims at fostering the enrichment of research data with metadata, and implements supporting services for all domains by consolidating the expertise of the six Helmholtz research areas in Metadata Hubs. It introduces the subject metadata to the stakeholders, highlights the importance of metadata within research data management, offers advice, and facilitates sustainable infrastructure services for the storage, reuse and international exchange of metadata. Supported by a reliable framework and aware of the scientific rewards, researchers are willing to invest in the metadata enrichment of their research data. The data are findable through metadata; they are accessible, interoperable, and can be reused in an interdisciplinary context. Citation and reuse of high-quality research data increase the international visibility and scientific reputation of researchers. Novel research insights with generalised concepts are enabled by the aggregation of data and metadata from various sources in knowledge-based information systems.

The HMC platform is based on a structure of central (HMC Office, technical components) and distributed (Metadata Hubs) components with static (Metadata Hubs, HMC Office, technical components) and dynamic elements (Inkubator projects). The research areas contribute competences, ideas and requirements of their domains via the Metadata Hubs. Generically usable processes, technical solutions as well as educational/consulting services, are created centrally and made available to the Metadata Hubs for specific adaptation and utilization.

1 Motivation

Forschungsdaten sind für eine Nachnutzung unter wissenschaftlichen Qualitätsstandards geeignet, wenn sie mit einer Beschreibung über die Art und Organisation der Daten, Information über deren Zustandekommen sowie Angaben, die Rückschlüsse auf die Präzision und Qualität der Daten zulassen, angereichert werden. Die Sammlung all dieser, einen Datensatz beschreibenden Attribute, definieren einen eigenen Datensatz, den man als Metadaten bezeichnet. Metadaten werden damit selbstverständlich erhoben, wenn Forschungsdaten organisiert, strukturiert erfasst und abgelegt werden. Der Prozess der Metadatenerhebung startet konsequenterweise schon weit vor der Erhebung der eigentlichen Experimental- und Simulationsdaten bei der Planung und Umsetzung wissenschaftlicher Experimente und Simulationen. Metadaten sind damit als umfassende und exakte Beschreibung des Datensatzes auch eine unabdingbare Komponente, ohne die eine Nachvollziehbarkeit bei der Datennahme, Auswertung, Wiederverwendung und Präsentation von Forschungsdaten im Sinne guter wissenschaftlicher Praxis nicht gegeben ist. Sie erlauben darüber hinaus eine weiterführende interdisziplinäre Nutzung von Forschungsdaten, da sie es ermöglichen, weitere Forschungsdatensätze zuzuordnen, wie beispielsweise solche, die synchron in Raum oder Zeit erhoben wurden.

In Zeiten der Digitalisierung und Big Data Analytics geht die Bedeutung von Metadaten über die qualitativ beschreibende Funktion hinaus. Metadaten müssen – wie die eigentlichen Forschungsdaten auch – auffindbar und automatisiert auswertbar zur Verfügung stehen. Nur so können, mittels der weitgehend automatisierten Erfassung von Metadaten, die durch sie beschriebenen Datensätze überhaupt suchbar und für eine Nachnutzung zugänglich gemacht werden. Den Metadaten kommt damit die Rolle eines entscheidenden Bindeglieds zu, welches Wissenschaftstreibende, die an der Nutzung von Daten interessiert sind, mit den Datenquellen selbst verbindet.

Spitzenforschung impliziert eine exzellente formale und inhaltliche Qualität aller im Forschungsprozess beteiligten Komponenten. Dies stellt insbesondere hohe Anforderungen an die Qualität von in wissenschaftlichen Projekten erhobenen Daten und Metadaten. Schon das Fehlen von wenigen, im wissenschaftlichen Prozess entstehenden, Informationen kann die Beurteilung wissenschaftlicher Ergebnisse erschweren, eine Nachvollziehbarkeit gegebenenfalls unmöglich machen, und Analysen aus einer Nachnutzung der Daten durch Dritte verfälschen. Dabei sind die Quellen von Metadaten verteilt. Informationen über die an einem Experiment beteiligten Personen, deren Aufgaben und Verantwortungen, speisen sich aus der Organisation von Großgeräten, Forschungszentren, -instituten und -gruppen. Standard Operation Procedures (SOPs) definieren Einstellungen von experimentellen Geräten, Verfahren zur Qualitätskontrolle und Kalibrierung sowie Prozessabläufe bei experimentellen Arbeiten. Angaben von Herstellern beschreiben Art und Zusammensetzung von Laborkits oder geben Aufschluss über Betriebs- und Funktionsweise experimenteller Anlagen. Experimentelle Abläufe werden derzeit meist in Laborbüchern dokumentiert und archiviert. Das Prinzip der verteilten Quellen von Metadaten gilt auch für die Modellierung, Simulation und Datenanalyse, deren Ergebnisse nur nachvollziehbar werden, wenn alle zur Datenproduktion notwendigen Informationen wie Software- und Hardwareumgebungen sowie genutzte Datensätze und Workflows zur Nachvollziehbarkeit zur Verfügung stehen. Da mittlerweile viele der wissenschaftlichen Arbeiten auf der zunehmend komplexen Analyse von Daten beruhen, wird die Dokumentation und Archivierung der dafür eingesetzten Softwarewerkzeuge und Parameter zunehmend wichtiger. Nur durch eine möglichst lückenlose Provenienz können die gewonnenen Erkenntnisse auch reproduziert, zugeordnet und nachverfolgt werden. Über die Provenienz hinaus muss der Einsatz von Softwarewerkzeugen sowohl zur Datenanalyse wie zur Simulation mit der zugrundeliegenden Daten-, Software- und Hardwareinfrastruktur sowie der genutzten Workflows annotiert werden. In Anlehnung an die SOPs müssen die zugehörigen Abläufe

soweit als möglich in Hinsicht auf die Metadatenabnahme automatisiert und standardisiert werden, um die Reproduzierbarkeit von Datenanalysen und Simulationen zu gewährleisten.

Qualitativ exzellente Metadaten erfassen nicht nur all diese, für die Beurteilung wissenschaftlicher Daten bedeutenden Aspekte, sondern setzen sie auch in einen Kontext, der eine Vergleichbarkeit über institutionelle und fachliche Grenzen hinaus ermöglicht. Dies wird durch Metadatenschemata erreicht, die Standards und Best Practices definieren und institutionsübergreifend abgestimmt werden. Damit ist unmittelbar klar, dass die Definition, Erfassung und Bereitstellung geeigneter Metadaten innerhalb des jeweiligen fachlichen, internationalen Kontextes erfolgen muss. Dies gilt insbesondere für Metadatenstandards, Ontologien und Minimalanforderungen zur Qualitätssicherung. Doch kann die Organisation von Metadaten nicht innerhalb der Grenzen fachlicher Disziplinen stehen bleiben. Gerade das Potenzial, das sich durch die Verknüpfung und Auswertung großer, unterschiedlicher Datensätze ergibt, setzt voraus, dass Daten auch über Fachgrenzen hinaus kombiniert und ausgewertet werden können. Damit stellen fachübergreifend aufgestellte, automatisiert lesbare Metadatenschemata eine der Grundvoraussetzungen dar, um Big Data Analytics innerhalb der Wissenschaft zu entfalten. Unterschiedliche Standardisierungen und Ontologien, die historisch aus verschiedenen Fachgebieten heraus entstanden sind, müssen so ergänzt werden können, dass es gelingt, beispielsweise Genomsequenzierungsdaten aus der Umweltforschung mit Datensätzen aus der biomedizinischen Forschung zu verbinden und diese im Kontext von sozialen- und Umweltveränderungen zu betrachten. Gerade an Großforschungsanlagen kann hierbei durch die gemeinsame Nutzung von Quellen und Instrumenten durch verschiedene wissenschaftliche Gemeinschaften ein erheblicher Fortschritt zur Vereinheitlichung von Seiten der Großgerätebetreiber in Zusammenarbeit mit den Großgerätenutzern gemacht werden, um Standards für die interdisziplinäre Metadatenerfassung zu setzen. Eine große Herausforderung sind dabei die hohen Raten, großen Mengen und die zunehmende Komplexität der Daten, die auch die automatische Metadatenabnahme zum Beispiel an Maschinen, großen Detektoren, Lichtquellen und zentralen Experimentalaufbauten zu einer sowohl technischen wie auch organisatorischen Herausforderung machen. Durch eine vereinheitlichte Metadatenerfassung an Großgeräten werden entscheidende Voraussetzungen geschaffen, um zum Beispiel Materialwissenschaften und Biologie sowohl in der Beschreibung sowie im Verständnis der betrachteten Systeme, basierend auf ihren atomaren Grundlagen, näher zusammenzubringen.

Die professionelle, effiziente und vorausschauende Erfassung von Metadaten scheitert heute oft am Mehraufwand, den eine geeignete Organisation und Speicherung der Metadaten unter Berücksichtigung vielfältiger, sich dynamisch verändernder, Standards erfordert. Dabei erlauben die Möglichkeiten der Digitalisierung schon heute eine Unterstützung dieser Prozesse durch geeignete Technologien und Datenservices. Viele der Prozesse und Werkzeuge, die zur Pflege guter Metadaten notwendig sind, unterscheiden sich dabei formal und technologisch nicht zwischen den verschiedenen Fachdomänen. Das bedeutet, es besteht ein erhebliches Potenzial durch die Erarbeitung allgemeiner, domänenunabhängiger Werkzeuge Synergien zu schaffen, die den zusätzlichen Aufwand reduzieren, der für eine zukunftsweisende Organisation von Metadaten unweigerlich erforderlich ist. Die am Endnutzer und am Ziel der vereinfachten Metadatenerfassung orientierte Entwicklung geeigneter Metadatenservices ist eine Grundvoraussetzung dafür, die notwendige Akzeptanz bei den Wissenschaftlerinnen und Wissenschaftlern zu schaffen, die Anreicherung ihrer Daten durch geeignete Metadaten über den unmittelbaren eigenen Nutzen hinaus zu betreiben. Dabei muss grundsätzlich die internationale Wissenschaftsgemeinschaft der jeweiligen Fachdomäne in den Blick genommen werden.

Von der Helmholtz-Gemeinschaft als Betreiber komplexer Großforschungsanlagen wird selbstverständlich erwartet, dass die Zentren der Gemeinschaft auch eine tragende Rolle bei der Aufarbeitung

und Bereitstellung qualitativ hochwertiger Forschungsdaten übernehmen. Eine solche Erwartung steht vor dem Hintergrund entsprechender Regelungen und Basisanforderungen an das Management von Forschungsdaten. Diese sind insbesondere die forschungspolitischen Leitlinien zum Umgang mit Forschungsdaten der EU (European Commission 2016), der Deutschen Forschungsgemeinschaft (Deutsche Forschungsgemeinschaft 2015), der Allianz der deutschen Wissenschaftsorganisationen (Franke u. a. 2015) und anderer Zuwendungsgeber. Auch die Helmholtz-Gemeinschaft hat sich die Selbstverpflichtung auferlegt, Richtlinien für das Forschungsdatenmanagement an den Zentren der Gemeinschaft zu etablieren (Helmholtz-Gemeinschaft 2016), und erarbeitet derzeit dazu eine übergreifende Digitalisierungsstrategie.

In diesem Kontext haben sich international die FAIR Data Prinzipien (Wilkinson u. a. 2016) als Leitmotive guten Forschungsdatenmanagements etabliert. Das Akronym FAIR steht dabei für die Prinzipien der Auffindbarkeit (Findable), Zugänglichkeit (Accessible), Interoperabilität (Interoperable) und Wiederverwendbarkeit (Reusable). Gerade die beiden letztgenannten Prinzipien beinhalten den Anspruch, Forschungsdaten automatisiert zusammenzuführen und auswerten zu können. Die Forderung, Forschungsdaten weltweit offen und harmonisiert einem breiten Publikum auch jenseits akademischer Grenzen zur Verfügung zu stellen, zieht sich auch durch die forschungspolitischen Empfehlungen des Rats für Informationsinfrastrukturen (RFII – Rat für Informationsinfrastrukturen 2017), die insbesondere zum Aufbau einer Nationalen Forschungsdateninfrastruktur (NFDI) führen sowie der Europäischen Kommission im Rahmen des Aufbaus der European Open Science Cloud (EOSC) mit dem 2018 verabschiedeten Handlungsplan „Turning FAIR into Reality“ (Hodson u. a. 2018).

Mit dem Aufbau der Helmholtz Metadata Collaboration (HMC) Plattform stellt sich die Helmholtz-Gemeinschaft dieser Herausforderung und übernimmt eine führende Rolle bei der praktischen Umsetzung. HMC nutzt dabei die Stärke der Gemeinschaft, durch die Zusammenführung fächerübergreifender Expertise Synergien zu schaffen, die innerhalb einer fachlichen Disziplin nicht oder nur sehr viel schwerer zu entfalten wären. Zugleich sind die Zentren der Helmholtz-Gemeinschaft innerhalb der jeweiligen Fachdomänen der Forschungsbereiche weltweit führend und in die internationale Forschergemeinschaft eingebunden. Damit wird es auf der einen Seite möglich, die Brücke zwischen den Anforderungen und der notwendigen Abstimmung innerhalb der Fachdisziplinen zu schlagen, und auf der anderen Seite durch die Schaffung allgemeiner Datendienste die praktische Umsetzung professionellen Datenmanagements insbesondere auf dem Feld der Metadaten zu beschleunigen. Gute Erfahrungen liegen innerhalb der Gemeinschaft bereits vor, wie beispielsweise im Rahmen des H2020 Flagship Human Brain Projects, des Portals Deutsche Meeresforschung und an den Großforschungsanlagen im Forschungsbereich Materie. Unter anderem als Anknüpfungspunkt zu HMC wird der Forschungsbereich Materie das Topic Data Management & Analysis im Programm Matter & Technologies einführen und so eine Koordination des HMC mit dem gesamten Forschungsbereich Materie anbieten können. Auch im FB Erde und Umwelt werden die Aktivitäten im Bereich des Datenmanagement gebündelt, und tragen zu einer vernetzten nationalen Dateninfrastruktur bei (NFDI Prozess).

Eine Kartierung und Zusammenführung einzelner Aktivitäten und Kompetenzen in der Entwicklung von Metadatenservices ist dabei das Gebot der Stunde. Die Gemeinschaft als Ganzes kann von einer Landkarte der metadatenrelevanten Kompetenzen erheblich profitieren, daher ist die Erstellung dieser eine zentrale Aufgabe des HMC. So adressiert beispielsweise der Rat für Informationsinfrastrukturen explizit die Gefahr der vorhandenen und weiter zunehmenden Fragmentierung von Infrastruktursystemen weltweit, die insbesondere aus den finanziellen Rahmenbedingungen durch den Vorzug der Projektförderung entsteht (Wedlich u. a. 2017). HMC steht diesem Trend diametral

entgegen, indem sich die Helmholtz-Gemeinschaft eine Plattform schafft, die das Thema Organisation und Bereitstellung von Metadaten im Rahmen eines professionellen Forschungsdatenmanagements mit seinen dynamischen Komponenten langfristig und nachhaltig finanziert. Durch das starke Engagement von Vertreterinnen und Vertretern der Gemeinschaft an Aktivitäten im Rahmen der Definition, Prozessierung und Standardisierung von Metadaten mitzuwirken, wie es im internationalen Kontext beispielsweise innerhalb der Research Data Alliance (RDA) geschieht, wird gleichzeitig sichergestellt, dass die Arbeit der HMC Plattform nicht zu einer weiteren, wenn auch interdisziplinären Insellösung innerhalb Deutschlands führt, sondern, dass die Sichtbarkeit der Helmholtz-Gemeinschaft genutzt werden kann, um das Thema Metadatenmanagement international entscheidend voranzutreiben.

In die Helmholtz-Gemeinschaft hinein wird HMC in dreifacher Hinsicht wirksam werden.

- Durch die Verankerung von Metadata Hubs in den Forschungsbereichen der Gemeinschaft wird sichergestellt, dass die HMC Plattform die Bedarfe der Forschenden bedient und den internationalen Rahmenbedingungen der jeweiligen Fachdomänen Rechnung getragen wird.
- Gleichzeitig sorgen die Metadata Hubs in den Forschungsbereichen dafür, dass allgemeine Standards, Best Practices, Prozesse und Werkzeuge auf die Bedürfnisse der Forschenden zugeschnitten werden und sie in ihrem Bemühen, Forschungsdaten professionell zu organisieren, optimal unterstützt werden. Die Metadata Hubs nutzen dabei bereits existierende Strukturen innerhalb der Forschungsbereiche zur Vernetzung.
- Die Verbindung der einzelnen Metadata Hubs in einer Helmholtz Metadata Collaboration Plattform wiederum sorgt für die notwendige Vernetzung und Bündelung der Aktivitäten und erlaubt es, gemeinsame Werkzeuge weiter zu entwickeln und zur Verfügung zu stellen, von denen das Forschungsdatenmanagement in allen Forschungsbereichen profitiert. So erhöht sich langfristig die Arbeitsteilung und Zusammenarbeit über die Forschungsbereiche hinweg.

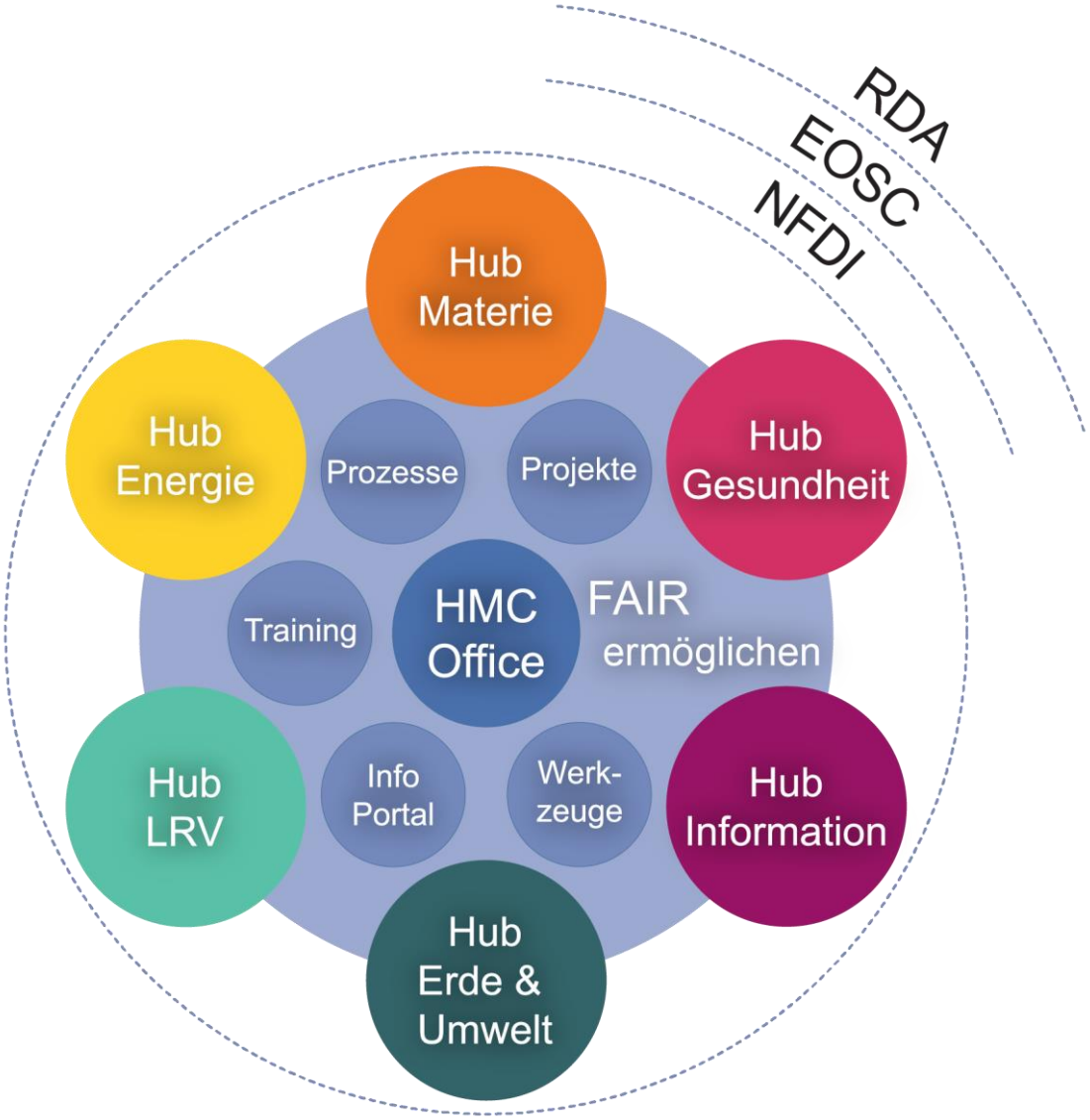

Abbildung 1: Struktur der Helmholtz Metadata Collaboration Plattform

HMC ist damit auch ein wichtiges Bindeglied in der Reihe der Plattformen des Helmholtz-Inkubators für Information & Data Science. Zum einen profitieren die innerhalb von HMC zu entwickelnden Services von den Diensten, welche durch die Technologieplattform HIFIS bereitgestellt werden. Damit wird für die Aktivitäten der HMC Plattform die technologische Basis gestellt. Zum anderen kann die Vermittlung von Fachkenntnissen und Expertise zum Thema Metadaten innerhalb der Akademie HIDA durchgeführt werden. Dies ermöglicht HMC, sich auf beratende, konkrete Projekte an den Zentren und unterstützende Aktivitäten im Rahmen von Consulting und Support zu fokussieren.

Darüber hinaus sind die beiden wissenschaftlich ausgerichteten Plattformen HIP und HAICU in hohem Maße darauf angewiesen, dass Forschungsdaten aus der Gemeinschaft den FAIR-Prinzipien folgend zugänglich gemacht werden. Insbesondere Big Data Analytics, wie sie im Rahmen von HAICU angestrebt werden, setzen die automatisierte Verfügbarkeit von Daten und deren beschreibende Daten

(=Metadaten) voraus. HMC schließt damit die Lücke zwischen Technologieplattformen und Wissenschaftsplattformen.

Im Ausbau der HMC Plattform geht es auch darum, einen Kulturwandel in der Zusammenarbeit und Vernetzung von Kompetenzen im Bereich von Forschungsdateninfrastrukturen zu erreichen, um kompatibel mit der Umsetzung einer NFDI zu sein und auch für die Helmholtz-Forschung optimal nutzen zu können. Die als dezentrale Komponenten über das HMC Office zur Verfügung gestellten technologischen Werkzeuge und Dienste sorgen dafür, die Umsetzungsschwelle in den Hubs und den Zentren niedrig zu halten. Werkzeuge zur Annotation von Forschungsdaten, zum Mapping von Datenschemata oder zur Sicherstellung der Provenienz von Workflows sind allgemeiner Natur und können, einmal entwickelt, an allen Zentren genutzt und auch Partnern übergeben werden. PID/DOI-Dienste und geeignete Repositorien machen einzelne Datenrepositorien und Forschungsdatensätze auffindbar und zuordenbar. Sie erlauben den Nutzerinnen und Nutzern schnell den Mehrwert einer Datenorganisation nachzuvollziehen. Bereits existierende Lösungen aus den Forschungsbereichen können aufgegriffen und an die Bedürfnisse und Anforderungen anderer Forschungsbereiche angepasst werden. HMC trägt dazu bei, die Anschlussfähigkeit zum Beispiel im Bereich des Ausbaus einer EOSC durch die Zuordnung von Daten (Forschungsdaten-Provenienz) herzustellen.

Leitlinien

Organisation und Aufbau der HMC Plattform orientieren sich dabei an fünf Leitlinien, die auch in der übergreifenden Digitalisierungsstrategie der Helmholtz Gemeinschaft abgebildet sind wie in den Zielen für den Aufbau der Nationalen Forschungsdaten Infrastruktur (NFDI).

Einbindung der Zielgruppen und Akteure

Die Arbeitsweisen aller Nutzer- und Anbietergruppen, von Datenerzeugern, Datennutzern, Infrastrukturanbietern bis zum Fördergeber, können durch die Einführung der Plattform beeinflusst werden. Nur eine dauerhafte, strukturelle Einbindung der Akteure in die Entwicklung verspricht eine hinreichend leistungsfähige, den komplexen und sich wandelnden Anforderungen der Forschung gemäße und nachhaltige Lösung (Wedlich u. a. 2017). Dazu gehört auch die Vernetzung der Akteure: hier begreift sich die HMC Plattform als Schnittstelle zwischen Wissenschaft und Forschungsdatenmanagement und nimmt diese mit Angeboten wie Schulungen, Workshops und Tagungen wahr.

Einfachheit, Offenheit und Transparenz

Durch effizientes Forschungs- und Metadatenmanagement wird die wissenschaftliche Arbeit letztlich vereinfacht und führt zu neuen Ergebnissen. Ein wesentlicher Baustein dafür sind einfach zu bedienende Werkzeuge. Ergebnisoffene und vorurteilsfreie Analyse des Stands der Technik, der Entwicklung von Diensten und Dienstleistungen sowie Transparenz bei langfristiger Bereitstellung von Werkzeugen und Diensten erhöhen das Vertrauen in HMC. Dabei enthalten die Metadaten auch Informationen, die die Prozesskette des Forschens und die Quelle der Daten offenlegt, um Qualitätsansprüche zu sichern und Nachnutzung zu erlauben.

Internationalität

Mit dem Anspruch der Helmholtz-Gemeinschaft, Anknüpfungspunkt für nationale und internationale Partner auf dem Gebiet der Meta- und Forschungsdaten zu sein, ist HMC mit internationalen Organisationen, z. B. der Research Data Alliance, CODATA und großen Forschungsdateninfrastrukturprojekten, wie EOSC und EUDAT CDI, eng vernetzt. Zusätzlich wird HMC durch die Beteiligung der Forschungsbereiche an internationalen Initiativen wie ELIXIR, AARC, ExPaNDS, etc. über die Metadata Hubs bereichert. Nur auf dieser Ebene etablieren sich im internationalen Kontext die besten Lösungen. Daher

muss recherchiert und geprüft werden, welche Elemente sich in diesem Umfeld nachnutzen und in internationale Gremien einbringen lassen. Auch aus Forscherperspektive ist der konkrete Austausch der Metadatenkompetenz auf internationaler Ebene entscheidend für bessere Ergebnisse, Sichtbarkeit und Reputation.

Interoperabilität

Zur Austauschbarkeit von Komponenten und Technologien sowie zur vernetzten Analyse der Metadaten ist Interoperabilität unerlässlich – und zugleich eine der größten Herausforderungen. Für ein zukunftsfähiges Konzept müssen Prozesse zur Erhebung, zum Management und zur Analyse von Metadaten definiert werden, die sich die Interoperabilität der verwendeten Methoden und technischen Komponenten zur Bedingung machen. Unter diesem Gesichtspunkt werden auch vorhandene Forschungsdateninfrastrukturen geprüft und ergänzt. Nur wenn die interdisziplinäre Austauschbarkeit auf internationalem Niveau gewährleistet ist, werden sich Standards und Technologien durchsetzen und von den Communities akzeptiert.

Nachhaltigkeit

Durch Integration des Umgangs mit Metadaten in die tägliche Arbeit von Forschenden wird eine nachhaltige Bereitstellung unabdingbar. Je mehr Ergebnisse anderer vorliegen, desto niedriger wird der Aufwand und umso höher der Nutzen für alle Beteiligten sein. Nur den Infrastrukturen, deren Betreuung auf lange Sicht gesichert ist, vertrauen wissenschaftliche Akteure ihre Daten an. Langfristige Verfügbarkeit und modulare technische Dienste ermöglichen auch kleineren Institutionen, Dienste bereitzustellen. Es muss von Beginn an vermittelt werden, dass die Infrastrukturen und Werkzeuge funktionstüchtig bleiben, gepflegt und erweitert werden.

2 Ziele

Eine umfassende Beschreibung der Rahmenbedingungen bei der Entstehung von Beobachtungs-, Mess- oder auch Analysedaten stellt die Forschenden vor ebenso große Herausforderungen, wie die Aufgaben im Rahmen des Managements von Metadaten oder der Umgang mit unbekannten Metadatenstandards im Rahmen von communityübergreifender Forschung.

Mit der Umsetzung der HMC Plattform werden in der Helmholtz-Gemeinschaft die methodisch-technischen Voraussetzungen sowie die strukturellen Angebote geschaffen, um

- richtungsweisend für die wissenschaftlichen Disziplinen und Forschungsinfrastrukturen die qualitative Beschreibung von Forschungsdaten durch Metadaten zu etablieren und damit auch eine communityübergreifende und internationale Vernetzung voranzutreiben,
- die Forschungsdaten so aufzubereiten, dass sie gefunden werden (Find), zugreifbar (Access) und (I)nteroperabel sind sowie disziplinär und interdisziplinär wiederverwendet (Re-use) werden können (FAIR-Prinzipien) (Wilkinson u. a. 2016), z. B. um eine automatisierte Erschließung und Zuordnung großer Datensammlungen zu ermöglichen (einschließlich von Ontologien und Provenienz),
- eine neue Forschungskultur zu fördern, in der die Forschenden aus der Helmholtz-Gemeinschaft stolz sind, ihre Daten öffentlich zugänglich zu machen. Die hohe Qualität der oft zitierten und nachgenutzten, auffindbaren Forschungsdaten erhöht ihre internationale Sichtbarkeit und wissenschaftliche Reputation und bildet so eine wichtige Grundlage für Open Science.

Für das Erreichen dieser Ziele sind verschiedene Handlungsstränge so zu kombinieren, dass die unterschiedlichen Zielgruppen (Forschende, Wissenschafts-Communities, Zentren, Helmholtz-Gemeinschaft, nationale und internationale Partner) Informationen, Werkzeuge und Dienste in leicht nutz- und verwertbarer Form und unter Zuhilfenahme von entsprechenden Beratungs- und Schulungsangeboten weiterentwickeln und in ihre Arbeitsabläufe integrieren können. Die HMC Plattform hat daher insgesamt eine Struktur von zentralen und dezentralen, statischen und dynamischen Leistungen gewählt, in der die Forschungsbereiche Kompetenzen ihrer Fachdomänen über lokale Hubs organisieren, und zentrale Aktivitäten durch ein HMC Office mit dezentralen Leistungen sowie dynamisch ausgeschriebenen Projekten für die Weiterentwicklung der Gemeinschaft als Ganzes abgebildet werden (Abbildung 1). Im Folgenden werden die wesentlichen Ziele und Aufgaben auf Basis des zu erwartenden Mehrwerts dargestellt.

Erschließung der Forschungsdaten durch Metadaten

In den verschiedenen wissenschaftlichen Communities werden Metadaten in unterschiedlicher Breite und Struktur erfasst. Eine einheitliche Datenstruktur für die Metadaten gibt es dabei selten. Als Ergebnis sind in der Vergangenheit unzählige Metadatenstandards und -ontologien entstanden, die nicht immer miteinander kompatibel sind. Allein in der marinen Community sind mehr als 150 bekannt. Alle Forschenden stehen demnach vor der Aufgabe, den für ihre Forschungsaufgabe passenden Metadatenstandard zu finden und anzuwenden, um eine Vergleichbarkeit innerhalb der Community und Anwendbarkeit für ähnliche Forschungsaufgaben sicherzustellen. Diese entwickeln sich jedoch dynamisch wie das Forschungsfeld selbst weiter. Dafür ist in der HMC Struktur eine Zusammenarbeit durch Verankerung dauerhafter Leistungen bei den Forschungsbereichen in Form der Metadata Hubs (Abbildung 1) vorgesehen.

Ein weiterer Aspekt ist die Erfassung der Metadaten. Alle Forschenden sowie die Betreiber von Großforschungsanlagen stehen vor der Frage, wie viele und welche Metadaten sie erfassen müssen. Für die Forschenden selbst ist bei der Verwendung ihrer eigenen Forschungsdaten häufig nur ein Bruchteil der Metadaten notwendig, da sie selbst über die übrigen Informationen als Teil der Hypothese beziehungsweise der Planung und Entwicklung des Experiments bereits verfügen. Dieses implizite Wissen steht Nachnutzenden nicht zur Verfügung. Mit der Explizierung dieses „zusätzlichen“ Wissens, in Form strukturierter Metadaten ist jedoch zusätzlicher Aufwand verbunden, der für die Forschenden selbst keinen unmittelbaren Mehrwert generiert (sieht man einmal von der noch zu etablierenden Anerkennung für qualitativ hochwertige Forschungsdaten inklusive detaillierter Metadaten ab). Jedoch ergibt sich dieser Mehrwert oft unmittelbar, wenn eine spätere Nachvollziehbarkeit der Datennahme während der Erstanalyse der Daten notwendig wird. Um das Verhältnis zwischen Aufwand und Nutzen auch für den einzelnen Forschenden selbst zu verbessern, sind Werkzeuge und Abläufe (Workflows) notwendig, die eine automatisierte Erfassung von Metadaten bestmöglich unterstützen oder die manuelle Erfassung von Metadaten erleichtern. Auch diese Leistungen sind zunächst von den Domänen zu erbringen und weiter zu entwickeln. Betreibern von Großforschungsanlagen kommt dabei eine besondere Rolle zu, da sie selbst Metadaten communityübergreifend aufnehmen, sammeln und zur Verfügung stellen. Hier zeigen sich zwischen den Forschungsbereichen sehr unterschiedlich ausgereifte Kompetenzen und Leistungen, die übergreifend (zentral) kartiert werden müssen, um die Effizienz und Leistungsfähigkeit der Helmholtz-Gemeinschaft in einer Arbeitsteilung zu erhöhen. Zur Zielerreichung wird das HMC Office mit statischen Leistungen eingerichtet, die nach den unterschiedlichen Kompetenzen der Zentren auch dezentrale Komponenten haben (Abbildung 1).

Um die *Forschenden* bei der Nutzung und Generierung von Metadaten durch Angebote zu unterstützen, wird HMC mit Hilfe der Metadata Hubs der Forschungsbereiche und in enger Anbindung an dort bereits bestehende Strukturen:

1. Eine umfassende Informationsbasis bereitstellen, die sowohl den aktuellen Stand hinsichtlich der in der Wissenschaft verwendeten Metadaten-Ontologien sowie Metadatenstandards enthält, als auch einen zentralen Zugang zu diesen Informationen ermöglicht.
2. Werkzeuge und Vorgehensweisen bereitstellen sowie bei deren Implementierung beraten, um Metadaten möglichst automatisiert zu erfassen und damit die Forschenden von Mehraufwand zu befreien.
3. Forscherinnen und Forschern die Vorteile der durch Orientierung an den FAIR Prinzipien gewonnenen Zitierfähigkeit und Nachnutzbarkeit ihrer Daten praktisch demonstrieren.

In manchen Communities ist hinsichtlich der Vielfalt der verschiedenen Metadaten-Ontologien und -standards seit geraumer Zeit der Versuch unternommen worden, die Anzahl solcher Standards zu begrenzen und sich auf möglichst wenige zu einigen. Metadatenstandards sind also auch selbst zum Gegenstand der Untersuchung, Überprüfung und gegebenenfalls Neuordnung geworden. In den Communities hat sich dafür Expertise herausgebildet, die für eine Bearbeitung von communityübergreifenden Fragen unbedingt zu beteiligen ist. Gleichzeitig kann für eine Herangehensweise zur Ausprägung von Quasi-Standards in den verschiedenen Wissenschaftsdisziplinen Methoden-Know-How genutzt werden, das zentral aufgebaut und vorgehalten wird. Für die *Wissenschafts-Communities* wird HMC

4. vor allem über die Metadata Hubs der Forschungsbereiche Methodenwissen zentral aufbauen und in Form von Beratung, Schulung, etc. bereitstellen, um z. B. in den Communities die Optimierung der Metadatenstandards zu unterstützen,
5. in die übergreifenden Fragestellungen zu Metadaten die Expertinnen und Experten aus den Communities einbeziehen und so Lösungen erarbeiten, mit denen communityübergreifende wissenschaftliche Fragestellungen beantwortet werden können,
6. Expertinnen und Experten benennen, die die Fachcommunities in nationalen und internationalen Prozessen der Generierung und Weiterentwicklung von Metadatenwerkzeugen, -diensten und -standards vertreten können.

Standards werden auch durch die Betreiber der Großforschungsanlagen in den einzelnen Forschungsbereichen in Zusammenarbeit mit den Nutzercommunities definiert und etabliert. Daher wird ein enger Kontakt zwischen HMC und den Betreibern der Großforschungsanlagen angestrebt. Dies geschieht vor allem über die Metadata Hubs der Forschungsbereiche sowie bereits in den Forschungsbereichen existierende Vernetzungsstrukturen, um so eine umfassende Koordination aller Aktivitäten in Bezug auf Metadaten zu fördern.

Forschungsdaten durch Metadaten verwertbar machen

Die Helmholtz-Gemeinschaft verfügt aus ihrer langjährigen Forschung an zentralen, gesellschaftlich bedeutenden Fragestellungen über einen großen Fundus unterschiedlicher Forschungsdaten. Der Wert einer solchen Datensammlung erschließt sich nur dann, wenn die Daten insgesamt auch auswertbar beziehungsweise analysierbar sind. Die Plattformen des Helmholtz-Inkubators widmen sich jeweils verschiedenen Schwerpunkten, so setzen HIP und HAICU auf Aspekte der Analyse von Daten im Bildbereich beziehungsweise durch Künstliche Intelligenz, HIFIS schafft technische Voraussetzungen für eine breitbandige Vernetzung der Helmholtz-Zentren und damit der Speicherorte von Forschungsdaten der Helmholtz-Gemeinschaft und stellt eine Plattform für Cloud-Dienste bereit.

In Bezug auf die Vernetzung der vorhandenen und künftigen Kompetenzen und Arbeitsteilung im Management von Forschungsdaten und Metadaten fehlt die Ebene, in der die Expertisen der verschiedenen Zentren so intelligent miteinander verknüpft werden, dass wesentliche Ziele für die gesamte Gemeinschaft und ihre Partner sowie für internationalen Aktivitäten erreicht werden. Für Forschende wird HMC dazu:

7. Dienste und Werkzeuge zur dezentralen Organisation und Verwaltung von Metadaten entwickeln und in enger Verbindung mit HIFIS bereitstellen und betreiben;
8. Prozesse, Werkzeuge, Dienste und Schnittstellen für die Erschließung, Verknüpfung und intelligente Aggregation der vorhandenen und künftigen Forschungsdaten aufbauen, besonders im Bereich der Datenprovenienz und der Ontologien;
9. über die dynamische Komponente der Projekte die methodisch-technischen Kompetenzen communityübergreifend aufbauen;
10. Schulungen, Training und Beratung zum Thema FAIR Data und Metadaten unter Einbeziehung nationaler und internationaler Expertinnen und Experten anbieten.

Anerkennung durch qualitativ hochwertige Metadaten für Forschungsdaten

Die steigende Bedeutung von Forschungsdaten geht Hand in Hand mit der wachsenden Bedeutung von Metadaten. In der Vergangenheit war der wesentliche Output der Forschenden die Publikation der wissenschaftlichen Ergebnisse in Textform. Forschungsdaten und genutzte Software bei der Erfassung und Analyse der Daten waren nachrangig in der Bewertung der wissenschaftlichen Leistung. Inzwischen ist der potenzielle Wert von Forschungsdaten allgemein anerkannt. Deshalb fordern die Forschungsförderer die Bereitstellung der Daten nicht nur für die Nachvollziehbarkeit von Forschungsergebnissen, sondern auch für eine Nachnutzung. Ausdruck der wachsenden Bedeutung der Daten ist außerdem die Entstehung und der große Erfolg von Datenjournalen, die qualitätsgeprüfte (peer review) Beschreibungen von Mess- und Beobachtungsdaten veröffentlichen. In dieser Form sind die Daten zugreifbar (siehe unter anderem „Earth Systems Science Data“ oder „Nature Scientific Data“). Wissenschaftliche Reputation durch die Nachnutzung von Forschungsdaten beruht ganz wesentlich darauf, dass die Qualität und Vollständigkeit der kontextuellen (inhaltsbezogenen) Metadaten eine Nachnutzung der Daten auch im interdisziplinären Umfeld erlauben. So können diese als Grundlage neuer Erkenntnisse dienen.

Die DFG hat in den „Empfehlungen zur gesicherten Aufbewahrung und Bereitstellung digitaler Forschungsprimärdaten“ bereits im Januar 2009 die Bedeutung der Metadaten als umfassende Beschreibung der Primärdaten aus der Forschung aufgezeigt und dabei die Einbeziehung community-spezifischer Voraussetzungen und Strukturen empfohlen. Daneben wird dort auch auf die internationale Zusammenarbeit schon eingegangen:

Zu den wesentlichen Zielen gehört, dass sich diese Strukturen an internationalen Maßstäben und Standards orientieren und in bereits vorhandene überregionale, internationale Strukturen einbetten. Für ihr nachhaltiges Wirken ist von Anfang an Sorge zu tragen (Deutsche Forschungsgemeinschaft 2009). Auch im Aufbau der NFDI spielt dieser Prozess eine tragende Rolle.

Um diesen Zielstellungen gerecht zu werden, wird HMC organisatorisch und koordinierend sowohl in die Helmholtz-Gemeinschaft wirken, und dabei die Expertise und das Know-How der Institute und Communities einbeziehen, als auch im internationalen Maßstab sichtbar agieren. Im Einzelnen werden für die Helmholtz-Gemeinschaft folgende Ziele verfolgt:

11. Für die Verbreitung von Informationen, Kenntnissen und Vorgehensweisen im Zusammenhang mit Metadaten werden Basisinformationen und Schulungsinhalte bereitgestellt, die in Zusammenarbeit mit den Community-Expertinnen und -Experten auf die jeweiligen Wissenschaftsgebiete angepasst werden und z. B. im Rahmen der HIDA als fachspezifischer Schulungsinhalte angeboten werden können.
12. Für die Einbeziehung der vorhandenen Expertise in den Forschungsbereichen ist der Aufbau und die nachhaltige Koordinierung einer eigenen „Metadaten-Community“ vorgesehen. Die Vernetzung der verschiedenen Akteure ist sowohl für die Bewertung und Bearbeitung genereller beziehungsweise übergreifender Fragestellungen notwendig, gleichzeitig sind die Mitglieder dieser „Metadaten-Community“ als Ansprechpartner eine direkte Verbindung in die Forschungsbereiche.
13. International wird die Plattform HMC community-übergreifende Metadatenmethoden und technische Dienste entwickeln und aktiv mit der Arbeit in internationalen Gremien (z. B. der Research Data Alliance) und internationalen Projekten (z. B. EOSC) vernetzen. So entstehen zukunftssichere Lösungen, die für die Zentren und Forschungsbereiche sowie den Bau und Betrieb komplexer Forschungsanlagen dringend notwendig sind. Die durch die Zentren in ihren jeweiligen internationalen Fachgemeinschaften geleistete Vernetzungsarbeit wird durch übergreifende Aktivitäten ergänzt.

Optimale Vernetzung und nationale/internationale Anschlussfähigkeit

Der Rat für Informationsinfrastrukturen (RfII – Rat für Informationsinfrastrukturen 2017) wie auch viele Forschungsförderer weltweit fordern einen offenen und harmonisierten Umgang mit wertvollen Forschungsdaten, um diese langfristig einem breiteren Publikum zur Nachnutzung zugänglich zu machen (Wedlich u. a. 2017). In Deutschland ist geplant, ab 2019 eine wissenschaftsgetriebene NFDI aufzubauen. Darüber hinaus verpflichtet sich die Helmholtz-Gemeinschaft zu den Prozessen von Open Science, insbesondere der EOSC der Europäischen Kommission. HMC ist bereits inhaltlich auf die zu erwartenden Empfehlungen des NFDI und EOSC ausgerichtet. So erarbeitet sich die Helmholtz-Gemeinschaft mit der Einrichtung der Plattform, in der die bestehende Expertise erstmals gebündelt und zu konkreten Lösungen geführt wird, eine Vorreiterrolle — bevor avisierte Projekte in EOSC und NFDI starten und die Rekrutierung ihrer Mitarbeiter beginnen. Gleichzeitig werden die Helmholtz-Zentren durch HMC attraktive Konsortialpartner im Kontext von EOSC und NFDI.

Die Plattform schafft eine Balance zwischen den Bedürfnissen der Forschungsdisziplinen und Infrastrukturanbieter wie auch den politischen Forderungen nach Nachhaltigkeit und bietet die Voraussetzungen für den offenen Zugang und Austausch von Forschungsdaten über Metadaten. Dazu wird HMC eng mit den Akteuren der H2020 FET Flagships Human Brain Project und Battery 2030+, der Erdsystemforschung, beispielsweise der Deutschen Allianz für Meeresforschung, und den Großgerätebetreibern im Forschungsbereich Materie sowie dem dort verorteten Topic „Data Management & Analysis“ zusammenarbeiten, um so die in solchen kollaborativen interdisziplinären Großprojekten erworbenen Kompetenzen zu nutzen und den Austausch auch über Forschungsfelder hinweg zu ermöglichen. HMC bietet besonders durch Organisation zentraler Prozesse einschließlich ihrer Kommunikation und Vertretung in Gremien konkrete Vorteile für die verschiedenen Akteure:

1. *Datenerzeuger* finden Empfehlungen und einheitliche Werkzeuge, um passende Metadaten-Standards auszuwählen, zu benutzen und um die gesammelten Meta- und Forschungsdaten für Menschen und Maschinen zugänglich, durchsuchbar und archivierbar zu machen. Das erleichtert ihre Arbeit und ermöglicht die Zitierung und Wiederverwendung der erzeugten Daten.

2. *Datennutzer* werden über die Weiterentwicklung nationaler und internationaler Forschungsdateninfrastrukturen und Informationssysteme informiert und bestmöglich nach Innen und Außen vertreten. Sie können eine Reihe von Diensten wahrnehmen und Kompetenzen erwerben sowie ihre Bedarfe anmelden.
3. *Institutionen* können zentral angebotene Informationen im Bereich Best Practice, Schulungen und Workshops für die Weiterentwicklung ihrer eigenen Kulturen im Forschungsdatenmanagement nutzen und adaptieren. Zudem erhalten sie einheitliche Bewertungskriterien beziehungsweise Metriken für die Qualität des Bereiches Metadaten in ihren Strukturen
4. Die *Helmholtz-Gemeinschaft* profitiert durch den disziplinübergreifenden Erfahrungsaustausch mit gemeinschaftsweiten Lösungen, durch die kostenintensive Parallelentwicklungen in den Forschungsbereichen und Programmen vermieden werden. Ihre Leistungen und die Bereitstellung ihrer Daten werden auffindbar und besser sichtbar, z. B. durch Datenprovenienz für Open Science.

3 Operationalisierung der Ziele

Um die beschriebenen Ziele zu erreichen und damit den größtmöglichen Nutzen für die Forschenden, die wissenschaftlichen Communities, die Helmholtz-Zentren und -Gemeinschaft aus der Plattform zu ziehen, müssen zumeist parallel organisatorische und technische Aufgaben erledigt werden. Gleiches gilt für eine Unterscheidung zwischen domänenspezifischen und generellen Aufgaben. Eine Zusammenstellung relevanter Aufgaben und ihrer Verankerung in den HMC Strukturen findet sich in Tabelle 1.

Alle Ziele, deren Erfüllung vor allem domänenspezifische Expertise erfordern, werden in operativen Einheiten – den Metadata Hubs – entlang der Forschungsbereiche gebündelt, um den fachspezifischen Charakteristiken der jeweiligen Anwendungsdomäne zu entsprechen. Hierzu gehört vor allem die Erarbeitung des methodischen Konzepts mit den notwendigen Festlegungen von Prozessen, Empfehlungen und den Best Practices zum Umgang mit Metadaten und Standards in den jeweiligen Domänen. Hinzu kommt die Metadatenexpertise aus den Domänen, die in Form von Informationen und Schulungsmaterialien gesammelt, aufbereitet und bereitgestellt werden, so dass alle für HMC notwendigen Wissensquellen über einen Ort erreichbar sind. Auch die Datensammlungen, die durch Portale erschlossen werden sollen, müssen in einer für die Forschenden jeweils fachlich sinnvollen Form zugänglich sein. Dabei haben die Forschungsbereiche ganz unterschiedliche Kompetenzen, Strukturen, Werkzeuge und Methoden, die insgesamt kartiert werden müssen, um Effizienz, Arbeitsteilung, Sichtbarkeit und Mehrwert zu fördern. Hier entstehen Schnittstellen-Aufgaben mit den zentralen Komponenten (HMC Office und dezentrale Leistungen: FAIR technisch ermöglichen).

Zu den operativen Einheiten, die sich der Erreichung der übergeordneten und damit generellen Ziele widmen, zählen vor allem die Entwicklung und der nachhaltige Betrieb übergreifender informationstechnischer Werkzeuge und Dienste: Die Portaltechnologie des Informationsportals, das den zentralen Zugang zu HMC für die Nutzenden darstellt (HMC Office) sowie der übergreifende Werkzeugkasten für die Bereitstellung, Anreicherung, Beschreibung und Nutzung der Metadaten (dezentrale Leistungen des Office: FAIR technisch ermöglichen). Ferner ist die Koordination und Steuerung der dezentralen Leistungen des HMC Office wie auch der regelmäßig auszuschreibenden dynamischen HMC-Projekte zentral unter Einbeziehung der Metadata Hubs zu gewährleisten, um sicherzustellen, dass die zu entwickelnden Produkte dieser Projekte in einen größeren Kontext eingeordnet werden können und einen Mehrwert für die gesamte Gemeinschaft erzeugen. Tabelle 1 zeigt die Arbeitspakete der einzelnen Bereiche und deren Verknüpfungen beziehungsweise Abhängigkeiten.

Tabelle 1: Übersicht der Arbeitspakete der drei HMC-Teilbereiche, deren Mittelbedarf und Verknüpfung. Eine detaillierte Aufstellung der Finanzen findet sich in Tabelle 3.

AP	Zielsetzung	
1	Zentrales HMC Office, zentrale dauerhafte Leistungen	4 FTE + 70 T€
1.1	Aufbau und Koordination der Geschäftsstelle	1.5 FTE
1.1.1	Aufstellen von und Abstimmen mit wissenschaftlichem Beirat	AP 3.1.1
1.1.2	Kommunikation und PR	AP 2-3
1.1.3	Zusammenarbeit mit anderen Aktivitäten des Helmholtz-Inkubators	AP 2-3
1.1.4	Internationale Vernetzung, Gremienarbeit, Harmonisierung und Standardisierung	AP 3.1.3
1.1.5	Aufbau und Betreuung einer „Metadaten-Community“ in der Helmholtz-Gemeinschaft	AP 3.1.2
1.1.6	Projektbüro – Ausschreibung, Überwachung der IVF Ausschreibung und Qualitätssicherung	AP 4, AP 3.1, AP 2
1.2	Wissen & Vermittlung	2 FTE
1.2.1	Kartierung der Forschungsdatenexpertisen	AP 3.2.2
1.2.2	Methodenwissen aufbauen und verfügbar machen	AP 3.2.1, AP 3.2.3
1.2.3	Basisinformationen und Schulungsinhalte – Wissen bezüglich Metadaten verbreiten	AP 3.2.1-3
1.2.4	Vermittlung durch Expertinnen und Experten	AP 3.3.4
1.2.5	Persönliche Beratung	AP 3.3.4
1.3	Komponenten & Prozesse	0.5 FTE
1.3.1	(Meta-)Daten haben eine eindeutige Lizenz	AP 3.1
1.3.2	FAIRe Vokabularien, Ontologien, Minimalstandards	AP 3.3.5-6, AP 2.1.2, AP 2.2.1
1.3.3	Empfehlungen zur Annotierung von Forschungsdaten mit umfangreichen Metadaten, unter anderem Provenienz	AP 3.1, AP 2
2	Dezentrale Komponenten des HMC Office: FAIR technisch ermöglichen	8 FTE
2.1	Übergreifende technische Dienste und Werkzeuge	4 FTE (anfangs 6)
2.1.1	Implementierung FAIR Data Object – Schnittstellen zu EOSC und NFDI	AP 1.2, AP 3.1.3, AP 4
2.1.2	Entwicklung und Betrieb technischer Dienste und Werkzeuge	AP 1.2, AP 3.1.1, AP 4
2.2	Komponenten und Prozesse – Realisierung der FAIR-Prinzipien	4 FTE (anfangs 2)
2.2.1	Findable: Forschungsdaten mittels Identifier und umfassenden Metadaten auffindbar machen	AP 3.3, AP 1.2
2.2.2	Accessible: offene und standardisierte Protokolle und Policies	AP 3.3.4, AP 3.1
2.2.3	Interoperable: FAIRe Wissensrepräsentation durch offene Vokabulare, Ontologien und Standards	AP 1.3.2, AP 3.2.1, AP 3.3
2.2.4	Re-Usable: Meta- / daten, ihre Provenienz und Nutzungslizenz sind detailliert beschrieben	AP 1.3.5, AP 3.3.6
3	Domänenspezifische Leistung – Metadaten Hubs	je 5 FTE + 50 T€
3.1	Koordination und Management	1 FTE pro Hub
3.1.1	Einbeziehung der Community-Expertise	AP 1.1.5, AP 1, AP 2.2
3.1.2	Aufbau und Betreuung einer „Metadaten-Community“ im jeweiligen Forschungsbereich	AP 1.1.5
3.1.3	Internationale Vernetzung, Gremienarbeit, Harmonisierung und Standardisierung	AP 1.1.4
3.2	Wissen & Vermittlung	2 FTE pro Hub
3.2.1	Aufbau und Bereitstellung einer Informationsbasis zu Metadaten(-vokabularien), Ontologien und Standards	AP 1.2.2-3
3.2.2	Landschaft der Forschungsdatenexpertisen	AP 1.1.5, AP 1.2.1, AP 1.2.4
3.2.3	Domänenspezifische Ergänzungen zu Methoden, Basisinformationen und Schulungsinhalten	AP 1.2.2-3
3.2.4	Beratung von Zentren, Projekten und Forschenden	AP 1.2.4-5
3.3	Komponenten & Prozesse	2 FTE pro Hub
3.3.1	Prozesse, Werkzeuge und Dienste zur Erschließung von Forschungsdatensammlungen aufbauen	AP 1.2.1-2, AP 2.2.1-4
3.3.2	Ingest: Werkzeuge zur Automatisierung der Erfassung von Metadaten	AP 1.3.3, AP 2.1.2, AP 2.2.1
3.3.3	Werkzeuge zur Handhabung verschiedener Metadatenstandards	AP 2.2.1-4
3.3.4	Zugang zu Forschungsdaten über Schnittstellen schaffen	AP 2.2.1-4
3.3.5	Erfassen der FAIRness der Daten	AP 1.3.4, AP 2
3.3.6	Provenienz	AP 1.3.3, AP 2.2
4	Inkubatorprojekte	1200 T€

4 Arbeitsprogramm

AP 1: Zentrales HMC Office

AP 1.1. Aufbau und Koordination der Geschäftsstelle

Die Geschäftsstelle ist das organisatorische und operative Zentrum der HMC-Plattform, es betreibt das Projektmanagement. Dort laufen alle Auftrags-, Priorisierungs- und Kommunikationsaufgaben zusammen, um eine moderierte und koordinierte Arbeitsweise der Plattform zu gewährleisten. Zudem sind die Berichtswege der Plattform gegenüber den zuständigen Gremien über die Geschäftsstelle organisiert. Der Aufbau beinhaltet demnach die Ausstattung des Office und die Etablierung der Kommunikationswege und Abläufe für ein reibungsloses Funktionieren der Plattform. Die Geschäftsstelle ist zudem für die Außendarstellung der Plattform verantwortlich, bereitet in Zusammenarbeit mit dem IVF die Projektausschreibungen vor und koordiniert diese.

Dieses Arbeitspaket erfordert eine enge Abstimmung mit AP 3.1.

AP 1.1.1 Aufstellen von und Abstimmen mit wissenschaftlichem Beirat

Das HMC Collaboration Board schlägt Kandidaten für den wissenschaftlichen Beirat vor, informiert den Beirat durch regelmäßige Berichte und organisiert die Beiratssitzungstermine.

Dieses Arbeitspaket erfordert eine enge Abstimmung mit AP 3.1.1.

AP 1.1.2 Kommunikation und PR

Die Kommunikationsstrategie, also wann, wie und wo Informationen verteilt werden, ist Teil dieses Arbeitspakets. Dazu wird unter anderem ein Logo für HMC erstellt, ein Corporate Design (in Absprache mit dem Helmholtz-Design), Flyer, Gadgets, Poster, etc. Die PR ist außerdem für die Verbreitung über Social Media verantwortlich sowie die Erstellung von kurzen Videos und Info-Texten für Presse und interne Publikationen. Generell ist dieses AP auch Ansprechpartner für die Expertinnen und Experten, die allgemeines Informationsmaterial zur Verteilung in die Programme und Forschungsbereiche benötigen. Für die zu erwartenden Materialkosten für Werbezwecke sind als Sachmittel in der Finanzierung vorgesehen.

Dieses Arbeitspaket erfordert eine enge Abstimmung mit AP 2 und 3.

AP 1.1.3 Zusammenarbeit mit anderen Aktivitäten des Helmholtz-Inkubators

Es ist ein enger Kontakt z. B. zu den koordinierenden Stellen der Plattformen HIDA, HIFIS, HIP und HAICU, zu pflegen. Insbesondere ist die automatisierte Abrufbarkeit („machine-to-machine communication“) von standardisierten und harmonisierten Daten und deren beschreibenden Daten (=Metadaten) eine Voraussetzung für Big Data Analytics unter anderem im Rahmen von HAICU. Die metadatenspezifischen technischen Dienste, die in HMC entwickelt werden, können als Services über HIFIS zur Verfügung gestellt werden. HMC stützt sich auf alle allgemeinen technischen Dienste (wie Authentifizierung, etc.) die im Rahmen von HIFIS entwickelt werden.

Dieses Arbeitspaket erfordert eine enge Abstimmung mit AP 2 und 3.

AP 1.1.4 Internationale Vernetzung, Gremienarbeit, Harmonisierung und Standardisierung

Soweit sinnvoll und möglich, werden zu den identifizierten Spezifikationen und Prozessen Arbeitsgruppen in Organisationen wie der RDA, CODATA, ISO, ICSU-WDS und W3C identifiziert oder gegründet, in denen HMC-Vertreter aktiv mitgestalten und die internationale Anschlussfähigkeit sicherstellen. Die Vernetzung und Verbreitung der Informationen ist Aufgabe des HMC Office. Entsprechend werden die Ergebnisse der Gruppen in HMC auf allen methodischen oder technischen Ebenen einfließen.

Dieses Arbeitspaket erfordert eine enge Abstimmung mit AP 3.1.3.

AP 1.1.5 Aufbau und Betreuung einer „Metadaten-Community" in der Helmholtz-Gemeinschaft

Dieses Arbeitspaket führt die übergreifenden und die lokalen Aktivitäten der Vermittlung zusammen. Langfristig ist es sinnvoll, die bereits im Inkubator-Prozess eingerichtete Arbeitsgruppe „Mehrwerte aus Forschungsdaten durch Metadaten" als Helmholtz-weites Netzwerk von Metadaten-Expertinnen und -Experten zu verfestigen, um Informationen, Lösungen und Fachwissen zum Metadatenmanagement auszutauschen. Regelmäßig stattfindende Workshops bringen diese im Laufe der Zeit wachsende Gruppe zusammen, um Synergieeffekte zwischen den Forschungsbereichen voranzutreiben und damit das HMC Office bei der herausfordernden Aufgabe zu unterstützen, den Informationsaustausch zu steuern. Aufgaben und Themen der Metadaten-Community sind:

- Auf Experten-Niveau werden die in den einzelnen Metadata Hubs erreichten Lösungen gemeinsam mit den technischen Angeboten evaluiert, fachlich diskutiert und sinnvoll zusammengeführt.
- Durch Aktivität der Expertinnen und Experten aus der Metadata-Community innerhalb internationaler Gremien wie EOSC, RDA, CODATA, GO FAIR etc. können Impulse in das HMC hineingegeben sowie umgekehrt für die Verbreitung von im HMC entwickelten Lösungen gesorgt werden.

Die Metadaten-Community vernetzt die Plattform zusätzlich zu den domänenspezifischen Metadata Hubs quer zu den Forschungsbereichen, schafft damit neue Perspektiven und eine Helmholtz-übergreifende Sichtbarkeit.

Langfristig ist darauf zu achten, immer wieder neue Köpfe für die Metadata-Community zu gewinnen und beispielsweise durch Gäste von außerhalb der Helmholtz-Gemeinschaft, vor allem international, zu erweitern.

Dieses Arbeitspaket erfordert eine enge Abstimmung mit AP 3.1.2.

AP 1.1.6 Projektbüro – Ausschreibung, Überwachung der IVF Ausschreibung und Qualitätssicherung

In diesem Arbeitspaket werden die zu klärenden Forschungsfragen und die Kriterien für die Vergabe von Forschungsaufträgen spezifiziert und Vorschläge für Kriterien bei der Vergabe von Projekten durch den Impuls- und Vernetzungsfonds erarbeitet. Das bedeutet unter anderem den Ausschreibungsprozess, von der Verbreitung des initialen Aufrufs für Bewerbungen, über die Annahme von Interessensbekundungen und transparente Klärung von Fragen, die Organisation der Gutachten des wissenschaftlichen Beirats zu begleiten. Bezüglich der dynamischen Mittel aus dem Impuls- und Vernetzungsfonds werden diese Aufgaben von der Helmholtz-Geschäftsstelle unterstützt und mit den förderrechtlichen Regularien abgestimmt.

Während der Durchführung koordiniert das HMC Office jedes vergebene Forschungsprojekt. Es verfolgt den Fortschritt, verfeinert die Spezifikation und hilft, den Verlauf an neue Erkenntnisse anzupassen. Dabei muss die Qualität der Ergebnisse geplant, gemessen und gesteuert werden. So wird sichergestellt, dass die gelieferten Ergebnisse für ihren disziplinübergreifenden Einsatz im HMC geeignet sind.

Dieses Arbeitspaket erfordert eine enge Abstimmung mit AP 2, AP 3.1 und AP 4.

AP 1.2. Wissen & Vermittlung

In diesem Arbeitspaket werden Expertise- und Wissenskartierungen aus den Metadata Hubs gebündelt und Trainings-, Vermittlungs- und Beratungsformate entwickelt und umgesetzt.

Dieses Arbeitspaket erfordert eine enge Abstimmung mit AP 3.2 und AP 3.3.

AP 1.2.1. Kartierung der Forschungsdatenexpertisen

Der jeweils im Forschungsbereich benannte Hub des HMC soll eine Übersicht über vorhandene Expertisen im Forschungsbereich erstellen. Diese sind Zentren-, Großforschungsanlagen- und domänenübergreifend zu erfassen. Die Informationen sollen an das HMC Office geliefert und transparent dargestellt werden. Dies kann beispielsweise im Informationsportal als strukturierte Information durchsuchbar hinterlegt sein und damit eingesehen werden. Dadurch bekommen Helmholtz-Forschende die Möglichkeit, sowohl innerhalb ihres Forschungsbereichs, aber auch darüber hinaus, vorhandene Expertise zu nutzen.

Dieses Arbeitspaket erfordert eine enge Abstimmung mit AP 3.2.2.

AP 1.2.2 Methodenwissen aufbauen und verfügbar machen

Hier sind sinnvolle Verbreitungswege zu beachten. Die Informationen zu Best Practices etc. werden gesammelt und für das Informationsportal zielgruppengerecht aufbereitet. Folgende Arten von Verbreitungswegen sind zuerst vorgesehen:

- Wissenschaftliche Untersuchungen in Form von Vorträgen und Veröffentlichungen im Kontext von Forschungsdatenmanagement und internationalen Foren wie RDA, aber auch im fachwissenschaftlichen Kontext (Zielgruppe: Metadaten-Expertinnen und -Experten und Community-Expertinnen und -Experten);
- Bereitstellung von professionellen Anleitungen für z. B. die Verwendung von Schnittstellen, Metadatenschemata oder -registries im eigenen Umfeld (Zielgruppe: Verantwortliche der Informationsinfrastruktur in den Forschungsbereichen, Forschungsdaten-Liaison-Officers);
- Tutorials zu Anwendungen und Werkzeugen, die z. B. die Suchfunktionen des Informationsportals erläutern (Zielgruppe: Forschende, die sich mit dem Thema Metadaten nur indirekt beschäftigen wollen).

Daneben ist die Sichtbarkeit der gesamten Plattform und der durch sie gewonnen Mehrwerte innerhalb der Helmholtz-Gemeinschaft sowie national und international zu erreichen. Dies wird zu einem kleineren Anteil über Öffentlichkeitsarbeit umgesetzt (lokal in den Metadata Hubs, übergreifend im HMC Office) und zu einem entscheidenden Anteil über Multiplikatoren (Metadaten-Community) und erfolgreiche Forschung durch effektiveren Umgang mit Forschungsdaten, die mit qualitativ hochwertigen Metadaten versehen sind (HMC-Projekte). Neben Informationen durch Flyer, Webseite, Branding, Merchandising, etc. ist die Präsenz der Mitarbeitenden in den Metadata Hubs innerhalb der unterschiedlichen Communities auf Konferenzen, in Projekten und im Consulting/Support entscheidend, um Schritt für Schritt flächendeckend die Arbeitsweise jedes einzelnen Forschenden in Richtung eines „Kulturwandels“ im Umgang mit Forschungsdaten zu beeinflussen.

Dieses Arbeitspaket erfordert eine enge Abstimmung mit AP 3.2.1 und AP 3.2.3.

AP 1.2.3 Basisinformationen und Schulungsinhalte – Wissen bezüglich Metadaten verbreiten

Das HMC Office etabliert und betreibt ein Informationsportal als zentrale Anlaufstelle. Es erstellt verständliche Anleitungen und Erklärungen zu Metadaten-Themen, z. B. Provenienz, administrativen Metadaten und strukturellen Metadaten. Eine regelmäßig aktualisierte und mit HIDA abgeglichene Liste von Lehr- und Trainingseinheiten, Schulungen, sowie Summer Schools zum Thema Forschungsdatenmanagement und FAIR Data, gibt Auskunft über aktuelle Veranstaltungen und Weiterbildungsmöglichkeiten.

Dieses Arbeitspaket erfordert eine enge Abstimmung mit AP 3.2.1 bis 3.2.3.

AP 1.2.4 Vermittlung durch Expertinnen und Experten

Die Zielgruppe sind hier Akteure aus dem Forschungsdatenmanagement, die als Multiplikatoren eine Schnittstelle zur Wissenschaft darstellen.

Die Ansprache der Expertinnen und Experten findet auf der Ebene eines beidseitigen Austauschs von Know-how statt und erfolgt in mehreren Schritten. Kenntnisse und Erfahrungen fließen in die Umsetzung der Plattform ein und umgekehrt besteht der Anspruch von HMC, alle Akteure in diesem Bereich auf den gleichen Wissensstand zu bringen. Die einzelnen Helmholtz-Mitarbeiterinnen und -Mitarbeiter müssen in ihrer Fachsprache, in ihrem fachlichen Kontext und ihrem Kenntnisstand entsprechend angesprochen und beraten werden. Daher spielt die Vermittlung durch Multiplikatoren eine wichtige Rolle innerhalb der Plattform.

Nachdem der Kenntnisstand der Expertinnen und Experten ermittelt und die Frage geklärt ist, ob zusätzliche Fachkräfte ausgebildet werden müssen, können diese in Zusammenarbeit mit HIDA geschult werden. Danach wird das gesammelte Know-how der Akteure sowohl durch die Informationsplattform, als auch in Schulungen und Beratungen weitergegeben. Als Vermittler zwischen Forschungsdatenmanagement und Wissenschaft sollen die Expertinnen und Experten das Wissen im Fachvokabular der einzelnen Disziplinen und auf die Bedürfnisse vor Ort angepasst transferieren.

Dieses Arbeitspaket erfordert eine enge Abstimmung mit AP 3.3.4.

AP 1.2.5 Individuelle Beratung

Individuelle Beratung und Problemlösung ist eine der besten Lehr- und Lernmethoden. Doch ist es aus Effizienzgründen schwierig, allen Forschenden der Helmholtz-Gemeinschaft einen Metadaten-Experten beziehungsweise eine Metadaten-Expertin zur Seite zu stellen. Consulting und Support muss so spezifisch wie nötig und so generell wie möglich geplant werden. Folgende Arten von Consulting und Support lassen sich unterscheiden:

- Projektbezogene Beratung sollte dezentral und disziplinspezifisch in den Forschungsbereichen angesiedelt und durchgeführt werden.
- Beratung von Multiplikatoren wie dem ortsansässigen RDM-Team, oder dem Forschungsdaten-Liaison-Officer (Hilfe zur Selbsthilfe) führt zu einer netzwerkartigen Verbreitung von Kenntnissen über Funktionalitäten von Werkzeugen und Diensten des HMC.
- Community-Expertinnen und -Experten werden zu Metadaten-Beratern und Beraterinnen auf ihrem domänenspezifischen Gebiet weitergebildet und sorgen damit für einen Domino- und Multiplikationseffekt.

Hier ergeben sich Schlüsselpositionen zwischen den eher technischen Lösungen des Metadatenmanagements und den domänenspezifischen Bedarfen. Es findet ein gegenseitiger Wissensaustausch statt, von dem beide Seiten, die Fachwissenschaft und das Metadatenmanagement, profitieren. Bedarfe und bereits entwickelte Lösungen werden an die dezentralen Leistungen des HMC Office weitergegeben und anwenderorientiertes Consulting und Support zu neuen, dort entwickelten Lösungen wird an die Forschungsbereiche zurückgegeben.

Dieses Arbeitspaket erfordert eine enge Abstimmung mit AP 3.3.4.

AP 1.3. Komponenten & Prozesse

Über dieses Arbeitspaket entwickelt das zentrale HMC Office Anforderungen und Empfehlungen z. B. zu Lizensierungsfragen, Vokabularien und Provenienz unter anderem auf Basis der Ergebnisse der dezentralen technischen Leistungen sowie der Metadata Hubs.

Dieses Arbeitspaket erfordert eine enge Abstimmung mit AP 2 und AP 3.3.

AP 1.3.1 (Meta-)Daten haben eine eindeutige Lizenz

Das HMC Office prüft bestehende Lizenzmodelle (wie Creative Commons) auf Eignung und erarbeitet auf deren Basis Richtlinien zur Nutzung im Kontext von Daten- und Metadaten, insbesondere eines „Helmholtz Defaults" für Lizenzen, so dass möglichst viele Wissenschaftlerinnen und Wissenschaftler davon profitieren. Eine einschränkende Lizenz für Metadaten sollte nur in Ausnahmefällen genutzt werden. Empfehlungen für Lizenzen zur (Sekundär-)Nutzung von Forschungsdaten sollten ebenfalls erarbeitet werden. Bereits bestehende Erfahrungen und Formulierungen, beispielsweise aus den Helmholtz AK Open Science und AK Recht werden integriert.

Dieses Arbeitspaket erfordert eine enge Abstimmung mit AP 3.1.

AP 1.3.2 FAIRe Vokabularien, Ontologien, Minimalstandards

Die Entwicklung von Strukturen für Metadaten, wie Vokabularien oder Ontologien, ist in vielen Fällen nur sinnvoll, wenn eine kritische Masse an Anwenderinnen und Anwendern daran mitarbeitet und die Ergebnisse übernimmt. Deshalb ist vor allem die Möglichkeit zur kooperativen Entwicklung solcher Strukturen zu schaffen, sowie das Bewusstsein für deren Wichtigkeit zu wecken.

In enger Absprache mit den Hubs werden die Anforderungen im Bereich Vokabularien, Ontologien und Minimalstandards erfasst. Nach Abgleich mit bereits bestehenden Initiativen (z. B. BioPortal und fairsharing) wird gegebenenfalls die Entwicklung neuer Ontologien und Standards angestoßen. Dabei erscheint ein modularer Aufbau sinnvoll, um gemeinsame „Kernelemente" vorgeben zu können, aber z. B. jedem Forschungsbereich oder jedem Zentrum Ergänzungen dazu zu ermöglichen.

Dieses Arbeitspaket erfordert eine enge Abstimmung mit AP 3.3.5, AP 3.3.6, AP 2.1.2 und AP 2.2.1.

AP 1.3.3. Empfehlungen zur Annotation von Forschungsdaten mit umfangreichen Metadaten, unter anderem Provenienz

In diesem Arbeitspaket werden Policies und Standardprozesse gesammelt, bewertet und optimiert in Form von Vorlagen an die Hubs weitergegeben. Damit wird die Arbeit der Annotation in den Metadata Hubs vorbereitet und koordiniert. Die Empfehlungen werden dann zusammengefasst und verbreitet.

Dieses Arbeitspaket erfordert eine enge Abstimmung mit AP 2 und AP 3.1.

AP 2: Dezentrale Leistungen HMC Office: FAIR technisch ermöglichen

Die technische Umsetzung der FAIR-Prinzipien bildet eine notwendige Basis für die Plattform HMC. Sie benötigt ein Portfolio von einfach zu verwendenden Werkzeugen, Diensten, Schnittstellen und IT-Infrastrukturkomponenten für Forschungsdaten und dessen Metadaten-Management, die nahtlos zusammenwirken und den Aufbau von FAIR-Prozessen und damit Informationsinfrastrukturen erleichtern.

Viele Spezifikationen und Empfehlungen für Dienste existieren bereits als Empfehlungen der Research Data Alliance, ISO, OAI, W3C, IETF und als ICT Technical Specifications der European Commission (European Commission 2017). Ebenso existieren zum Teil Implementierungen für Dienste und Werkzeuge aus großen Forschungsdateninfrastrukturprojekten, z. B. EOSC, ARDC, die nachgenutzt werden

können. So können z. B. die Spezifikations-Empfehlungen der Research Data Alliance für den Aufbau von Katalogen, Registries und Practical Policies (Moore und Stotzka 2015) direkt verwendet werden.

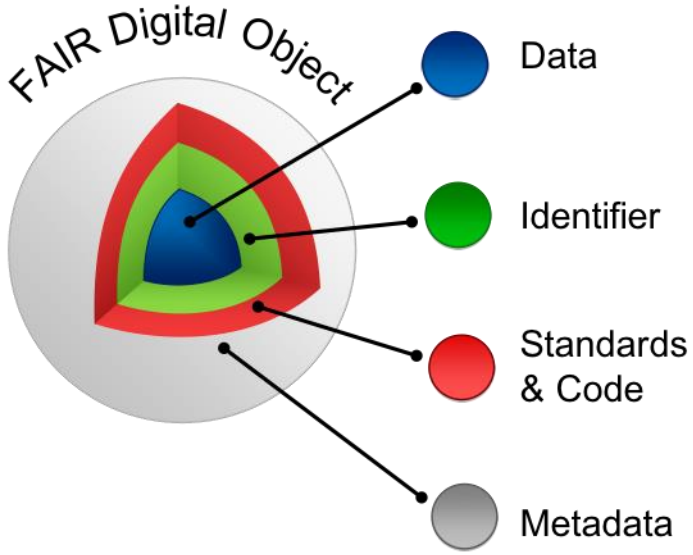

Abbildung 2: FAIR Digital Object Model: Metadaten befinden sich nicht nur in der äußeren grauen Schale, die man als inhaltlich beschreibende Metadaten der Daten interpretieren kann, sondern „verstecken" sich auch in den verwendeten Datentypen, im Identifier und Standards & Code (Hodson u. a. 2018)

Das FAIR Digital Object Model wurde in der Research Data Alliance entwickelt und bildet die Grundlage des technischen Ökosystems für FAIR Data (Hodson u. a. 2018). Es erlaubt aufgrund seiner Struktur die Realisierung der FAIR-Prinzipien, benötigt aber sichere, vertrauenswürdige und zuverlässige technische Dienste, die dezentral betrieben werden können. Es muss sichergestellt werden, dass die Komponenten interoperabel zu Daten-Diensten im internationalen Umfeld sind. Dies erfordert internationale Abstimmungsprozesse und ist die Basis für die technologische Zukunftsfähigkeit von HMC.

Während zurzeit in EOSC eher prototypische Lösungen entwickelt werden, benötigt die Helmholtz-Gemeinschaft zeitnah nutzbare und nachhaltige Dienste für den Produktionsbetrieb. Dies kann in HMC in enger Kooperation mit dem zukünftigen Forschungsbereich Information, Programm „Engineering Digital Futures", Topic „Enabling Computational- & Data-Intensive Science and Engineering" (zurzeit im Programm „Supercomputing & Big Data") und durch Anknüpfungspunkte in anderen Forschungsbereichen (z. B. im Forschungsbereich Matter, Programm „Matter and Technology", Topic „DMA – Data Management and Analysis") sowie EOSC Projekten wie z. B. EOSC-Hub, EOSC-Pillar, EOSC-Synergy und ExPaNDS (EOSC Photon and Neutron Data Services) realisiert werden. Weiterhin werden in GO FAIR (C2CAMP: Active GO FAIR Implementation Network) Testbeds entwickelt und aufgesetzt, an denen der Forschungsbereich Information aktiv beteiligt ist.

Die Dienste von HMC werden je nach Beschaffenheit dezentral in den Datenzentren der Helmholtz-Gemeinschaft betrieben, z. B. in lokalen IT-Zentren, die Basisdienste wie Server, Netzwerke, Speicher, Datenbanken, Rechenkapazitäten und die Authentifizierungs- und Autorisierungs-Infrastruktur zur Verfügung stellen oder in der Helmholtz Data Federation.

Um Aufwand und Kosten zu minimieren und um die internationale Anschlussfähigkeit zu wahren, werden eigene Entwicklungen und Software hauptsächlich dazu dienen, existierende Lösungen auf die speziellen Bedarfe anzupassen und Lücken in der Dienstelandschaft zu schließen. Die Werkzeuge und Dienste sollen, wann immer technisch realisierbar,

- als Open-Source-Komponenten zum Download angeboten werden, um eine größtmögliche internationale Verbreitung zu erreichen,
- als Dienste- und Werkzeug-Zugänge über das Informationsportal von HMC beziehungsweise über HIFIS in der Helmholtz-Gemeinschaft bereitgestellt und betrieben werden. So können alle Mitarbeiterinnen und Mitarbeiter die Ergebnisse direkt nutzen. Weiterhin wird so eine Erweiterung des Betriebs zu Schaffung der NFDI ermöglicht;
- Produktionsqualität besitzen, um einen langfristigen Betrieb und Wartung durch betreibende Datenzentren zu gewährleisten und so vertrauenswürdige, nachhaltige Meta- und Forschungsdaten-Infrastrukturen aufzubauen.

Das folgende Bild (Abbildung 3) veranschaulicht die Aufgaben in AP 2 zum Aufbau der dezentralen Komponenten des HMC Office, um FAIR technisch zu ermöglichen. Unten liegen als Basis die allgemeinen IT Dienste, die von den lokalen oder föderierten Rechenzentren, HIFIS, HDF und anderen zur Verfügung gestellt werden.

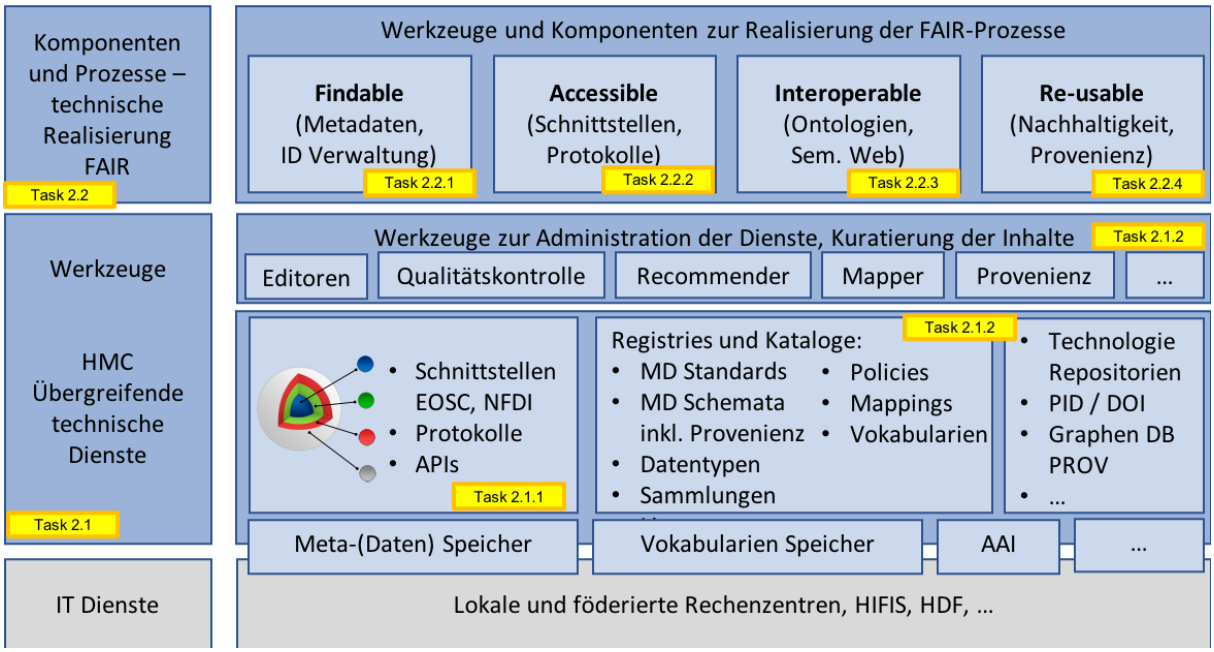

Abbildung 3: Aufbau der dezentralen Komponenten des HMC Office: technische Realisierung FAIR

AP 2.1 Übergreifende technische Dienste und Werkzeuge

Die mittlere Schicht (siehe Abbildung 3) stellt die technischen Grundlagen zur Realisierung der FAIR-Prinzipien bereit und teilt sich in folgende Sub-Tasks auf:

AP 2.1.1 Implementierung FAIR Data Object – Schnittstellen zu EOSC und NFDI

Das FAIR Digital Object, seine Schnittstellen und APIs werden in enger Zusammenarbeit mit EOSC und der Research Data Alliance entworfen und implementiert. Die Ergebnisse dienen als Blaupause für die technischen Dienste der NFDI.

Dieses Arbeitspaket erfordert eine enge Abstimmung mit AP 1.2, AP 3.1.3 und AP 4.

AP 2.1.2 Entwicklung und Betrieb technischer Dienste und Werkzeuge

Zusätzlich zu generischen Speicherdiensten für Metadaten und Vokabularien werden Kataloge und maschinenlesbare Registries für verschiedene Informationen über Metadaten-Standards, Schemata, Datentypen, etc., ähnlich dem Semantic Web, benötigt, um die FAIR Digital Objects zu unterstützen. Werkzeuge unterstützen bei der Administration der Dienste und Pflege der Informationsinhalte.

Dieses Arbeitspaket erfordert eine enge Abstimmung mit AP 1.2, AP 3.1.1 und AP 4.

AP 2.2 Komponenten und Prozesse – Realisierung der FAIR-Prinzipien

Basierend auf den übergreifenden technischen Diensten und Werkzeugen lassen sich anhand des Modells der FAIR Digital Objects (Abbildung 2) technische Prozesse zur Realisierung der einzelnen FAIR-Prinzipien beschreiben und das Zusammenspiel der Komponenten automatisieren. Dies ist in Abbildung 3 in der oberen Schicht angedeutet.

Um z. B. inhaltliche, beschreibende Metadaten einer Sammlung von Objekten in einem durchsuchbaren Verzeichnis automatisch zu registrieren und indizieren (FAIR F3), ist eine Kette von Diensten und Werkzeugen nötig:

- Datentypen können automatisch in einer Data Type Registry maschinenlesbar nachgeschlagen werden, um die Inhalte (Beispiel Typ: Metadata, Schema: „xy“) durch Programme lesen zu können;
- Metadatenschemata werden nur dann automatisch lesbar, wenn die Metadata Schema Registry existiert und angesprochen werden kann;
- Einträge der einzelnen FAIR Digital Objects können gelesen und anhand der bekannten Schemata in einem Verzeichnis registriert und indiziert werden;
- Verzeichnisse sind durchsuchbar und die Sammlung der FAIR Digital Objects kann inhaltlich erschlossen werden.

Ähnliche Prozesse lassen sich für sehr viele FAIR-Komponenten darstellen und automatisieren. Die fachspezifische Adaption der Dienste und Werkzeuge wird in den Hubs mit Unterstützung von AP 2 durchgeführt.

So verteilen sich die Sub-Tasks für die Realisierung der Prozesse und zusätzlicher Komponenten auf die vier FAIR-Prinzipien. Weiterhin werden die Prozesse ausführlich dokumentiert und über ein Informationsportal den Anwenderinnen und Anwendern sowie den Hubs zur Verfügung gestellt.

AP 2.2.1 Findable: Forschungsdaten mittels Identifier und umfassenden Metadaten auffindbar machen

Für die Auffindbarkeit von Daten und Metadaten sind in erster Linie persistente Identifier (PID), durchsuchbare Verzeichnisse und Referenzierungen notwendig. Hier kommen gegebenenfalls Metadata Harvester, Suchfunktionalitäten und Repositorien zum Einsatz. Die umfangreiche Beschreibung der Daten durch Metadaten selbst ist eine Aufgabe, die nur gemeinsam mit den Metadata Hubs in den Forschungsbereichen sowie dem AP 1.2 und AP 3.2 Wissen & Vermittlung erreicht werden kann. Einzelne Forschungsbereiche wie Erde & Umwelt können mit Pilotprojekten hier ein Vorbild sein (vgl. FREYA).

Dieses Arbeitspaket erfordert eine enge Abstimmung mit AP 1.2 und AP 3.2.

AP 2.2.2 Accessible: offene und standardisierte Protokolle und Policies

Für den Zugriff auf Meta-(Daten) sind Protokolle und Schnittstellen notwendig, die anderen zur Verfügung gestellt werden können, und auch im Rückgriff auf AAI-Systeme die Authentifizierung und Rechteverwaltung ermöglichen (vgl. z. B. aus dem Forschungsbereich Gesundheit ELIXIR oder aus dem

Forschungsbereich Information „Authentication and Authorisation for Research and Collaboration“ AARC). Eine Policy des nachhaltigen Zugriffs auf Metadaten, selbst wenn beschlossen wird, die Forschungsdaten nicht länger vorzuhalten, kann beispielsweise über ein Repositorium mit Exitstrategie technisch realisiert werden.

Dieses Arbeitspaket erfordert eine enge Abstimmung mit AP 3.1 und AP 3.3.4.

AP 2.2.3 Interoperable: FAIRe Wissensrepräsentation durch offene Vokabularien, Ontologien und Standards

Um Vokabularien, Ontologien und domänenspezifische Standards zu erschaffen, benötigen die einzelnen Metadata Hubs (AP 3) Werkzeuge und Prozesse zur Erzeugung, Verwaltung und Interaktion von Vokabularien (im Einzelnen sind dies z. B. Editor, Registry, Werkzeuge zur Einbindung in Web-Formulare, Mapper, Storage). Diese Aufgabe zeigt ganz besonders die Abhängigkeit der nutzerorientierten, einfachen Werkzeuge vom Gesamtgeflecht der technischen Komponenten.

Dieses Arbeitspaket erfordert eine enge Abstimmung mit AP 1.3.2, AP 3.2.1 und AP 3.3

AP 2.2.4 Re-Usable: Meta-(Daten), ihre Provenienz und Nutzungslizenz sind detailliert beschrieben

Meta-(Daten) nachnutzbar zu machen, bedeutet auch, den Weg ihrer Entstehung in allen Einzelheiten nachverfolgen zu können – bis hin zur Reproduzierbarkeit von Forschungsergebnissen, die entscheidend zur Qualitätssicherung beitragen. Während die Beschreibung der Meta-(Daten) sowie die Community Standards bei den Metadata Hubs liegen müssen, stellen die technischen Komponenten Policy-Registry und Werkzeuge zur Provenienz wichtige Voraussetzungen zur Nachnutzbarkeit der Meta-(Daten) dar.

Eine vollständige Digitalisierung bis hin zur Provenienz erlaubt es, Reproduzierbarkeit als automatisierten Prozess aus der Verbindung von Daten, Metadaten, Provenienz, Methoden und Werkzeugen zu entwickeln, die alle Schritte in Richtung eines umfassenden Datenverständnisses reproduzierbar digital zur Verfügung stellt und so die Entstehung von Forschungsergebnissen aus Daten vollständig transparent und durch geeignete Technologien nachvollziehbar im Sinne eines digitalen Workflows zu machen.

Dieses Arbeitspaket erfordert eine enge Abstimmung mit AP 1.3.5 und AP 3.3.6.

AP 3: Domänenspezifische Leistung – Metadata Hubs

Die domänenspezifischen Leistungen werden in den Metadata Hubs erbracht, die pro Forschungsbereich aufgebaut werden. Deren Aufgaben orientieren sich an der Metadatenerzeugung und -nutzung in den Domänen. Diese liegen in den Bereichen Koordination und Management, Wissenssammlung und -vermittlung, praktische Nutzung und Beratung sowie entsprechender informationstechnischer Unterstützung. Die Ausprägung der einzelnen Aufgaben innerhalb der Forschungsbereiche kann variieren, um kulturellen und strukturellen Unterschieden gerecht werden zu können, im Grundsatz sind die Aufgabenbereiche pro Forschungsbereich jedoch vergleichbar. Allen gemeinsam ist die enge Einbindung der Expertise und der Bedarfe der jeweiligen Community. In einigen Aufgabenbereichen entwickeln die Hubs zentrale Elemente weiter, beziehungsweise diversifizieren sie, in anderen Aufgabenfelder liefern sie den zentralen Bereichen zu. In jedem Fall ist eine enge Verzahnung der zentralen und domänenspezifischen Aufgaben zu gewährleisten.

AP 3.1. Koordination und Management

Die Koordination und das Management eines Metadata Hubs, gegebenenfalls über die beteiligten Zentren hinweg, sind entscheidend für dessen Erfolg nach Innen und Außen. Daher ist dies eine kritische Aufgabe im Rahmen der Kommunikation und Vernetzung der spezifischen Aktivitäten innerhalb eines Forschungsbereichs sowie für HMC insgesamt. Bei den domänenspezifischen Koordinationsstellen entscheidet sich, ob der notwendige enge Austausch und die Abstimmung der inhaltlichen und dateninfrastrukturtechnischen Entwicklungen (a) innerhalb des Hubs und (b) übergreifend mit den anderen Metadata Hubs sowie (c) die Verknüpfung zu den dezentralen und zentralen Aufgaben des HMC Office funktioniert.

Dieses Arbeitspaket erfordert eine enge Abstimmung mit AP 1.1.

AP 3.1.1 Einbeziehung der Community-Expertise

Alle technischen und organisatorischen Maßnahmen in HMC werden sich an der Akzeptanz innerhalb der jeweiligen Community orientieren. Ein wesentlicher Hebel ist daher die Identifikation und aktive Einbeziehung der Community-Expertise. Diese wird häufig durch Akteure repräsentiert, die sowohl in der Forschung als auch in den Metadatenspezifika der Domäne aktiv sind und somit eine optimale Schnittstelle zwischen Forschung und Metadateninfrastruktur darstellen. Die Koordinationsstelle baut Mechanismen auf, die den Einfluss dieser Akteure sowohl auf die Entwicklung als auch die Nutzbarkeit von Katalogen und Werkzeugen sicherstellt. Dies kann auf der Ebene der AG Metadaten (s. AP 1.1.5), durch das dynamische Instrument der internen Projektförderung in einzelnen Zentren oder durch die Einbindung im Rahmen der Entwicklung domänenweit nutzbarer Mechanismen zur Qualitätssicherung der Daten (Kuration) geschehen.

Dieses Arbeitspaket erfordert eine enge Abstimmung mit AP 1.1.5., AP 1.2. und AP 2.2.

AP 3.1.2 Aufbau und Betreuung einer „Metadaten-Community" im jeweiligen Forschungsbereich

Ziel der geplanten Community-Maßnahmen, wie Bestandsaufnahme der inhaltlich-technischen Metadatenexpertise, Wissensvermittlung sowie lokale Beratung und Projektförderung, ist eine stetig wachsende Metadatenexpertise unter den Forschenden. Aufgabe der Koordinationsstelle eines Metadata Hubs ist es, dieser wachsenden domänenspezifischen Metadaten-Community über das Informationsportal (s. AP 1.2.3) sowie Meetings und Workshops ein Forum zu bieten als auch Austausch und Kooperation über Forschungsbereiche hinweg zu fördern. Die Angebote entlang der Bedarfe der Forschenden sollen im Verlauf des Plattformaufbaus zunehmen und schließlich in Form verlässlicher Programme zyklisch und unter kontinuierlicher Anpassung stattfinden.

Dieses Arbeitspaket erfordert eine enge Abstimmung mit AP 1.1.5.

AP 3.1.3 Internationale Vernetzung, Gremienarbeit, Harmonisierung und Standardisierung

Dieses Arbeitspaket deckt die domänenspezifische Partizipation an internationalen Entwicklungen ab. Davon sind vor allem die Themenfelder der Harmonisierung und Standardisierung betroffen, aber auch die generelle Lobbyarbeit für die HMC Plattform in den jeweiligen Communities. Initiale Aufgabe ist die Identifikation domänenspezifischer Initiativen und Akteure in den Kontexten von FAIR Data (GO FAIR, Enabling FAIR, fairsharing.org, re3data, FAIRsFAIR), der Research Data Alliance (RDA), EOSC Projekten, etc. Wo notwendig, müssen neue Akteure für eine Beteiligung der Helmholtz-Forschungsbereiche in spezifischen internationalen Metadatenprojekten benannt werden.

Dieses Arbeitspaket erfordert eine enge Abstimmung mit AP 1.1.4.

AP 3.2 Wissen & Vermittlung

Im Arbeitspaket Wissen & Vermittlung wird das metadatenspezifische Methodenwissen der Community gesammelt, aufgebaut und über das Informationsportal (s. AP 1.2.3) verfügbar gemacht. Darüber hinaus werden domänenspezifische Schulungsangebote sowie Projekt- und Individualberatung entwickelt und angeboten. Im Fokus steht die Wissensvermittlung zu domänenspezifischen Vokabularien, Ontologien, Methoden, Workflows und Werkzeugen, die im Metadata Hub gesammelt und weiterentwickelt werden. Ein weiteres Ergebnis sind domänenspezifische Empfehlungen, Prozess- und Best Practice Beschreibungen.

Dieses Arbeitspaket erfordert eine enge Abstimmung mit AP 1.2.

AP 3.2.1 Aufbau und Bereitstellung einer Informationsbasis zu Metadaten(-vokabularien), Ontologien und Standards

Metadaten sind für viele Forschende eine sperrige und abstrakte Thematik. Ziel der Metadateninformationsbasis ist es daher, die domänenspezifischen Informationen zu Metadaten, einschließlich einschlägigen Standards und Ontologien (die in AP 1.3.2 und domänenspezifisch in 2.2.3 erstellt werden) zu sammeln und so aufzubereiten, dass sie Forschenden nicht nur Informationen über Metadaten über das Informationsportal bereitstellt, sondern insbesondere eine Orientierung in Bezug auf deren Nutzung in ihrem jeweiligen Forschungsgebiet liefert.

Dieses Arbeitspaket erfordert zudem eine enge Abstimmung mit AP 1.3.2 und AP 2.2.3.

AP 3.2.2. Landschaft der Forschungsdatenexpertisen

Die Identifikation der existierenden Datensammlungen und die Erfassung der notwendigen beschreibenden Metadaten und Datentags etc. führt zu einer umfassenden Kartierung der Forschungsdatenexpertise innerhalb der Helmholtz-Gemeinschaft. Allein in re3data, einem durchsuchbaren Portal, in dem mit strukturierten Metadaten beschriebene „Steckbriefe“ für Datenrepositorien und Portale auffindbar sind, können 106 Repositorien und Portale identifiziert werden, die mit Beteiligung von Helmholtz-Zentren betrieben werden. Das umfangreiche Metadatenmodell von re3data (re3data.org – Registry of Research Data Repositories 2015) beschreibt allein die registrierten Portale durch 41 Metadatenfelder zu allgemeinen Informationen, Zuständigkeiten, Policies und rechtliche Aspekte, aber auch zu technischen und Qualitäts-Standards, wie Schnittstellen und Metadatenstandards. Jeder Eintrag wird durch eine DOI eindeutig referenziert. Ursprünglich von der DFG gefördert und mit Beteiligung zweier Helmholtz Zentren entwickelt, ist der Dienst seit 2016 in DataCite integriert. Diese Basis wird erweitert und ausgebaut.

Durch fachspezifische Portale kann die Gesamtheit aller Datenschätze aus Helmholtz erschlossen und in der EOSC und anderen Open Science Portalen sichtbar gemacht und verfolgt werden. Voraussetzung ist dabei durch einheitliche Formate die Datenprovenienz zu sichern. In der Folge werden die Daten, ihre Erzeuger und die Helmholtz-Gemeinschaft als Datenproduzent sichtbarer. Die Nachnutzung wird ermöglicht und die Zentren können diese Kartierung als Pluspunkt bei der Beteiligung übergeordneter Initiativen, wie der NFDI, einbringen.

Die Kartierung wird möglichst auch die Erfassung der personellen Expertise innerhalb des Forschungsbereiches abbilden, wie sie in AP 3.1.1 identifiziert und in AP 3.1.2. weiterentwickelt wird.

Dieses Arbeitspaket erfordert eine enge Abstimmung mit AP 1.1.5, AP 1.2.1 und AP 1.2.4.

AP 3.2.3 Domänenspezifische Ergänzung zu Methodenwissen, Basisinformationen und Schulungsinhalten

Ergänzend zu den allgemeinen Wissenssammlungen (s. AP 1.2.2 und 1.2.3) werden hier die domänenspezifischen Aspekte aufgebaut und gemeinsam für alle Forschungsbereiche über das Informationsportal verfügbar gemacht sowie Schulungen über Plattform HIDA angeboten. Die Inhalte beantworten Fragen zur Wahl und Nutzung der für eine Subdomäne passenden Werkzeuge und Workflows.

Dieses Arbeitspaket erfordert eine enge Abstimmung mit AP 1.2.2 und AP 1.2.3.

AP 3.2.4 Beratung von Zentren, Projekten und Forschenden

Der Beratung von Zentren, Projekten und Forschenden innerhalb der Metadata Hubs kommt eine große Bedeutung zu, da dieses Format besonders geeignet ist, praktische Probleme im Forschungsalltag der Subdomänen zu identifizieren und individuelle Lösungsmöglichkeiten zu entwickeln.

Die Zentrenberatung fokussiert auf die organisatorische (infra-)strukturelle Verankerung des Themas Metadaten innerhalb der Einrichtung. Welche Subdomänen und Forschenden sind aktiv eingebunden, welche zentralen Aufgaben können am Zentrum erfüllt werden, welche nicht? Ziel ist es, in allen Zentren auf ein ähnliches und verlässliches Maß an Know-How für dieses zentrale Zukunftsthema der Auffindbarkeit eigener Daten zu kommen. Die Beratung bedingt eine enge Abstimmung mit bereits bestehenden Strukturen in den Zentren, z. B. Forschungsdatenmanagementteams.

Die Projektberatung begleitet den gesamten Prozess der Nutzung der HMC-internen Projektförderung. Der Fokus liegt auf der inhaltlichen Nutzbarkeit der Fragestellungen und der Einordnung ihrer Umsetzung im Kontext der bereits existierenden Infrastrukturen, Werkzeuge, Dienste, Standards, etc. Die Beratung findet in Form von Gesprächen statt und wird im Rahmen der Organisation der Projektausschreibung und -begleitung proaktiv angeboten.

Die individuelle Beratung zielt auf einzelne Forschende oder Forschungsgruppen, die beispielsweise ein spezifisches Problem lösen möchten, ohne die Projektförderung in Anspruch nehmen zu können, oder die Hilfestellung bei konkreten Schritten der Integration der von HMC angebotenen Dienstleistungen in ihre eigenen Workflows haben, etc.

Dieses Arbeitspaket erfordert eine enge Abstimmung mit AP 1.2.4 und AP 1.2.5

AP 3.3. Komponenten und Prozesse

In den Forschungsbereichen werden bereits vereinzelt Daten-Workflows, Metadatenstandards und Werkzeuge zu deren Verknüpfung mit Daten entwickelt und/oder eingesetzt. In diesem Arbeitspaket werden die vorhandenen technischen Komponenten und etablierten Prozesse erfasst, Lücken identifiziert und Szenarien zur Ergänzung und Weiterentwicklung in den Domänen entworfen. Ziel ist die einfache und FAIRe Erschließung und Nutzbarkeit vorhandener und zukünftiger Datensammlungen der Forschungsbereiche sowie die Befähigung der Forschenden, FAIRe Daten (semi-)automatisch zu erstellen. Hierzu benötigen sie neben dem notwendigen Wissen (s. AP 3.2.) auf ihre Anforderungen zugeschnittene Werkzeuge und Prozesse. Dieses Arbeitspaket erfordert eine enge Zusammenarbeit mit Arbeitspaket 2, da gemeinsam die Anforderungen aus den Bedürfnissen der Domänen formuliert werden, um sie im Arbeitspaket 2 in technische Lösungen umzusetzen und in Arbeitspaket 3.3 zu nutzen. Als Basis können die Prozesse und Werkzeuge aus 2.2.1 – 2.2.4 dienen.

Dieses Arbeitspaket erfordert eine enge Abstimmung mit AP 2.1 und AP 2.2.

AP 3.3.1 Prozesse, Werkzeuge und Dienste zur Erschließung von Forschungsdatensammlungen aufbauen

Ziel ist es, bestehende Prozesse, Werkzeuge und Dienste zur Erschließung von Forschungsdatensammlungen pro Forschungsbereich zu erheben, zu beschreiben, nutzbar zu machen und, wo notwendig, aufzubauen. Dabei ist eine Verschränkung dieser für HMC zentralen Aufgabe mit den dezentralen und zentralen Leistungen des HMC Office notwendig: (1) die Beschreibungen für die Sammlungen von bestehenden Prozessen, Diensten und Werkzeugen werden übergreifend abgestimmt, damit es über die gesamte Plattform hinweg ein einheitliches Format gibt. (2) die Nutzbarmachung erfolgt in Form eines technischen Formats und Zugangs, die nach Absprache aller Hubs mit den dezentralen Leistungen (AP 2) durch letztere umgesetzt und domänenspezifisch angepasst werden. Kritisch ist die Abstimmung des Anforderungsmanagements innerhalb der Hubs sowie zwischen den Arbeitsbereichen.

Dieses Arbeitspaket erfordert eine enge Abstimmung mit AP 1.2.1, AP 1.2.2 und AP 2.2.1 bis 2.2.4.

AP 3.3.2 Ingest: Werkzeuge zur Automatisierung der Erfassung von Metadaten

Metadaten sollen, soweit wie möglich, automatisiert erfasst und zu den entsprechenden Daten annotiert werden, was für viele datenerzeugende Geräte möglich ist, aber oft in einem spezifischen Umfeld umgesetzt beziehungsweise auf Basis bestehender Lösungen angepasst werden muss. Nach der Erhebung bestehender Lösungen werden deren Entwicklungs- und Adaptionspotentiale bewertet und auf einige typische, domänenspezifische Probleme angewendet. Die Auswahl wird nach Rückkopplung aller Forschungsbereiche innerhalb des einzelnen Forschungsbereichs festgelegt. Auch Projekte (s. Arbeitsbereich 4) können zielführende Anwendungsfälle mit sich bringen.

Dieses Arbeitspaket erfordert eine enge Abstimmung mit AP 1.3.3, AP 2.1.2 und AP 2.2.1

AP 3.3.3 Werkzeuge zur Handhabung verschiedener Metadatenstandards

Metadatenstandards gibt es viele, auch disziplinenspezifisch herrscht oft Vielfalt vor. Werkzeuge zur Handhabung verschiedener Metadatenstandards unterstützen daher die Arbeit mit (a) Daten unterschiedlicher Quellen und damit oft unterschiedlicher Annotationen, z. B. zur Verbesserung der Vergleichbarkeit, und (b) unterschiedlichen Metadatenstandards für dieselben Datensätze, wie es für unterschiedliche Kontexte notwendig sein kann. Die Werkzeuge werden entlang der in einer Domäne tatsächlich verwendeten Standards (weiter-)entwickelt, um die praktische Nutzbarkeit sicherzustellen. Auch werden zunächst prototypisch geeignete Beispiel pro Hub in gegenseitiger Absprache umgesetzt.

Dieses Arbeitspaket erfordert eine enge Abstimmung mit AP 2.2.1 bis 2.2.4.

AP 3.3.4 Zugang zu Forschungsdaten über Schnittstellen schaffen

Ziel ist es, dass alle Forschungsdaten über offene, gebräuchliche Schnittstellen einfach zugänglich sind, ohne auf spezielle Werkzeuge für den Zugang angewiesen zu sein. Gleichzeitig soll eine Authentifizierung für solche Daten möglich sein, die nicht offen sind beziehungsweise sein dürfen. Entsprechend müssen alle (Meta-)Datenrepositorien, die im Rahmen von HMC genutzt werden, insbesondere die bereits bestehenden, entsprechend weiterentwickelt werden, sollten sie diesem Ziel nicht entsprechen. Ferner sollten Metadaten auch dann noch zugänglich sein, wenn die Daten selbst nicht mehr verfügbar sind, z. B. Urheber oder verknüpfte Publikationen (vgl. AP 2.2.2 z. B. durch ein Repositorium mit Exitstrategie).

Diese Kriterien werden in der Sammlung bestehender Prozesse und Werkzeuge überprüft und gegebenenfalls ergänzt.

Dieses Arbeitspaket erfordert eine enge Abstimmung mit AP 2.2.1 bis 2.2.4.

AP 3.3.5 Erfassen der FAIRness der Daten

Mit der Erschließung von Forschungsdatensammlungen wird auch der Zustand ihrer Beschreibung bezüglich der FAIR-Prinzipien erfasst. Hierzu wird in AP 2.1.2 ein einfaches Bewertungswerkzeug bereitgestellt, das eine semi-automatische Bestandsaufnahme entlang der FAIR-Kriterien unterstützt und die Ergebnisse zugleich für die Entwicklung der Werkzeuge in 2.2 nutzbar macht.

Dieses Arbeitspaket erfordert eine enge Abstimmung mit AP 1.3.4 und AP 2.

AP 3.3.6 Provenienz

Datenprovenienz und Datentracking sind essentiell für die lückenlose Rückverfolgung von Daten bis zu ihren Ursprüngen, sowohl, was die Erzeugung der Daten angeht, ihre Urheber, ihre Verwendung im Laufe der Zeit, die Zitation und mögliche Versionen, Prozessierungen, Abwandlungen und Aggregationen. Wie werden beispielsweise unterschiedliche Daten verfolgbar gemacht, die mit Hilfe unterschiedlicher Analysemethoden durch unterschiedliche Personen an verschiedenen Orten auf Basis ein und derselben, mehrfach geteilten Probe durchgeführt werden?

Die Umsetzung einer lückenlosen Verfolgbarkeit der Daten ist sehr komplex und greift an vielen Schritten des (Meta-)Datenworkflows ein. Für jede mögliche Veränderung müssen Felder/Beschreibungen und Policies entwickelt und vorgehalten werden. Daher ist diese Aufgabe sehr eng gekoppelt an die domänenspezifischen Workflowbeschreibungen, die Werkzeugentwicklung in 2.2 sowie Provenienzrichtlinien, wie sie in AP 1.3.3 erarbeitet werden. Die technischen Lösungen dafür sind für alle Zentren und die Helmholtz Gemeinschaft insgesamt von höchster Relevanz für Qualitätssicherung, das Monitoring und die Steuerung von Leistung in Bezug auf Daten und den digitalen Kulturwandel insgesamt.

Dieses Arbeitspaket erfordert eine enge Abstimmung mit AP 1.3.3 und AP 2.2.

AP 4: HMC Projekte

Neben den statischen Komponenten realisiert durch die Metadata Hubs sowie das HMC Office, verfügt HMC über dynamische Elemente in Form der IVF-finanzierten Projekte (diese machen anteilig 25% der Gesamtfinanzierung von HMC aus und erfordern bei einer 50%igen Finanzierung durch den IVF 50% Eigenfinanzierung aus den sich beteiligenden Zentren). Diese übernehmen zielgerichtet die Entwicklung spezifischer Werkzeuge und Lösung von Aufgaben, die sich im operativen Betrieb von HMC ergeben und dessen Angebot dynamisch ergänzen und adaptieren. Die Projekte werden regelmäßig durch das HMC Office ausgeschrieben und in der Umsetzung begleitet und gesteuert. So wird HMC zu einer wachsenden Plattform, die sich den Entwicklungen in den Forschungsbereichen und der Gemeinschaft als Ganzes anpassen und angemessen darauf reagieren kann.

Für die Zielgruppe der Forschenden ist der wissenschaftliche Mehrwert, der durch das Auffinden (von Daten und Metadaten) und die Auswertung der Metadaten sowie durch die Möglichkeit interdisziplinärer Arbeit erzielt wird, das entscheidende Argument. Für die Projekte können sich fortlaufend Forschende bewerben (Vergabe erfolgt über den Impuls- und Vernetzungsfonds), die konkrete Leistungen für die Lösung von Metadatenproblemen benötigen.

Jährlich sind je nach Umfang der Projekte etwa vier Projekte aus verschiedenen Forschungsbereichen vorgesehen, die aus dem Impuls- und Vernetzungsfonds finanziert werden sollen. Insgesamt ist damit zu rechnen, dass nach erfolgreichem Aufbau des HMC der Wunsch nach Problemlösungen innerhalb der Helmholtz-Gemeinschaft größer wird und mehr Bewerbungen für Projekte eingehen.

Pro Projekt ist (zusätzlich zu den dynamischen Mitteln aus dem Impuls- und Vernetzungsfonds) Mitarbeit aus den Metadata Hubs vorgesehen, um die Lösungsansätze direkt in die Plattform zu integrieren und um darauf zu achten, dass die Lösungen interoperabel sind. Im Anschluss an die erfolgreiche Durchführung der Projekte muss außerdem aus dem Bereich Vermittlung dafür Sorge getragen werden, dass die Ergebnisse sichtbar gemacht werden. Die Anfragen zu Projekten nimmt das HMC Office entgegen.

Beispiele für Projekte aus den Forschungsbereichen

In mehreren Forschungsbereichen werden aktuell Problemlösungen mit Hilfe von Linked Open Data oder für das Tracking von Analysedaten innerhalb eines Data Flow Frameworks gesucht. Letzteres betrifft ebenfalls das Thema der Provenienz von Daten. Die semantische Verknüpfung und Automatisierung führt zu erweiterten Analysemöglichkeiten, die neue wissenschaftliche Erkenntnisse befördern. Aus dem Bereich Luft-, Raumfahrt und Verkehr werden Methoden der künstlichen Intelligenz benötigt, um Metadaten aus Rohdaten zu extrahieren und auszuwerten. Dies könnte in enger Kooperation mit HAICU geschehen.

Best Practices zu Sensor- und Detektorbeschreibungen

Sensoren werden in verschiedenen Forschungsbereichen eingesetzt. Sowohl bei ihrem Einsatz als auch bei der Auswertung der Daten müssen Forschende sie eindeutig und auswertbar beschreiben. Zur Erarbeitung von Best Practices werden eindeutige Sensorbeschreibungen und Metadatenstandards ausgewählt und Prozesse erarbeitet, die z. B. Registrierung, Arbeitsweise, Messmethode, Einsatzgeschichte oder Anwendungsszenarien für Sensoren klar definieren. Vorhaben und Ergebnisse werden mit internationalen Gremien (z. B. Arbeitsgruppen der RDA wie „Persistent Identification of Instruments“ oder „From Observational Data to Information“) abgestimmt beziehungsweise dort eingebracht. Entstandene Werkzeuge werden in Kooperation mit HMC möglichst universell gestaltet, um auch in anderen Forschungsbereichen anwendbar zu sein.

Detektoren als aus einzelnen Sensoren zusammengesetzte Diagnostikinstrumente bedürfen darüber hinaus eigener Beschreibungen, die die Beziehungen der einzelnen Komponenten zueinander erfassen. Dies betrifft auch wichtige Leistungsdaten wie Kalibrierung, Dynamikbereich, Auflösung in Zeit- und Raum, Totzeitverhalten oder maximale Ereignisrate. An Großgeräten im Forschungsbereich Materie existieren bereits umfangreiche Datensammlungen zu Detektoren inklusive vollständige funktionale und strukturelle Metadaten, deren physikalische Eigenschaften in die Messung miteinbezogen und dokumentiert werden, jedoch ständig ergänzt, erneuert und erweitert werden müssen.

Best Practices für integrierte Versuchsdokumentation mit umfassenden Metadaten

Unter Einbeziehung internationaler Empfehlungen werden Best Practices zur integrierten Versuchsdokumentation mit umfassenden Metadaten erarbeitet, die Experimente reproduzierbar und auswertbar beschreiben. Der Fokus liegt dabei auf einfachen Workflows für Anwender, welche einen größtmöglichen Grad an Automatisierung besitzen. Daher werden Schnittstellen zu anderer Software (z. B. von Messgeräten oder Kontrollsystemen großer Experimente) analysiert und inhaltlich spezifiziert. Durch Daten- und Metadatenübernahme sollen Doppeldokumentation und hohe Dokumentationsaufwände vermieden werden. Die Integration von Nutzerinput in standardisierter Weise ist extrem wichtig und eine besondere Herausforderung an Großgeräten mit hoher Nutzerfluktuation und hohen Datenraten. Durch Integration eines Vokabulars kann die Qualität der Metadaten gesteigert werden. Erstellte Best Practices oder Teile daraus werden frühzeitig als Vorschläge in einen internationalen Konsensprozess eingebracht (u. a. bei RDA und ISO), um eine Interoperabilität sowohl auf Prozess- als auch technischer Ebene zu erreichen.

Prozesse zur Reproduzierbarkeit von Analysen mit Metadaten zur Provenienz

In vielen Forschungsbereichen ist die Provenienz von Daten ein wichtiges Problem, z. B. die Beschreibung der Provenienz, welche nachgelagerte, teilweise manuelle, und zeitlich stark versetzte Prozessierungsschritte der Daten dokumentiert. Das Ziel ist, die Vergleichbarkeit von Ergebnissen herzustellen, indem die Entstehung der Daten und deren Analyseschritte reproduzierbar in geeigneten Werkzeugen (z. B. Versionskontrollsystemen oder Workflowbeschreibungssystemen) dokumentiert werden. Diese Metadaten zur Provenienz ermöglichen die Einschätzung, ob Ergebnisse miteinander vergleichbar sind oder wie sie weiter (automatisch) ausgewertet und interpretiert werden können, und stellen zugleich die gesamte Kette der Verarbeitungsschritte bis zum ursprünglichen Forschungsdatensatz dar. Dazu müssen die Prozesse der Datenentstehung und der Analyse durch die automatische Metadatenerhebung in hoher Qualität ergänzt werden. Ein wichtiger Aspekt ist die Integration dieser Metadaten mit den passenden Datenformatbeschreibungen, Lese- wie Auswertungssoftware und speziellen technischen Gegebenheiten der Analyse (beispielsweise die Notwendigkeit eines bestimmten Hochleistungsrechners, oder die Einbettung interaktiver Analyseschritte) zu verbinden, die integraler Bestandteil einer reproduzierbaren Datenanalyse sind. Die Entwicklung von einheitlichen Vokabularien zur Beschreibung und Klassifikation von Analyseprozessen ermöglicht das Auffinden (Find) nicht nur von Primärdaten, sondern erlaubt auch die spezifische Suche nach bereits auf bestimmten Daten ausgeführten Analyseprozessen aufgrund der in den Metadaten hinterlegten Provenienz, und erleichtert somit den kollaborativen wissenschaftlichen Austausch. Auch hier wird der Ansatz verfolgt, spezifische Anforderungen bestimmter Bereiche von generell anwendbaren Techniken zu trennen. Auf diese Weise werden Analyseskripte verständlicher, der Bezug zu den Ausgangsdaten explizit und eindeutig, Metadaten austauschbar sowie Visualisierungsschritte und Validierungsworkflows vereinfacht, beispielsweise im Rahmen internationaler Konsortien wie den H2020 FET Flagships Human Brain Project, Battery 2030+ und an den Großforschungsanlagen im Forschungsbereich Materie.

Empfehlungen zu domänenübergreifenden Analysen

Daten aus verschiedenen Bereichen zusammenzuführen, erfordert umfangreiche Metadaten, um vergleichbare Datenelemente zu identifizieren. Es ist zu erarbeiten, welche Schritte generell für solche Analysen empfehlenswert sind, um verlässliche Ergebnisse zu erhalten. Dazu reicht es nicht, die technischen Herausforderungen zu lösen, sondern es muss ein starker Fokus auf die heterogene Terminologie der einzelnen Domänen gelegt werden, um diese angleichen zu können. Dazu gehören auch Qualitätsmetriken für Metadaten und Daten, sowie einzuhaltende Prozesse. Daraus entstehende Empfehlungen ermöglichen anderen Forschenden, ihren Umgang mit Metadaten zu verbessern. Teilaufgaben dieser Art könnten als Projekte ausgestaltet und beantragt werden. In Zusammenarbeit mit Großgerätebetreibern können community-übergreifende Standards als Grundlage für die genannte Diversifikation dienen, da ihnen grundlegend die Nutzung der Großgeräte wie vorhandenen Messplätze gemein ist, jedoch eine starke Diversifikation in ihrer Nutzung in den einzelnen Wissenschaftsdisziplinen zu finden ist. Domänenübergreifende Analysen werden dabei in erster Linie auf bereits durch spezifisch von einzelnen Domänenwissenschaftlern und Domänenwissenschaftlerinnen ausgewerteten Daten geschehen. Eine Erweiterung auf andere Domänen in einem frühen Stadium der Erstanalyse von Daten mitzudenken, ist eine erhebliche Herausforderung, die bisherige wissenschaftliche Arbeitsweisen nachhaltig verändern kann.

Best Practices für Metadaten zu Messreihen

Messreihen werden vermehrt in speziellen Zeitreihendatenbanken gespeichert. Die Metriken der Zeitreihen und weiterer Parameter sind für die Suchanfragen relevant und variieren je nach Anwendung. Damit wird eine generische Beschreibung dieser Datensätze durch Metadaten erforderlich. Eine solche Beschreibung kann sich auch unterschiedlichen Speicherstrukturen anpassen. Beispielsweise werden

zeitlich hochaufgelöste Langzeitmessungen oft in Dateien gespeichert, deren Namen aus einzelnen Metadateneinheiten zusammengesetzt sind und damit sowohl den Dateninhalt beschreiben als auch Lokalisierungsfunktion haben. Ziele wie Skalierbarkeit und Performance sollen auch durch geeignete, beispielsweise Microservice-basierte Architekturen adressiert werden. Wesentlich ist auch die Untersuchung der Schnittstellen zu Datenanalyseservices und Visualisierungen oder bei der Erfassung zu den Datenmodellen der Messgerätekommunikation. Zeitreihen, oftmals auch mit sehr großem Datenvolumen, spielen in vielen Forschungsbereichen eine große Rolle. Gerade an Großgeräten ist eine synchronisierte Abnahme zeitabhängiger Daten und Metadaten in großer Menge und Volumina in Zukunft unumgänglich. Dabei ist zu beachten, dass die kontrollierte Abnahme von Zeitreihen die Integration vieler zeitabhängiger Daten wie Metadaten nötig macht, um eine Integrität über den gesamten Messverlauf zu gewährleisten, insbesondere in einer Nachanalyse bei Auftreten von Anomalien.

Qualitätskriterien für Metadaten

Die Qualität, respektive (Nach-)Nutzbarkeit, „fitness for use“, eines Datensatzes erschließt sich den Nutzerinnen und Nutzern durch die Beurteilung der Metadaten. Diese sollten im Idealfall nicht nur Aufschluss über die Herkunft der Daten (wer hat was wann und wo genommen) geben, sondern es sollten auch mögliche Fehlerquellen und Geräte- und Messtoleranzen bis hin zu (bekannten) Kalibrierungsfehlern vermerkt sein. Bei prozessierten und/oder aggregierten Daten sollte die Provenienz möglichst lückenlos vermerkt sein. Das Zusammenspiel von Daten und Metadaten erlaubt es dann individuell zu entscheiden, ob ein bestimmter Datensatz für die geplante wissenschaftliche Nutzung/Analyse als Ausgangsbasis dienen kann oder nicht. Diese Entscheidung ist in hohem Maße nichttrivial und hängt entscheidend von der Qualität der Metadaten ab.

Daher muss man systematisch erfassen, welche Metadatenparameter für die Qualitätsbeurteilung von Daten und für deren Nachnutzung essentiell sind. Dabei sollten verschiedene Nachnutzungsszenarien, sowie domänenspezifische Spezifika in Betracht gezogen werden. Die daraus resultierenden Empfehlungen würden anschließend im Wissenspool zur Vermittlung zur Verfügung stehen. Sie würden auch Grundlage für Handlungsempfehlungen und Best-Practices für die Kuration von Metadaten bei der Archivierung und Publikation der Daten liefern. Zudem ist es notwendig, die Qualitätsanalyse in einem konsequenten und so gut es möglich ist automatischem oder durch Vorgaben eingeschränkten Verfahren zu einem dauerhaften, nebenläufig zur Metadatenerfassung ablaufenden, Verfahren zu entwickeln, damit eine stets einheitlich hohe Qualität gewährleistet werden kann. An Großforschungsanlagen mit hoher Fluktuation von Nutzern unterschiedlicher Provenienz müssen sinnvolle und praktikable Standards für standardisierte Nutzungsansprüche etabliert werden.

Automatische Annotation von Daten

Aus gelabelten Trainingsdaten können mittels Machine-Learning-Verfahren Klassifikations- oder Regressionsverfahren zur Vorhersage von bestimmten Metadatenattributen trainiert werden. Z. B. könnte das „environment / sampling location“ (soil, marine, human skin, etc.) von Metaomics-Daten mikrobieller Gemeinschaften vorhergesagt werden (z. B. bei Omics-Daten von Pathogenen (Resistenzphänotyp bei klinischen, bakteriellen Isolaten) von Patienten). Mithilfe derartiger Methoden könnten anschließend manuell erstellte, unvollständige Metadatensätze automatisch erweitert oder homogenisiert werden, und sie dienten als Konsistenzüberprüfung für manuell erstellte Metadaten.

Die automatische Annotation von Daten mit Metadaten wird insbesondere an Großforschungsanlagen von herausragender Bedeutung werden, wenn hohe Datenraten menschliche Eingriffe erschweren. Sie kann auch dazu genutzt werden, Hilfestellung für eine heterogene Nutzerschaft anzubieten und durch intelligente Vorschlagsysteme die Qualität von Annotationen sowie die Bereitschaft der Nutzung von Annotationen seitens der Nutzerinnen und Nutzer zu unterstützen. Diese Systeme können dann auch

eingesetzt werden, um menschliche Annotationen in ihrer Qualität zu prüfen und so auch bei großen (Meta-) Datenmengen Qualitätsstandards überprüfbar zu machen. Die Gefahr fehlerhafter automatischer Annotationen steht und fällt dabei mit der vorhandenen Grundmenge an Beispieldaten und ihrer Qualität, die insbesondere an Großforschungslagen gegeben ist, an denen bereits ein hohes Maß an Standardisierung aller Abläufe gegeben ist und viele Daten digital verfügbar sind.

Qualitätsgesicherte Erfassung von Proben

Die Erstellung, Charakterisierung und Beschreibung von Proben ist für alle empirisch arbeitenden Disziplinen ein wesentlicher Bestandteil der wissenschaftlichen Arbeit. Die Erforschung von Proben an Großforschungsanlagen bedingt diese Erfassung über eine Vielzahl von Disziplinen und ist eng mit den Messmodalitäten an den einzelnen Messplätzen verbunden. Beispielsweise werden in der Energieforschung neue Materialien für Hochleistungsbatterien entwickelt, modelliert und ihre Eignung für die Energiespeicherung vermessen. Hierfür werden Materialproben erstellt und unter anderem mittels elektronenmikroskopischer Verfahren charakterisiert. Proben werden bei derartigen Forschungsansätzen interdisziplinär betrachtet und durchlaufen oft mehrere experimentelle Stufen.

Es besteht das Risiko, dass Informationen zu Proben, die nicht gleich zu Beginn des Arbeitsablaufes richtig aufgezeichnet wurden, verloren gehen. Ebenso wichtig ist eine lückenlose Dokumentation der Arbeitsschritte, welche die Probe bereits durchlaufen hat. Ansätze zur Vereinheitlichung und disziplinübergreifenden Methodik werden in Organisationen wie der Research Data Alliance (RDA) oder CODATA entwickelt. Eine Übertragung der Ergebnisse in die experimentelle Praxis bleibt jedoch bisher Aufgabe der einzelnen Labore.

Daher ist das Ziel, eine einheitliche Methodik und universell einsetzbare Softwarewerkzeuge zur Beschreibung und systematischen Erfassung von Metadaten zu experimentellen Proben innerhalb der Helmholtz-Gemeinschaft unter Nutzbarmachung existierender Standards und Best Practices zu schaffen.

Metadaten zu Maschinen und Messplätzen an Großforschungsgeräten

Ein vollständiger Messprozess an Großforschungsanlagen schließt die Einbeziehung von Metadaten der eigentlichen Großgeräte mit ein. Damit besteht eine umfassende Charakterisierung aus Metadaten über Zustand und Nutzungsweise des Großgeräts, dem spezifischen Experimentaufbau an konkreten Messplätzen, Informationen zur Probe sowie den genutzten Detektoren und dem Input der Großgerätenutzer selbst. Dieser umfassende Blick auf Metadaten verlangt neben einer konsequent koordinierten Aufzeichnung aller Metadaten eine umfassende Fusion aller Metadaten. In diesem Prozess werden Messdaten zum Beispiel aus der Maschine selbst zu Metadaten einer Analyse von Experimentaldaten. Die eigentliche Wissensextraktion bezieht dann selbst Daten aus Modellrechnungen und Simulationen mit ein, die mit Metadaten annotiert werden müssen. Die Nachvollziehbarkeit und Reproduzierbarkeit sowie die Wiederverwendbarkeit so gewonnener Daten ist nur in der Gesamtschau aller verfügbarer Metadaten möglich und setzt dabei eine „holistische“ Metadatenstrategie voraus.

Metadaten zu Analyseprozessen

Die Nachvollziehbarkeit und Überprüfbarkeit von Analysen hängt neben der Verfügbarkeit von Metadaten über den Messprozess ebenfalls von der Verfügbarkeit von Metadaten über die genutzten digitalen Verfahren ab. Dazu zählen grundlegende Dinge wie die eingesetzte Software inklusive Versionierung und genauem Einsatz, aber auch eine vollständige Dokumentation des Analyseprozesses. Gerade im maschinellen Lernen ist zum Beispiel die Qualität eines Ergebnisses nur aus dem Wechselspiel aus Eingabedaten, deren Auswahl und Qualität sowie den Lernparametern und der Überprüfbarkeit des erreichten Lernerfolges nachvollziehbar. Eine vollständige Erfassung des Analyseprozesses in all seinen

Facetten (auch im Bereich Provenienz) muss in einem digital verfügbaren Workflow Niederschlag finden, der alle zur Reproduktion der Analyse notwendigen Schritte zur Nachnutzung verfügbar macht. Hierzu fehlen bislang entscheidend Werkzeuge und Verfahren, die dies auch unerfahrenen Nutzern erleichtern und im besten Fall nebenläufig zur Analyse automatisiert verlaufen.

Skalierbare Sammlung, Fusion, Klassifikation und Auffindbarkeit von Daten- und Metadaten

Die für Datenerfassung nach FAIR-Prinzipien notwendigen Verfahren müssen mit steigenden Raten und Volumina an Daten so skalieren, dass menschliche Einflussnahme und Nutzung möglich bleiben und durch skalierbare automatische Prozesse so gut es geht unterstützt werden. Die Entwicklung hinsichtlich der Skalierbarkeit setzt in zunehmendem Maße intelligente Verfahren mit äußerst geringer Fehleranfälligkeit voraus. Die Entwicklung solcher Verfahren ist eine zukünftige Herausforderung, um auch einzelnen Wissenschaftlern und Wissenschaftlerinnen den sinnvollen Umgang mit großen Datenmenge zu ermöglichen.

5 Organisationsstruktur

Die Organisationsstruktur der HMC Plattform orientiert sich an der oben geschilderten Aufgabenverteilung und soll dabei die Integration der Metadaten-Kompetenz aus der gesamten Gemeinschaft in optimaler Weise ermöglichen. Die den Forschungsbereichen zugeordneten Metadata Hubs bilden zusammen mit dem HMC Office (mit seinen zentralen und dezentralen Leistungen) die HMC Plattform, die auch die dynamische Inkubatoren-Projekte beherbergt. Neben der Koordination und zentralen Repräsentanz ist das HMC Office für die Umsetzung der übergreifenden Aufgaben verantwortlich. Durch die Einrichtung der Metadata Hubs kann gesichert werden, dass die Expertise der Forschungsbereiche schon während der Umsetzung der HMC Plattform optimal eingebunden wird.

Die Metadata Hubs bilden das Bindeglied zwischen allen Zentren und den übergeordneten, an die gesamte Helmholtz-Gemeinschaft gerichteten Aktivitäten der HMC Plattform.

HMC Office

Als Plattform, die der gesamten Helmholtz-Gemeinschaft dient und deren Sichtbarkeit nach außen erhöht, stellt das HMC ein Office, das zentraler Ansprechpartner ist, Aktivitäten und Informationsfluss koordiniert, und für Personal- und Ressourcenmanagement sowie den Betrieb zuständig ist. Zentrale Kommunikation mit internen und externen Partnern sowie transparente Berichterstattung gegenüber dem Helmholtz-Inkubator ermöglichen ein effizientes und planbares Arbeiten. Damit etabliert sich das HMC als Vernetzungs-, Kommunikations- und Dienstleistungszentrum und Ansprechpartner für Helmholtz in Sachen Meta- und Forschungsdatenmanagement.

Das zentrale HMC Office erzielt Synergieeffekte, indem eine Verdopplung von Tätigkeiten vermieden wird, wo sich Aufgaben unabhängig von der jeweiligen Anwendungsdomäne angehen lassen. Insbesondere bei der Entwicklung von technologischen Werkzeugen zum Harvesting von Repositorien und beim Aufsetzen von technologischen Workflows zur automatisierten Integration von Metadaten sollen Software-Werkzeuge gemeinsam entwickelt und für alle Metadata Hubs nutzbar gemacht werden. Eine weitere zentrale Aufgabe ist das Screening neuer Technologien und die Sicherstellung des Informationstransfers aus internationalen, allgemeinen Entwicklungen in die fachspezifische Umsetzung in die Metadata Hubs hinein.

Leistungen des HMC Offices und seiner dezentralen Komponenten

- Koordination des HMC, Kommunikation mit Metadata Hubs, Berichterstattung an die zuständigen Gremien, zentraler Anlaufpunkt für Metadaten in der Helmholtz-Gemeinschaft

- Entwicklung dezentraler Dienste und Werkzeuge sowie Koordination bei der Entwicklung und beim Betrieb dezentraler Dienste
- Qualitätsmanagement und Koordination der dynamischen IVF Projekte
- Koordination der Zusammenarbeit mit den anderen Plattformen des Helmholtz-Inkubators

Metadata Hubs

Die in Forschungsbereiche zusammengefassten Zentren sind die wichtigsten Kompetenzträger für domänenspezifisches Metadatenwissen. Sie kennen ihren Community-spezifischen Stand, formulieren Bedarfe und Anforderungen. In den Metadata Hubs wird daher die domänenspezifische Kompetenz für die einzelnen Forschungsbereiche aggregiert und damit in die Arbeit der Plattform integriert. Gleichzeitig sind die Hubs ein wichtiger Kanal, um Informationen aus übergreifenden HMC-Erkenntnissen und -Entwicklungen in die Communities und die Zentren zu spiegeln. Es sind verlässliche, dauerhafte Einrichtungen, die für den Austausch von Leistungen, Informationen und Beratung zwischen den anderen Einrichtungen der Plattform und den Zentren sorgen. Viele Aufgaben (z. B. die Sammlung von Vokabularien, die in den einzelnen Disziplinen genutzt werden, die Entwicklung von Ontologien und eines Metadatenmodells) können nur durch aktive Verknüpfung zwischen dem Fachwissen in den Zentren und der gesamten HMC Plattform über die Brücke der Hubs zum Erfolg geführt werden.

Leistungen der Metadata Hubs

Die Einrichtung von einem Metadata Hub pro Forschungsbereich verortet wichtige Arbeitsbereiche der HMC Plattform direkt in den Programmen und Domänen.

- Koordination, Kommunikation und Vermittlung zwischen den Forschenden in einem Forschungsbereich und Metadaten-Expertinnen und -Experten in den verschiedenen Arbeitsbereichen der Plattform. Ein domänenspezifischer Metadata Hub fungiert als Bindeglied zwischen den übergreifenden Angeboten von HMC und den Forschenden in den Zentren. Dazu gehört auch die Adaption von technischen Werkzeugen und Diensten auf die domänenspezifischen Besonderheiten.
- Vernetzung der domänenspezifischen Aktivitäten und Bedarfe im Bereich Metadaten (bezüglich größerer Forschungsvorhaben wie Exzellenzcluster, ESFRIs, Flagships, Versuchskampagnen, Labore, etc.). Hier ist vor allem Informationsaustausch und Bedarfsermittlung gefragt.
- Sammlung der Best Practices und der Expertise aus dem Forschungsbereich, um sie in die allgemeinen Empfehlungen auf dem Informationsportal von HMC und in internationale Aktivitäten (RDA, EOSC, GO FAIR, re3Data, fairsharing.org) einzubinden.

Mehrwerte

- Reduzierung des Arbeitsaufwands (Kostenreduzierung, Effizienzsteigerung, Motivationseffekte)
- Forschende werden in ihrer Arbeitsumgebung (Domäne) „abgeholt“ und durch Vermittlung der Metadata Hubs mit den passenden Werkzeugen und Lösungen versorgt.
- Etablierung der qualitativen Beschreibung von Forschungsdaten durch Metadaten innerhalb des Forschungsbereichs erlaubt einen besseren interdisziplinären Informations- und Innovationsaustausch innerhalb eines Forschungsbereichs und eine bessere Vernetzung.
- Consulting: Schulung auf verschiedenen Ebenen etabliert Metadatenmanagement zielgruppengerecht in den Alltag der Forschenden.
- Im Metadata Hub werden die bisher verteilten Expertisen zu Metadaten innerhalb eines Forschungsbereichs gebündelt und vermittelt, so dass alle Programme und wissenschaftliche Projekte auf die Erfahrungen zugreifen können.

- Arbeitsabläufe werden innerhalb des Forschungsbereichs angepasst und optimiert. Z. B. geben Metadaten darüber Auskunft, ob ähnliche Messungen, Experimente, etc. an anderer Stelle bereits durchgeführt wurden. Redundanzen können vermieden werden.
- Synergieeffekte durch Vernetzung
- Optimierung der Kommunikationswege: Anbindung der Plattform an die Forschungsbereiche verkürzt die inhaltlichen und persönlichen Wege und integriert gleichzeitig bereits existierende Strukturen.
- Infrastrukturdienste und Werkzeuge können zentral entwickelt und dezentral genutzt beziehungsweise im Metadata Hub auf die Bedarfe des Forschungsbereichs angepasst werden.
- Existierende Lösungen können innerhalb der Domäne und der Programme leichter ausgetauscht werden (Aufgabe Koordination).
- Helmholtzweite Aktivitäten, wie z. B. das Helmholtz Open Science Koordinationsbüro finden Ansprechpartner mit direkter Anbindung an die Akteure im einzelnen Zentrum (Multiplikationseffekte).
- Austausch zwischen internationalen Aktivitäten und den Bemühungen innerhalb der Domänen wird erleichtert: Metadaten-Domänen-Expertinnen und -Experten können einschätzen, welche Lösungen für die eigene Arbeit im Forschungsbereich übernommen werden können und welche selbst entwickelten Lösungen lohnenswert für eine internationale Verbreitung erscheinen.
- Aufbau und Förderung der Aktivitäten im Bereich Forschungsdatenmanagement innerhalb der Zentren wird mit Expertise und Ergebnissen zum Thema Metadaten unterstützt.

Die Rolle des HMC im Helmholtz-Inkubator

Als Plattform, die der gesamten Helmholtz-Gemeinschaft dient und deren Sichtbarkeit nach außen erhöht, bestehen große Synergien mit den anderen Inkubator-Plattformen (siehe Tabelle 2).

Tabelle 2: Wechselbeziehungen zu anderen Inkubator Aktivitäten

	HMC liefert für andere Inkubator-Aktivitäten...	**HMC profitiert von anderen Inkubator-Aktivitäten...**
HIDA	Trainer, Schulungen, gemeinsame fachübergreifende Themen in der Aus- und Weiterbildung	Rekrutierung, Nachwuchs, gemeinsame fachübergreifende Themen in der Aus- und Weiterbildung
HIFIS	Interoperabilität der Metadaten (Schnittstellen, Prozesse, Werkzeuge und Dienste), Auffindbarkeit von Software durch Metadaten	Virtuelle Maschinen, Softwarerepositorien, Datenbanken, Authentifizierungs- und Autorisierungs- Infrastruktur, Backbone Services, Cloud Services
HIP	Interoperabilität der Metadaten, Maschinenlesbarkeit, Unterstützung beim Aufbau des „Metadata Portal for Imaging Data“	Zusammenarbeit Adoptionsprojekte, Analysemöglichkeiten
HAICU	Interoperabilität der Metadaten, Maschinenlesbarkeit, Unterstützung beim Aufbau vom Trainingsdaten-repositorium	Zusammenarbeit Adoptionsprojekte, Analysemöglichkeiten

6 Governance

Das HMC Office ist die zentrale Plattform der Gemeinschaft, die den strategischen und operativen Prozess der Entwicklung der Plattform umsetzt. Die Verortung und Entwicklung der Metadatenhubs

liegt in der Verantwortung der Forschungsbereiche. Die Ansiedlung der verschiedenen Aufgaben des HMC Office soll idealerweise im Konsens und kompetenzgetrieben zwischen den Forschungsbereichs-Hubs verteilt werden. Stellen die Hubs hier keinen Konsens her, soll die Geschäftsstelle ein wettbewerbliches Auswahlverfahren durchführen. Die nach dieser Verortung ermittelten finanziellen Bedarfe werden dem jeweiligen Helmholtz-Zentrum bedarfsorientiert zugeordnet. Das Personal, das über HMC eingestellt wird, untersteht damit der disziplinarischen Hoheit des jeweiligen Helmholtz-Zentrums. Im Rahmen der mit der Helmholtz-Gemeinschaft vereinbarten Ziele und Leistungen bezüglich HMC soll das jeweilige Helmholtz-Zentrum die Teil-Plattform operativ steuern.

Jedes Trägerzentrum benennt eine/n Standortkoordinator/in. Die Standortkoordinatorinnen und -koordinatoren führen die Beiträge der beteiligten Zentren zur HMC-Plattform zusammen. Sie bilden zusammen das HMC Collaboration Board und sind damit gemeinsam zuständig für die tägliche, operative Leitung. Das HMC Collaboration Board ist für fachliche Entscheidungen zuständig – im Rahmen der von den Vorständen der Helmholtz-Zentren, die das HMC tragen, eingeräumten Rechte und Pflichten. Das HMC Collaboration Board ernennt eine Sprecherin oder einen Sprecher und wird vom HMC-Office unterstützt.

Begleitet wird das HMC Collaboration Board von einem Lenkungskreis, der die Helmholtz-Gemeinschaft vertritt. Dieser kann Maßnahmen von seiner Zustimmung abhängig machen, hat Prüfungspflichten (insbesondere der aus den Arbeitspakten abzuleitenden Deliverables und Meilensteine für das Projektmanagement, grober Zeitplan s. Anhang 1) sowie Berichtspflichten an den Helmholtz-Inkubator. Der Helmholtz-Inkubator berichtet wiederum an die Mitgliederversammlung. Der Lenkungskreis bereitet die Entscheidung über die Vergabe der dynamischen Mittel vor und wird dazu vom wissenschaftlichen Beirat beraten. Der Präsident der Helmholtz-Gemeinschaft trifft die Förderentscheidung im Einklang mit den Regularien des Impuls- und Vernetzungsfonds. Die Besetzung des Lenkungskreises wird von der Mitgliederversammlung beschlossen. Der Lenkungskreis soll sich aus Mitgliedern des Helmholtz-Inkubators zusammensetzen und möglichst alle Forschungsbereiche repräsentieren. Der Sprecher des HMC Collaboration Boards und die Geschäftsstelle sollen an den Sitzungen des Lenkungskreises als Gäste teilnehmen. Der Lenkungskreis tagt mindestens einmal jährlich.

Der wissenschaftliche Beirat soll sich aus fünf bis sieben HGF-externen Expertinnen und Experten zusammensetzen, beispielsweise gewonnen aus internationalen Organisationen und Projekten aus den Bereichen Forschungsdatenmanagement und Metadaten. Die Mitglieder des wissenschaftlichen Beirats sollen vom Präsidenten der Helmholtz-Gemeinschaft bestimmt werden. Der Beirat sollte mit Blick auf Kontinuität besetzt werden und wird regelmäßig informiert, um allgemein beraten zu können. Der wissenschaftliche Beirat steuert die Ausschreibungen und Begutachtungen nach einem Kriterienkatalog (siehe unten) und tritt in der Regel halbjährlich zusammen. Idealerweise vertreten Mitglieder des Beirats Expertise aus dem NFDI Prozess sowie aus internationalen Metadaten-Prozessen.

Vergabeprozess für dynamische Mittel aus dem Impuls- und Vernetzungsfonds

Die auszuschreibenden Projekte sollen eigene Ergebnisse erarbeiten, die über das HMC wiederum anderen zur Verfügung stehen und generell genutzt werden können.

Der Richtwert für die Projektdauer sind 12-24 Monate. Der Lenkungskreis kann einen anderen Zeitraum bewilligen, wenn dies fachlich begründet ist. Ausschreibungen sollen zweimal jährlich stattfinden. Es wird angestrebt, zwischen zwei und vier Projekte gleichzeitig laufen zu lassen. Dazu sollen die Ausschreibungen kontinuierlich fortgesetzt werden, weil gerade durch die Projekte große Effekte auf die Vermittlung der innerhalb von HMC entwickelten Lösungen in die gesamte Gemeinschaft erwartet werden.

Das HMC Office entwickelt in Absprache mit dem Lenkungskreis und dem Impuls- und Vernetzungsfonds (IVF) einen Bewertungsbogen mit wenigen, gewichteten Kriterien, um die Begutachtung der Projektanträge zu vereinfachen. Diese Kriterien sollen sich auch an den Vorschlägen der Expert Group der EU-Kommission zu FAIR Data (Hodson u. a. 2018), sowie den sich im Rahmen der Umsetzung der NFDI ergebenden Richtlinien orientieren. Die endgültige Entscheidung über die Zuweisung von Mitteln zu den Projekten liegt beim Präsidenten der Helmholtz-Gemeinschaft.

7 Finanzplan

Tabelle 3: Finanzplanung für die HMC Platform

Plattformanteil	**Kosten p.a. in k€**
Arbeitsbereich	
Kostenart	
HMC Office	4 FTE
Koordination, Management und Controlling	
Sachkosten	30
Personalkosten	180
Wissen und Vermittlung	
Sachkosten	5
Personalkosten	90
Komponenten und Prozesse	
Sachkosten	5
Personalkosten	90
Travel	15
Workshops	15
Summe	**430**
Dezentrale Dienste – FAIR technisch	8 FTE
Übergreifende Technische Dienste und Werkzeuge	
Sachkosten	20
Personalkosten	360
Technische Realisierung der FAIR-Prinzipien	
Sachkosten	15
Personalkosten	360
Travel	25
Workshops	20
Summe	**800**
5 x Metadata Hub (1 x pro FB, lokal) (hier ohne DLR)	25 FTE
Koordination und Management	
Sachkosten	25
Personalkosten	450
Wissen und Vermittlung	

	Sachkosten	50
	Personalkosten	900
	Komponenten und Prozesse	
	Sachkosten	50
	Personalkosten	900
	Travel	75
	Workshops	50
Summe		**2,500**
Projekte (dynamisch nach Ausschreibung)		
	IVF-finanzierte Projekte (zuzüglich 50% Eigenfinanzierung aus sich beteiligenden Zentren)	
Summe		**1,200**
Summe pro Jahr		**4,930**
	davon:	
	Sachkosten (grundfinanziert)	400
	Personalkosten (grundfinanziert)	3,330
	Projektmittel (IVF finanziert)	1,200

Anmerkungen zu **Fehler! Verweisquelle konnte nicht gefunden werden.**:

- Die grundfinanzierten, statischen Mittel werden für den Aufbau und Betrieb der HMC Plattform verwendet mit dem zentralen HMC Office sowie den dezentralen Komponenten des HMC Office und den Metadata Hubs in den Forschungsbereichen.
- Zusammengefasst sind die geplanten Kosten der Plattformteile mit den Arbeitsbereichen.
- Basis für die Berechnung bilden die Personalmittelsätze der DFG für das Jahr 2019 plus jeweils ein Overheadsatz von 25 % auf Personal in den statischen Komponenten (im Mittel € 90.000).
- Sachkosten setzen sich zusammen aus:
    - Personalbezogene Sachkosten von ca. 10.000 € pro Jahr und pro FTE beinhalten Verbrauchsmittel, Büroausstattung, wissenschaftliche Hilfskräfte und Reisemittel.
    - Sachkosten im HMC Office von 30.000 € beinhalten PR-Materialien und andere Drucksachen sowie insgesamt Mittel für Workshops.
- Die Finanzierung des Metadata Hubs für DLR ist hier nicht im Gesamtvolumen berücksichtigt.

Referenzen

Deutsche Forschungsgemeinschaft. 2009. „Empfehlungen zur gesicherten Aufbewahrung und Bereitstellung digitaler Forschungsprimärdaten“. Deutsche Forschungsgemeinschaft. https://www.dfg.de/foerderung/programme/infrastruktur/lis/veroeffentlichungen/index.html.

———. 2015. „Leitlinien zum Umgang mit Forschungsdaten“. DFG. http://www.dfg.de/download/pdf/foerderung/antragstellung/forschungsdaten/richtlinien_forschungsdaten.pdf.

European Commission. 2016. „Guidelines on FAIR Data Management in Horizon 2020“. European Commission. http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf.

———. 2017. *Commission Implementing Decision (EU) 2017/1358 of 20 July 2017 on the Identification of ICT Technical Specifications for Referencing in Public Procurement (Text with EEA Relevance.). EU 2017/1358*. Bd. 190. http://data.europa.eu/eli/dec_impl/2017/1358/oj/eng.

Franke, Michael, Stefan Heinzel, Reiner Mauer, Janna Neumann, Heike Neuroth, Hans Pfeiffenberger, Henriette Senst, u. a. 2015. „Positionspapier ‚Research data at your fingertips‘ der Arbeitsgruppe Forschungsdaten“. https://doi.org/10.2312/allianzfd.001.

Helmholtz-Gemeinschaft. 2016. „Positionspapier ‚Die Ressource Information besser nutzbar machen!‘“ https://www.helmholtz.de/os-positionspapier/.

Hodson, Simon, Sandra Collins, Sarah Jones, Genova Francoise, Natalie Harrower, Leif Laaksonen, Daniel Mietchen, Rūta Petrauskaité, und Peter Wittenburg. 2018. „Turning FAIR into Reality“. Brüssel.

Moore, Reagan, und Rainer Stotzka. 2015. „Practical Policy Recommendations“. RDA. 26. März 2015. https://www.rd-alliance.org/group/practical-policy-wg/outcomes/practical-policy.

re3data.org – Registry of Research Data Repositories. 2015. „re3data.org Metadata Schema 3.0 XML Schema“. https://doi.org/10.2312/re3.009.

Wedlich, Doris, Lars Bernhard, Frank Oliver Glöckner, Gregor Hagedorn, Hans-Josef Linkens, Otto Rienhoff, Petra Gehring, und Norbert Lossau. 2017. „Entwicklung von Forschungsdateninfrastrukturen im internationalen Vergleich“. http://www.rfii.de/de/dokumente/.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, u. a. 2016. „The FAIR Guiding Principles for Scientific Data Management and Stewardship“. *Scientific Data* 3 (März): 160018. https://doi.org/10.1038/sdata.2016.18.

Anhang 1: Detailübersicht der Arbeitspakete

	Arbeitsplan Helmholtz Metadata Collaboration Platform										
			1 - 2 a	3 - 5 a	10 a	Jahr 1	Jahr 2	Jahr 3	Jahr 4	Jahr 5	Verknüpfungen mit
1	Zentrales HMC Office, zentrale dauerhafte Leistungen	430 T€ p.a. / 4 FTE + 70 TsD	4,0 FTE + 70 TsD								
1.1	Aufbau und Koordination der Geschäftsstelle		1,5 FTE								AP 3.1
1.1.1	Aufstellen von und Abstimmen mit wissenschaftlichem Beirat	durchgehend	0.1	0.1	0.1	Initialisierung und Detailabstimmung der Programme		Organisation der Beiratstreffen und Einspeisung der Empfehlungen in HMC Programme			AP 3.1.1
1.1.2	Kommunikation und PR	durchgehend	0.3	0.3	0.3	Konzeption und Aufsetzen der Kampagnen		Kontinuierlich Umsetzung und Weiterentwicklung der Kampagnen			AP 2-3
1.1.3	Zusammenarbeit mit anderen Aktivitäten des Helmholtz Inkubators	durchgehend	0.1	0.1	0.1	Identifikation der Schnittstellen und gemeisame Planung		Kontinuierliche Zusammenarbeit weiterentwickeln			AP 2-3
1.1.4	Internationale Vernetzung, Gremienarbeit, Harmonisierung und Standardisierung	2 Jahre, dann durchgehend mehr	0.2	0.4	0.4	Gremien besetzen	HMC-spez. Einfluss	kont. Interessenvertretung / aktive Abstimmung mit internationalen Akteuren			AP 3.1.3
1.1.5	Aufbau und Betreuung einer „Metadaten-Community" in der Helmholtz-Gemeinschaft	durchgehend	0.3	0.3	0.3	Formate für Information und Austausch etablieren		Regelmäßiges Networking über F2F- und Online-Formate			AP 3.1.2
1.1.6	Projektbüro - Ausschreibung, Überwachung der IVF Ausschreibung und Qualitätssicherung	durchgehend	0.5	0.5	0.5	Entwicklung 1. Ausschreibung, Etablierung der Prozesse		Zyklische Ausschreibung und Betreuung			AP 4, AP 3.1, AP 2
1.2	Wissen & Vermittlung		2,0 FTE			Aufbau und erstes Angebot von Vermittlungsformaten		Vermittlungsformate stehen Online und F2F kontinuierlich und aktuell zur Verfügung			AP 3.2.3
1.2.1	Kartierung der Forschungsdateneexpertisen	2 Jahre, dann durchgehend weniger	0.5	0.3	0.3	Koordination der Kartierung	Verknüpfung aller Hubs	Nutzbarkeit für AP 2 und andere Hubs ist gegeben, kontinuierliche Aktualisierung			AP 3.2.2
1.2.2	Methodenwissen aufbauen und verfügbar machen	durchgehend	0.5	0.5	0.5	Sammeln und aufbereitet bereitstellen		Kontinuierliche Aktualisierung der Informationen und Anpassung der Angebote			AP 3.2.1, AP 3.2.3
1.2.3	Basisinformationen und Schulungsinhalte – Wissen bezüglich Metadaten verbreiten	durchgehend	0.3	0.3	0.3	Entwicklung der Vermittlungsformate		Zyklisches Angebot der domänenspez. Vermittlungsformate ausrollen und betreuen			AP 3.2.1-3
1.2.4	Vermittlung durch Expertinnen und Experten	2 Jahre, dann durchgehend mehr	0.2	0.4	0.4	Identifikation und Vernetzung Multiplikatoren		Pool von Expertinnen kann vermittelt werden			AP 3.3.4
1.2.5	Persönliche Beratung	durchgehend	0.5	0.5	0.5	Beratungsformate entwickeln und erproben		Beratungsangebot (aktiv und passiv) kontinuierlich anbieten und weiterentwickeln			AP 3.3.4
1.3	Komponenten & Prozesse		0,5 FTE								AP 3.3, AP 2
1.3.1	(Meta-)Daten haben eine eindeutige Lizenz	durchgehend	0.1	0.1	0.1	Entwurf einer Richtlinie nach Evaluation best. Lizenzmodelle		Nutzung der Richtlinie betreuen und weiterentwickeln			AP 3.1
1.3.2	FAIRe Vokabularien, Ontologien, Minimalstandards	durchgehend	0.3	0.3	0.3	Anforderungskatalog mit den HUBs entwickeln		Kontinuierliche Bereitstellung erarbeiteter Lösungen			AP 3.3.5-6, AP 2.1.2, AP 2.2.1
1.3.3	Empfehlungen zur Annotierung von Forschungsdaten mit umfangreichen Metadaten, u.a. Provenienz	durchgehend	0.1	0.1	0.1	Empfehlungen auf Basis praktischer Beispiele entwickelt		Monitoring der Nutzbarkeit der Empfehlungen sowie Weiterentwicklung			AP 3.1, AP 2
2	Dezentrale Komponenten des HMC Office: FAIR technisch ermöglichen (statische Finanzierung)	800 T€ p.a. / 8 FTE	8 FTE								
2.1	Übergreifende technische Dienste und Werkzeuge		6 FTE	4 FTE	4 FTE						
2.1.1	Implementierung FAIR Data Object – Schnittstellen zu EOSC und NFDI	2 Jahre, dann durchgehend weniger	3	2	2	Implementierung des Fair Data Object Model		Weiterentwicklung in Koop. EOSC, RDA, NFDI; Integration der Lsg aus Inkubatorprojekten			AP 1.2, AP 3.1.3, AP 4
2.1.2	Entwicklung und Betrieb technischer Dienste und Werkzeuge	2 Jahre, dann durchgehend weniger	3	2	2	Dienste zur Verwendung der Kataloge/Registries		Weiterentwicklung, Support, Kuratierung; Integration der Lsg aus Inkubatorprojekten			AP 1.2, AP 3.1.1, AP 4
2.2	Komponenten und Prozesse – Realisierung der FAIR-Prinzipien		2 FTE	4 FTE	4 FTE						
2.2.1	Findable: Forschungsdaten mittels Identifier und umfassenden Metadaten auffindbar machen	2 Jahre, dann durchgehend mehr	0.5	1	1	Angebot für PID-Dienste, Repositorien existiert		Ausbau der Menge der erfassten Metadaten, Zusammenarbeit mit Metadata Hubs			AP 3.3, AP 1.2
2.2.2	Accessible: offene und standardisierte Protokolle und Policies	2 Jahre, dann durchgehend mehr	0.5	1	1	Protokolle und Schnittstellen stehen zur Verfügung		Policies werde erarbeitet, Repositorien werden mehr und mehr mit Exitstrategie betrieben			AP 3.3.4, AP 3.1
2.2.3	Interoperable: FAIRe Wissensrepräsentation durch offene Vokabulare, Ontologien und Standards	2 Jahre, dann durchgehend mehr	0.5	1	1	Werkzeuge, Prozesse zur Erzeugung, Verwaltung, Interaktion		Unterstützung bei der Anbindung an existierende und neue Ontologien			AP 1.3.2, AP 3.2.1, AP 3.3
2.2.4	Re-Usable: Meta- / daten, ihre Provenienz und Nutzungslizenz sind detailliert beschrieben	2 Jahre, dann durchgehend mehr	0.5	1	1	Policy Registry, Werkzeuge zur Provenienz von Meta (Daten)		Weiterentwicklung in Zusammenarbeit mit Metadata Hubs			AP, 1.3.5, AP 3.3.6,
3	Domänenspezifische Leistung – Metadaten Hubs	2.500 T€ p.a. pro Hub 5 FTE + 50 TsD.	pro Hub 5 FTE plus 50 Tsd								
		Ausprägung pro Hub variable!	Ausprägung/Hub variable								
3.1	Koordination und Management		1 FTE			Kommunikation und Maßnahmen etablieren		Fortlaufende Koordination u. Management der Maßnahmen und Kommunikationslinien			AP 1.1
3.1.1	Einbeziehung der Community-Expertise	2 Jahre, dann durchgehend weniger	0.4	0.2	0.2	Etablierung Integrationsmechanismen		Kontinuierliche Entwicklungsabstimmung durch Gremien, Projekte, ...			AP 1.2, AP 2.2
3.1.2	Aufbau und Betreuung einer „Metadaten-Community" im jeweiligen Forschungsbereich	durchgehend	0.4	0.4	0.4	Formate für Information und Austausch etablieren		Regelmäßiges Networking über F2F- und Online-Formate			AP 3.1.5
3.1.3	Internationale Vernetzung, Gremienarbeit, Harmonisierung und Standardisierung	2 Jahre, dann durchgehend mehr	0.2	0.4	0.4	Gremien besetzen	Domänenspez. Einfluss	kont. Interessenvertretung / aktive Abstimmung mit internationalen Akteuren			AP 3.1.4
3.2	Wissen & Vermittlung		2 FTE			Aufbau domänenspez. Vermittlungsformate		Vermittlungsformate stehen Online und F2F kontinuierlich und aktuell zur Verfügung			AP 1.2
3.2.1	Aufbau und Bereitstellung einer Informationsbasis zu Metadaten) vokabularien), Ontologien und Standards	2 Jahre, dann durchgehend weniger	0.5	0.25	0.25	Sammeln und aufbereitet bereitstellen		Kontinuierliche Aktualisierung der Informationen und Anpassung der Angebote			AP 1.3.2, AP 2.2.3
3.2.2	Landschaft der Forschungsdateneexpertisen	2 Jahre, dann durchgehend weniger	0.5	0.25	0.25	Aufbau der Kartierung	Verknüpfung mit Hubs	Nutzbarkeit für AP 2 und andere Hubs ist gegeben, kontinuierliche Aktualisierung			AP 1.1.5, AP 1.2.1, AP 1.2.4
3.2.3	Domänenspezifische Ergänzungen zu Methoden, Basisinformationen und Schulungsinhalten	2 Jahre, dann durchgehend weniger	0.5	0.5	0.5	Ergänzung der Vermittlungsformate um Domänenwissen		Zyklisches Angebot der domänenspez. Vermittlungsformate ausrollen und betreuen			AP 1.2.2-3
3.2.4	Beratung von Zentren, Projekten und Forschenden	2 Jahre, dann durchgehend mehr	0.5	1	1	Beratungsformate entwickeln und erproben		Beratungsangebot (aktiv und passiv) kontinuierlich anbieten und weiterentwickeln			AP 1.2.4-5
3.3	Komponenten & Prozesse		2 FTE								AP 3.1-2
3.3.1	Prozesse, Werkzeuge und Dienste zur Erschließung von Forschungsdatensammlungen aufbauen	2 Jahre, dann durchgehend mehr	0.5	0.75	0.75	Erhebung und Beschreibung	Anforderungsanalyse	Nutzbarmachung, Anpassung, Weiterentwicklung			AP 1.2.1-2, AP 2.2.1-4
3.3.2	Ingest: Werkzeuge zur Automatisierung der Erfassung von Metadaten	durchgehend	0.25	0.25	0.25	Erhebung und Evaluation	Anwendung auf zwei Bsp.	Domänenspez. Anpassungen und Anwendung auf weitere Bsp			AP 1.3.3, AP 2.1.2, AP 2.2.1
3.3.3	Werkzeuge zur Handhabung verschiedener Metadatenstandards	2 Jahre, dann durchgehend weniger	0.25	0.1	0.1	Konzeption	Umsetzung	Kontinuierliche Anpassung			AP 2.2.1-4
3.3.4	Zugang zu Forschungsdaten über Schnittstellen schaffen	durchgehend	0.2	0.2	0.2	Erhebung und Evaluation	Anwendung auf zwei Bsp.	Domänenspez. Anpassungen und Anwendung auf weitere Bsp			AP 2.2.1-4
3.3.5	Erfassen der FAIRness der Daten	2 Jahre, dann durchgehend weniger	0.3	0.2	0.2	Werkzeugentwicklung	Anwendung auf > 2 Bsp.	FAIRness der Daten kontinuierlich erfassbar, kontinuierliche Aktualisierung			AP 1.3.4, AP 2
3.3.6	Provenienz	durchgehend	0.5	0.5	0.5	Konzeption und prototypische Anwendung		Weiterentwicklung und kontinuierliche Ausweitung der Anwendung			AP 1.3.3, AP 2.2
4	Inkubatorprojekte, Ausschreibung koordiniert vom HMC Office (dynamisch), 50% Eigenleistung	1.200 T€ alt 1.200 T€									AP 1.1.6

Größere Version verfügbar in separater Datei

Anhang 2: Umfeldanalyse (intern und extern)

Research Data Alliance (RDA) wurde 2012 mit dem Ziel gegründet, Forschungsdaten gemeinsam zu nutzen und die sozialen und technischen Voraussetzungen dafür zu schaffen. Mittlerweile ist RDA eine weltumspannende Organisation mit ca. 7000 Datenexpertinnen und -experten, die in Arbeitsgruppen Empfehlungen für den einheitlichen Umgang mit Forschungsdaten erarbeiten und so eine bessere Vernetzung und Wiederverwendung der Daten zu ermöglichen. Einige davon werden Ausgangspunkt für Empfehlungen des HMC sein und Arbeiten aus dem HMC werden in Gremien der RDA eingebracht, um internationale Anschlussfähigkeit zu erreichen und zu behalten. In einigen Bereichen können Helmholtz-Wissenschaftlerinnen und -Wissenschaftler hier durch erfolgreiche Gremienarbeit eine führende Rolle in der internationalen Standardisierung und Verbreitung von Leitlinien zum Forschungsdatenmanagement einnehmen.

European Open Science Cloud (EOSC) hat das Ziel, eine offene Plattform zum Austausch von Forschungsdaten aufzubauen, um Forscherinnen und Forscher in Europa zu verbinden. Dabei spielen Metadaten eine wichtige Rolle. Empfehlungen und Werkzeuge, die vom HMC erarbeitet werden und z. B. in RDA Arbeitsgruppen diskutiert werden, könnten mittelbar in der EUDAT CDI in Lösungen zum Data Access & Re-Use, Verarbeitung oder Analyse einfließen. Technologien und Dienste, die im Rahmen der EOSC bereitgestellt werden, sollen soweit möglich vom HMC genutzt werden. Nutzer der HMC Dienste und Werkzeuge sollen möglichst einfach EOSC Dienste, wie z. B. Speicher (B2SHARE) oder bestehende Suchfunktionen (B2Find), nutzen können.

Cross Continental Collection Access and Management Pilot (C2CAMP) ist eine Initiative internationaler Experten und Expertinnen aus fünf Kontinenten, die gemeinsam ein flexibles und erweiterbares Testbed für digitale Datenobjekte basierend auf der Global Digital Object Cloud beziehungsweise FAIR Data Objects aufbaut. So sollen technische Entwicklungen für EOSC vorangetrieben und auf eine breitere Basis gestellt werden. In Europa entspricht C2CAMP einem Implementation Network der GO FAIR Initiative.

Novel Materials Discovery (NOMAD) Laboratory stellt für die größte Datensammlung innerhalb der Materialwissenschaften nicht nur Speicher für Daten und Metadaten zur Verfügung, sondern auch Werkzeuge, um diese online zu analysieren. Für das HMC können einige Faktoren übernommen werden, die in dieser Plattform gut umgesetzt sind. Die Suche nach Datensätzen anhand von strukturierten Metadaten mit Vokabularien (u. a. dem Periodensystem) kann z. B. ein Vorbild sein. Aber auch die Ansätze für Big Data Analysen sind richtungsweisend, setzen aber zwingend ein gutes Metadatenmanagement voraus, um valide Ergebnisse zu produzieren. Diese Vorarbeiten können durch das HMC ermöglicht oder verein- facht werden.

re3data ist ein globales Verzeichnis von Forschungsdaten-Repositorien und Portalen über alle wissenschaftlichen Disziplinen. Ursprünglich gefördert von der DFG und entwickelt unter Beteiligung zweier Helmholtz-Zentren, wird der Betrieb und die Weiterentwicklung seit 2016 von DataCite koordiniert. Die strukturierte Beschreibung der Repositorien und Portale enthält auch technische Informationen zu Schnittstellen und Metadaten Standards, welche als Teil der Bestandsaufnahme der Helmholtz Datensammlungen in das HMC fließen.

Das Helmholtz Open Science Koordinationsbüro (HOSK) vernetzt und unterstützt seit 2005 Wissenschaftlerinnen und Wissenschaftler sowie deren Helmholtz-Zentren bei der Umsetzung von Open Science mit besonderem Augenmerk auf dem Umgang mit Forschungsdaten. Das HMC profitiert von den existierenden Netzwerken und wird mit dem HOSK zusammenarbeiten.

Erfolgreiche Projekte, die Metadaten umfangreich nutzen wie **DARIAH** (Digital Research Infrastructure for the Arts and Humanities) oder **CLARIN** (Common Language Resources and Technology Infrastructure), werden dem HMC als Inspiration dienen. Hieraus entstandene technische Konzepte oder Werkzeuge (z. B. für Vokabularien) werden evaluiert und gegebenenfalls übernommen, aber auch Vermittlungskonzepte oder organisatorische Best Practices aufgegriffen, wenn sich diese bewährt haben.

Anhang 3: Verzeichnis der verwendeten Abkürzungen und Eigennamen

AAI	Authentication and Authorization Infrastructure
AARC	Authentication and Authorisation for Research and Collaboration, https://aarc-project.eu/
AG	Arbeitsgruppe
AK	Arbeitskreis
AP	Arbeitspaket
API	Application programming interface
ARDC	Australian research data commons, https://ardc.edu.au/
B2Find	Discovery service angeboten von EUDAT, https://www.eudat.eu/services/b2find
B2SHARE	Datenspeicherservice angeboten von EUDAT, https://b2share.eudat.eu/
Bioportal	Ein Repositorium von biomedizinischen Ontologien, https://bioportal.bioontology.org/
C2CAMP	Cross Continental Collection Access and Management Pilot, https://github.com/c2camp/core/wiki
CLARIN	European Research Infrastructure for Language Resources and Technology, https://www.clarin.eu/
CODATA	Committee on Data of the International Council for Science, http://www.codata.org/
DARIAH	Digital Research Infrastructure for the Arts and Humanities, https://www.dariah.eu/
DataCite	bietet Persistent identifiers (DOIs) für Forschungsdaten, https://www.datacite.org/
DFG	Deutsche Forschungsgemeinschaft, http://www.dfg.de/
DLR	Deutsches Zentrum für Luft- und Raumfahrt

DOI	Digital Object Identifier
EOSC	European Open Science Cloud, https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud
ELIXIR	Internationale Organisation, die Life Science Ressourcen aus ganz Europa verbindet. https://elixir-europe.org
Enabling FAIR	Enabling FAIR data project, http://www.copdess.org/enabling-fair-data-project/
ESFRI	European Strategy Forum on Research Infrastructures, http://www.esfri.eu/
EUDAT CDI	Collaborative Data Infrastructure, https://www.eudat.eu/eudat-collaborative-data-infrastructure-cdi
ExPaNDS	EOSC Photon and Neutron Data Services
FAIR	Leitlinien für Findable, Accessible, Interoperable, Re-Usable Datenpublizieren, https://www.force11.org/fairprinciples
FAIRSFAIR	Europäisches Projekt, das Regeln zur Teilnahme an EOSC aufstellt
fairsharing.org	Resource zu Daten- und Metadatenstandards, https://fairsharing.org/
FREYA	Horizon 2020 Projekt mit Fokus auf Open Identifiers, https://www.project-freya.eu
FTE	Full-Time Equivalent
GO FAIR	GO FAIR, internationaler Ansatz für die praktische Implementierung des EOSC, https://www.go-fair.org/
H2020 FET	Horizon 2020 Future And Emerging Technologies, https://ec.europa.eu/programmes/horizon2020/en/h2020-section/future-and-emerging- technologies
HAICU	Helmholtz Artificial Intelligence Cooperation Unit
HGF	Helmholtz Gemeinschaft, https://www.helmholtz.de
HIDA	Helmholtz Information & Data Science Academy

HIFIS	Helmholtz Infrastructure for Federated ICT Services
HIP	Helmholtz Imaging Platform
HMC	Helmholtz Metadata Center
HOSK	Helmholtz Open Science Koordinationsbüro, http://os.helmholtz.de/
ICSU-WDS	International Council for Science's World Data System, https://www.icsu-wds.org/
ICT	Information and Communications Technology
IETF	Internet Engineering Task Force, https://www.ietf.org/
ISO	International Organization for Standardization, https://www.iso.org
IVF	Impuls- und Vernetzungsfonds der Helmholtz-Gemeinschaft, https://www.helmholtz.de/ueber_uns/die_gemeinschaft/impuls_und_vernetzungsfonds/
NFDI	Nationale Forschungsdateninfrastruktur, http://www.rfii.de/de/themen/
NOMAD	Novel Materials Discovery Laboratory, https://nomad-repository.eu/
OAI	Open Archives Initiative, https://www.openarchives.org/
PID	Persistent Identifier
RFII	Rat für Informationsinfrastrukturen, http://www.rfii.de/
RDA	Research Data Alliance, https://www.rd-alliance.org/
RDM	Research data management
re3data	Registry of Research Data Repositories, https://www.re3data.org/
SOP	Standard operating procedure
W3C	World Wide Web Consortium, https://www.w3.org/