HELMHOLTZ RESEARCH FOR GRAND CHALLENGES

HMC REPORT | 2 | Data infrastructures

Die Dateninfrastruktur (DIS)-Erhebung im Forschungsbereich Erde und Umwelt der Helmholtz-Gemeinschaft

Erste Ergebnisse und Schlussfolgerungen

December 2022

HMC | HELMHOLTZ METADATA COLLABORATION

Keywords: FAIR, HMC Hub Earth and Environment, Data Infrastructures, Helmholtz, Survey

DOI: 10.3289/HMC_publ_06

Citation: Helmholtz Metadata Collaboration, Hub Earth and Environment; Die Dateninfrastruktur (DIS)-Erhebung im Forschungsbereich Erde und Umwelt der Helmholtz-Gemeinschaft: Erste Ergebnisse und Schlussfolgerungen, 2022.

Acknowledgement
This publication was supported by the Helmholtz Metadata Collaboration (HMC), an incubator-platform of the Helmholtz Association within the framework of the Information and Data Science strategic initiative.

Version: 1.0

Authors (ORCID)**:** Emanuel Söding (0000-0002-4467-642X), Andrea Pörsch (0000-0003-4502-6223), Pier Luigi Buttigieg (0000-0002-4366-3088), Helen Kollai (0000-0003-0214-1336), Martin Weinelt (0000-0003-3896-0261), Sören Lorenz (0000-0001-8577-6614)

HMC group: Hub Earth and Environment, https://www.helmholtz-metadata.de/earthandenvironment

Contact: HMC Office
GEOMAR Helmholtz Centre for Ocean Research Kiel
Wischhofstr. 1-3
24148 Kiel, GERMANY
E-mail: info@helmholtz-metadaten.de

www.helmholtz-metadaten.de

Inhalt

Zusammenfassung

Die Anschlussfähigkeit bestehender Helmholtz-Dateninfrastrukturen an nationale und internationale Initiativen (z.B. NFDI, EOSC und andere), ist ein wichtiges zentren-übergreifendes Ziel der Helmholtz-Gemeinschaft. Dafür ist es entscheidend, die bestehenden Zusammenhänge der Forschungsdaten-erhebenden und -nachnutzenden sowie die Dateninfrastrukturen (DIS) entlang der FAIR-Prinzipien[1] zu ordnen. Ziel dieser Erhebung ist es, einen Überblick über bestehende Praktiken und Entwicklungs-stände der Dateninfrastrukturen (DIS) in der Helmholtz Gemeinschaft im Bereich Erde und Umwelt zu erhalten, um eine konsistente Strategie zur Umsetzung eines FAIRen Datenraumes entwickeln zu können.

Es stellt sich heraus, dass, den DIS die FAIR-Prinzipien bekannt sind, sie sich mehrheitlich dazu bekennen und bestrebt sind, diese bestmöglich umzusetzen. Dabei wählen sie zum Teil sehr unterschiedliche Umsetzungsstrategien und treffen auf unterschiedliche Herausforderungen und Hindernisse. Dabei wird deutlich, dass die Schwierigkeiten in der Umsetzung mit der Tiefe und organisatorischen Komplexität bei der Implementierung der FAIR-Prinzipien zunehmen. Während beispielsweise einige Kriterien zur Auffindbarkeit (Findability) und Zugriffsmöglichkeit (Accessability) bereits bei vielen DIS z.B. durch den Einsatz von DOIs und Metadatenstandards erfüllt sind, ist es offensichtlich schwieriger, die Interoperabilität (Interoperabilty) und Nachnutzbarkeit (Reusability) der Daten umzusetzen.

Der HMC Hub Erde und Umwelt (EuU) kann aus der Erhebung für den Forschungsbereich EuU eine Reihe von Empfehlungen ableiten. Dazu gehört unter anderem die Notwendigkeit eines einheitlichen Verständnisses von FAIR. Innerhalb des Forschungsdatenmanagements (FDM) sollte der Fokus insbesondere auf eine einheitliche Umsetzung persistenter Identifier (PIDs) / Handles, FAIR Digital Objekts (FDOs), semantischer Konzepte und Provenienz gerichtet werden, weil deren Einsatz wichtige Bausteine für die Umsetzung der FAIR-Prinzipien sind. Infolgedessen könnte die Rolle von HMC sein, den einheitlichen Umgang mit Forschungsdaten innerhalb der Helmholtz-Gemeinschaft zu moderieren und zu koordinieren. Dadurch würden Prozesse zu etabliert die es erlauben, dass alle erhobenen Daten in den verschiedenen Forschungsbereichen gefunden, verwendet und einheitlich referenziert werden können.

Langfristig gilt es einen Community Co-Design Prozess zu etablieren, der es erlaubt, möglichst konkrete, gemeinschaftlich getragene Vereinbarungen zum Datenaustausch zu treffen und diese in Helmholtz Empfehlungen zum Umgang mit Forschungsdaten, also Policies, umzusetzen. Dadurch würde ein für alle nachvollziehbarer, interoperabler Helmholtz Datenraum geschaffen.

[1] Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). DOI:10.1038/sdata.2016.18.

Summary

The connectivity of existing Helmholtz data infrastructures to national and international initiatives (e.g. NFDI, EOSC and others), is an important goal across all centres of the Helmholtz Association. To achieve this, it is critical to organize the existing interrelationships of research data-collection, research data-reuse, and data infrastructures (DIS) along the FAIR principles. The goal of this survey is to obtain an overview of existing practices and the development status of data infrastructures (DIS) in the Earth and Environment domain of the Helmholtz Association in order to develop a consistent strategy to implement a FAIR data space.

It turns out that, the DIS are aware of the FAIR principles, the majority of them are committed to them and strive to implement them in the best possible way. In doing so, they sometimes choose very different implementation strategies and encounter different challenges and obstacles. It is clear that the difficulties in implementation increase with the depth and organizational complexity of implementing the FAIR Principles. For example, while some findability and accessibility criteria are already met in many DIS, e.g., through the use of DOIs and metadata standards, it is obviously more difficult to implement a thorough interoperability and reusability of data without extensive alignment with partners.

The HMC Earth and Environment (EuU) Hub can derive a number of recommendations from the survey for the EuU research area. Among these is the need for a unified understanding of FAIR. Within research data management (RDM), particular focus should be placed on a uniform implementation of persistent identifiers (PIDs) / handles, FAIR digital objects (FDOs), semantic concepts, and provenance because their use are important building blocks for the implementation of FAIR principles. As a result, the role of HMC could be to moderate and coordinate the uniform handling of research data within the Helmholtz Association. This would establish processes that allow all collected data to be found, used, and referenced consistently across the different research areas.

In the long run, a community co-design process has to be established, which allows to reach concrete, jointly supported agreements on data exchange and to implement them in Helmholtz recommendations on the handling of research data, i.e. policies. This practice would create an interoperable Helmholtz data space that is comprehensible to all.

1 Einleitung

Die Anschlussfähigkeit bestehender Helmholtz-Dateninfrastrukturen untereinander, wie auch die Anknüpfung an nationale und internationale Initiativen (z.B. NFDI, EOSC und andere), ist ein wichtiges zentren-übergreifendes Ziel der Helmholtz-Gemeinschaft. Dafür ist es entscheidend, die Landschaft der Forschungsdatenerzeuger, -nachnutzer und Dateninfrastrukturen (DIS) entlang der FAIR Prinzipien zu ordnen, Vereinbarungen zum Datenaustausch zu treffen und diese für alle nachvollziehbar und umsetzbar zu dokumentieren.

Der HMC Hub Erde und Umwelt (E&U) ist daher bestrebt, einen FAIRen Helmholtz Datenraum (Helmholtz FAIR Data Space, HFDS) zu definieren, zu schaffen und zu aktivieren. Dieser soll als eine "dezentrale Infrastruktur für die vertrauenswürdige gemeinsame Nutzung und den Austausch von Daten in Datenökosystemen auf der Grundlage gemeinsam vereinbarter Prinzipien"[2] dargestellt werden. Innerhalb der HMC E&E besteht der Datenraum aus gemeinsamen Vereinbarungen zu Verfahren und Prozessen in der Dokumentation von Forschungsdaten. Die Vereinbarungen beziehen sich sowohl auf die Nutzung standardisierter Metadaten, z.B. in PID Systemen, als auch auf die Nutzung abgestimmter Semantiken. Die Umsetzung dieser Vereinbarungen werden zu einer Interoperabilität der Daten innerhalb des Datenraumes führen.

Durch den Datenraum wird es auch möglich werden, einen einheitlichen Umgang mit Forschungsdaten innerhalb der Helmholtz-Gemeinschaft zu etablieren, sodass diese Daten von Forschern gefunden, verwendet und referenziert werden können. Dazu ist es nötig, Forschungsdaten mit beschreibenden Informationen, also Metadaten, zu versehen. Metadaten sind – in diesem Kontext - essentielle, standardisierte Informationen über Forschungsdaten und ermöglichen deren Auffinden sowie deren Vernetzung und Nachnutzung.

Die qualitative Anreicherung von Forschungsdaten durch Metadaten muss daher vorangetrieben sowie organisatorisch und technisch umgesetzt werden. Dazu führt HMC die wissenschaftliche Expertise zum Thema Metadaten aus den Fachdomänen in den Metadaten-Hubs der Forschungsbereiche zusammen und zeigt die Bedeutung von Metadaten innerhalb des Forschungsdatenmanagements auf. Um das weitere Vorgehen planen zu können, ist es wichtig, Informationen zum Entwicklungsstand der Dateninfrastrukturen (DIS) zu bekommen. Welche Aspekte werden bereits berücksichtigt und umgesetzt? Welche Techniken sind implementiert? Sind (Meta-) Daten maschinenlesbar? Wie ist deren Qualität?

Um zu diesen Fragen Hinweise zu bekommen, wurden vom Hub Erde und Umwelt der Helmholtz Metadata Collaboration (HMC) viele Informationen aus öffentlich zugänglichen Quellen über unterschiedliche Dateninfrastrukturen der Helmholtz-Gemeinschaft zusammengetragen (siehe z.B. https://helmholtz-metadaten.de/en/earth-and-environment/Helpful_Information).

Die im Folgenden vorgestellte Status- und Bedarfserhebung dient dem Ziel, einen detaillierten Einblick in die existierenden Komponenten und Datensammlungen, jenseits der öffentlich verfügbaren Informationen zu erhalten, um daraus Bedarfe und weitergehende Aktivitäten ableiten zu können.

Dabei sind unter Dateninfrastrukturen in diesem Kontext sowohl Softwarekomponenten als auch Dienste gemeint, welche dafür geeignet sind, wissenschaftliche Daten und Metadaten, Software oder Simulationen von Experimenten und Messungen zu speichern, zu beschreiben, zu transportieren oder zu verarbeiten. Das können Repositorien, Datenbanken, Visualisierungsanwendungen, Data Discovery Portale, Katalogdienste oder Vergleichbares sein.

[2] Nagel L., Lycklama D. (2021): Design Principles for Data Spaces. Position Paper. Version 1.0. Berlin, DOI: 10.5281/zenodo.5105744

2 Ziele der Erhebung

Die Erhebung verfolgt das Ziel, Informationen zu den folgenden Themenbereichen zusammenzustellen, insbesondere um:

1. einen Überblick darüber zu erhalten,
    a. welche DIS in der Helmholtz-Gemeinschaft genutzt werden.
    b. wie das Selbstverständnis der jeweiligen DIS gestaltet ist, d.h. ob die FAIR-Prinzipien bekannt sind und FAIR ein angestrebtes Ziel ist und ob Kooperationen und Austausch mit anderen DIS einen hohen Stellenwert haben.
    c. welche inhaltlichen Kategorien existieren, d.h. welche Zielgruppen angesprochen werden, welche thematischen oder methodischen Fokusse existieren und welche Dienste angeboten werden.
    d. wie nachhaltig der Betrieb und die Laufendhaltung der DIS sind. Dazu gehören Informationen dazu, welche Finanzierungsmodelle bestehen, wer fördert, der Zeitrahmen der Finanzierung, welche Perspektiven gesehen und angestrebt werden.

2. folgende Fragestellungen zu technisch implementierten Aspekten zu klären:
    a. Welche Metadatenstandards, PID Systeme, APIs, Werkzeuge, Viewer werden angeboten bzw. dokumentiert?
    b. Ist die DIS zertifiziert, und wird sie gegenüber einer verabredeten Metrik gemessen?
    c. Wie sehen die jeweiligen Prozesse zur Erhebung und Kontrolle aus, wer betreut diese?
    d. Werden Empfehlungen der nationalen und internationalen FDM Gemeinschaft berücksichtigt?
    e. Wie ist das Qualitätsmanagement gestaltet?

3. die DIS-Koordinations- oder Unterstützungsbedarfe zu ergründen:
    a. Welche Ziele verfolgen die DIS selbst und wie gehen sie damit um?
    b. Welche Herausforderungen werden in technischer und struktureller Hinsicht gesehen?
    c. In welchen Bereichen sehen die DIS-Verantwortlichen Handlungsbedarf, welcher z.B. von HMC abdeckt werden könnte.

Mit Hilfe dieser Informationen lassen sich möglicherweise Hinweise auf folgende Fragen ableiten:

1. Wie ist die Anschlussfähigkeit der DIS im Sinne der FAIR Prinzipien einzuschätzen?
2. Gibt es Aspekte, die von einigen DIS bereits umgesetzt wurden und können diese als Umsetzungsbeispiele für andere DIS dienen?
3. Kann ein Dialog initiiert werden, um für die geäußerten Bedarfe gemeinsam Lösungen zu entwickeln?
4. Welche Handlungsfelder ergeben sich daraus für ein modernes Forschungsdatenmanagement und welche Rolle kann HMC dabei spielen.

3 Rahmenbedingungen und Einschränkungen

Für diese Umfrage wurden die administrativ und technisch informierten Kontakte der Infrastrukturen angesprochen, um möglichst genaue Angaben zum Aufbau und Zielen der DIS zu erhalten. Die Teilnahme an der Erhebung war freiwillig. Infolgedessen haben nicht alle angeschriebenen DIS die Umfrage beantwortet. Die Erhebung soll ausdrücklich nicht dazu dienen, einzelne DIS zu evaluieren, d.h. eine vergleichende Auswertung über DIS wurde nicht durchgeführt. Die Erhebung wurde anonymisiert durchgeführt, d.h. es ist nicht ersichtlich, welche Person die Informationen für eine DIS eingegeben hat. Es werden im Nachgang keine Datensätze veröffentlicht, die auf eine bestimmte DIS rückschließen lassen. Die Publikation der Erhebung erfolgt anonymisiert und ggf. aggregiert. Das Umfrageergebnis soll vielmehr den befragten DIS helfen, durch die Veröffentlichung der Auswertungsstatistiken ihren eigenen Stand einschätzen zu können und auf diese Weise den Austausch mit anderen DIS anregen.

Einige der Infrastrukturen, welche Daten aus Projekten oder von verschiedenen Communities visuell aufbereitet in Portalen zur Verfügung stellen, erwecken von außen betrachtet den Eindruck, als wären sie eigenständige Infrastrukturen. Sie basieren aber auf derselben technischen Installation bzw. besitzen sie dieselbe Datenquelle. Um die Ergebnisse in der Statistik möglichst unverfälscht zu erhalten, sollte pro technisch distinkter Dateninfrastruktur jeweils nur **ein** Datensatz ausgefüllt werden.

Die Erhebung enthielt bewusst keine obligatorisch auszufüllenden Fragen, um den Befragten jederzeit die Möglichkeit zu geben, Angaben zu bestimmten Themen abzulehnen. Daher ist es möglich, dass sich die Summe der gegebenen Antworten nicht immer mit der Summe der Teilnehmer (37) deckt. Es war an vielen Stellen auch möglich anzugeben, dass die Antwort „nicht zutreffend“ oder „nicht bekannt“ sei.

Da es sich bei der Erfassung um eine Selbsteinschätzung der DIS handeln sollte, wurden alle Eintragungen unverändert ausgewertet, wie sie erfasst wurden, d.h. die Angaben wurden weder überprüft, noch auf Plausibilität untersucht oder korrigiert. Bei einigen Antworten wird deutlich, dass Fragen möglicherweise missverstanden wurden oder die Assoziation, die mit der Fragestellung verbunden war, unterschiedlich ausgelegt wurde. Ein Beispiel dafür ist das Verständnis von Provenienz, das von einigen Ausfüllenden als Versionierung verstanden wurde, von anderen wiederum als Abstammung der Daten. Diese unterschiedlichen Interpretationen geben Einblick darin, welche Themen in den DIS präsenter sind als andere.

4 Themenblöcke der erhobenen Informationen

In der Erhebung wurden Informationen zu den folgenden Themenblöcken erhoben:

1. Allgemeine Angaben zur Infrastruktur
2. Angaben zum Selbstverständnis
3. Angaben zum Kurationsprozess
4. Angaben zu verwendeten Technologien und Praktiken
5. Angaben zu Zugriffsmöglichkeiten und Schnittstellen auf die DIS
6. Angaben zu Nachnutzung der Dateninfrastruktur und ihrer Daten
7. Angaben zu möglichen Bedarfen und Verbesserungen der Infrastruktur

5 Ergebnisse

In Abstimmung mit den Forschungsdatenmanagement-Verantwortlichen der verschiedenen EuU-Zentren wurden 57 DIS-Kontakte angeschrieben. Davon haben 37 den Fragebogen vollständig ausgefüllt. Die Antworten in den einzelnen Bereichen ergeben folgendes Bild:

Allgemeine Angaben zur Infrastruktur

In diesem Themenblock wurden allgemeine Informationen über die DIS erhoben. Dazu gehörten Bezeichnung, URL und Trägerinstitution. Diese Informationen dienten der Identifizierung der DIS um Doppelungen erkennen zu können.

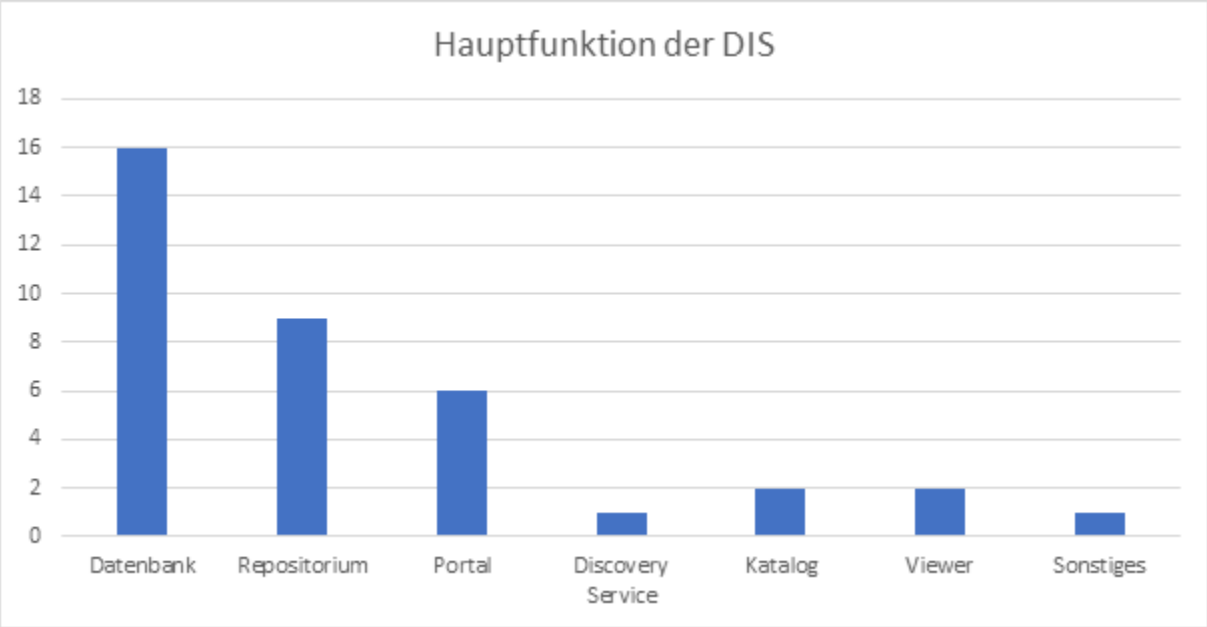

Abb. 1: *Welche Hauptfunktion oder Typ kommt - Ihrem Verständnis nach - Ihrer Infrastruktur am nächsten.*

Insgesamt wird als Hauptfunktion der DIS überwiegend ‚Datenbank' (16 Nennungen) angegeben, gefolgt von ‚Repositorium' (9) und ‚Portal' (6), siehe Abb. 1.

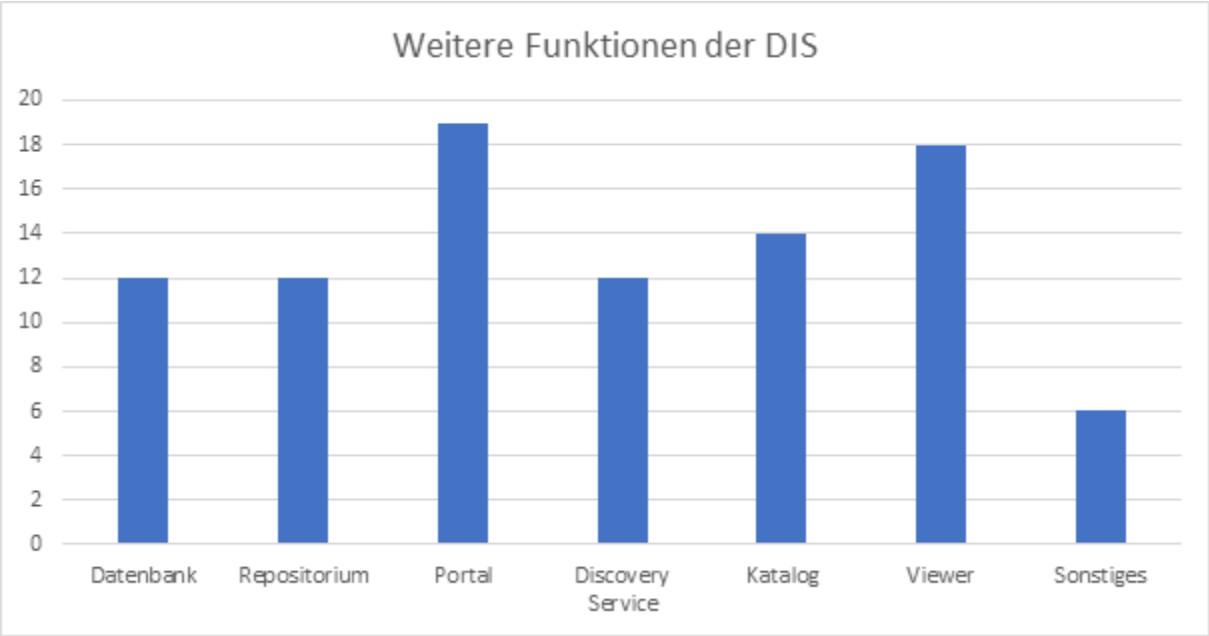

Abb. 2: *Erfüllt Ihre Infrastruktur weitere Funktionen? (Mehrfachauswahl möglich)*

Berücksichtigt man auch Nennungen weiterer Funktionen der DIS zeigte sich das Bild wie in Abb. 2 dargestellt mit einem breiten Anwendungsspektrum.

Die einzuordnenden Kategorien wurden in der Erhebung nicht gesondert erläutert. Es handelt sich daher um eine Selbsteinschätzung der DIS auf Basis dessen, wie die Kategorien aus ihrer Sicht gesehen werden. Die Streuung bei der Auswahl der Kategorien deutet darauf hin, dass innerhalb einer DIS verschiedene Funktionen in Nutzung sind.

Zur Speicherung der Daten gaben 31 der Befragten an, dass die Informationen in ihren Systemen gespeichert werden. Zwei Systeme enthalten Verweise auf externe Datenquellen, ein System dient nur der Präsentation von Forschungsdaten. Die Speicherung und Versionierung erfolgen in diesem Fall an anderer Stelle.

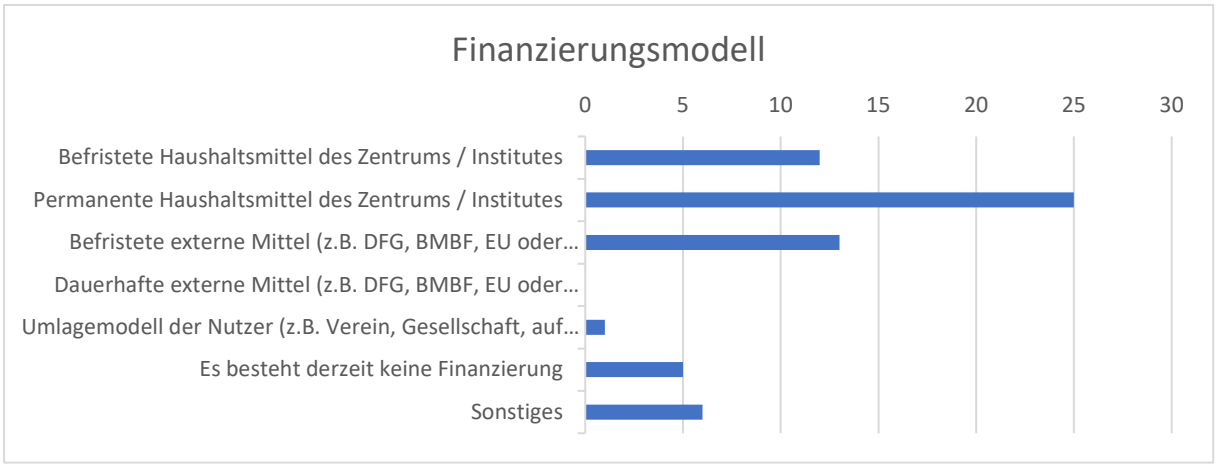

Abb. 3: *Finanzierungsmodelle der DIS (Mehrfachnennungen möglich)*

Die Finanzierungsmodelle der DIS sind vielfältig. Der Betrieb speist sich bei etwa der Hälfte der DIS aus mehreren Quellen und sowohl befristeten, wie auch unbefristeten Mitteln (Abb 3). Dauerhafte Zuwendungen aus Haushalten der jeweiligen Zentren erhalten 25 der DIS. Nahezu alle DIS werden als frei verfügbarer Service für ihre jeweiligen Communities betrieben (siehe Abschnitt 2). Nur ein Service nutzt ein Umlagemodell der Nutzer, um Haushaltsmittel zu ergänzen; ein anderer erhält freiwillige Zuwendungen von Nutzern. Sechs der befragten DIS arbeiten derzeit ohne Finanzierung. Nicht immer

sind Finanzierungsmodelle für eine DIS formal abgesichert. Manche Infrastrukturen werden aus dem laufenden Etat anderer Zentrumseinrichtungen betrieben.

Insgesamt ergibt sich für den langfristigen Betrieb und die Laufendhaltung der DIS ein gemischtes Bild. Während einige DIS verbindlich Ressourcen aus Zentren und Fachcommunities erhalten, beruht der Betrieb anderer DIS weitgehend auf persönlichem Engagement, eingeworbenen Mitteln und freiwilligen Beiträgen.

Selbstverständnis und inhaltliche Ausrichtung

In diesem Block wurden Angaben zum Selbstverständnis, z.B. FAIRness der Metadaten sowie Zielgruppen und der methodische bzw. fachliche Fokus abgefragt.

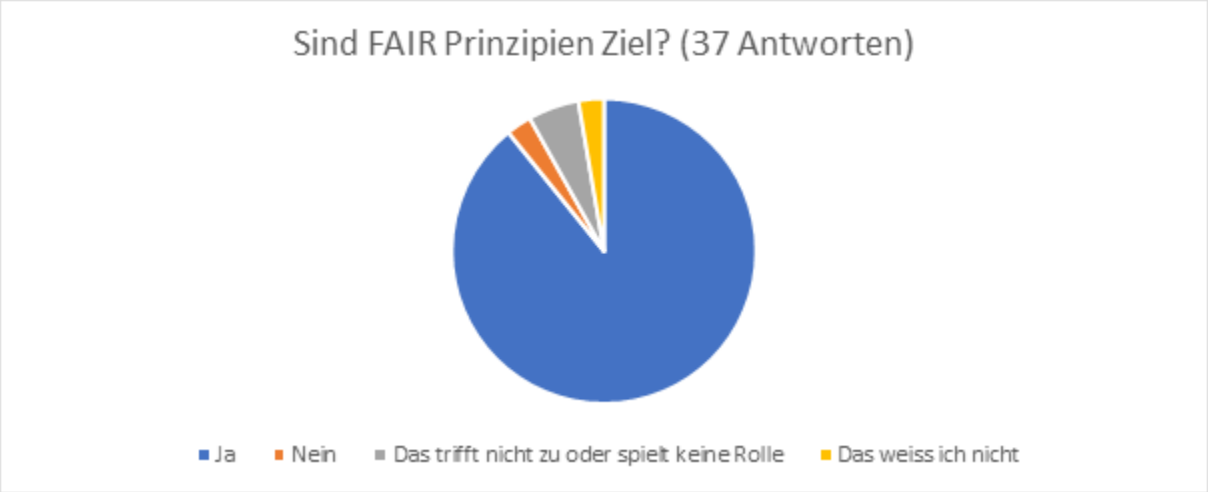

Abb. 4: *Hat Ihre Infrastruktur das Ziel, Daten nach den FAIR-Prinzipien zu behandeln? (Optionsfelder)*

Insgesamt ist das Bekenntnis zu den **FAIR-Prinzipien** sehr groß, siehe Abb. 4. Rund neunzig Prozent der DIS (33) geben die Antwort „Ja“ auf die Frage, ob ein Ziel die Umsetzung der FAIR-Prinzipien in ihrer Infrastruktur ist. In der Selbsteinschätzung, inwieweit die DIS auf diesem Ziel fortgeschritten sind, antworten mehr als die Hälfte (21) der Befragten, dass in ihrer DIS bereits mindestens 70 Prozent der FAIR-Prinzipien umgesetzt sind.

Die Arten von Informationen, die in den DIS vorgehalten werden, sind sehr divers. Die meisten DIS stellen mehrere **Informationsarten** bereit. Es werden vorwiegend gemessene Daten gespeichert,

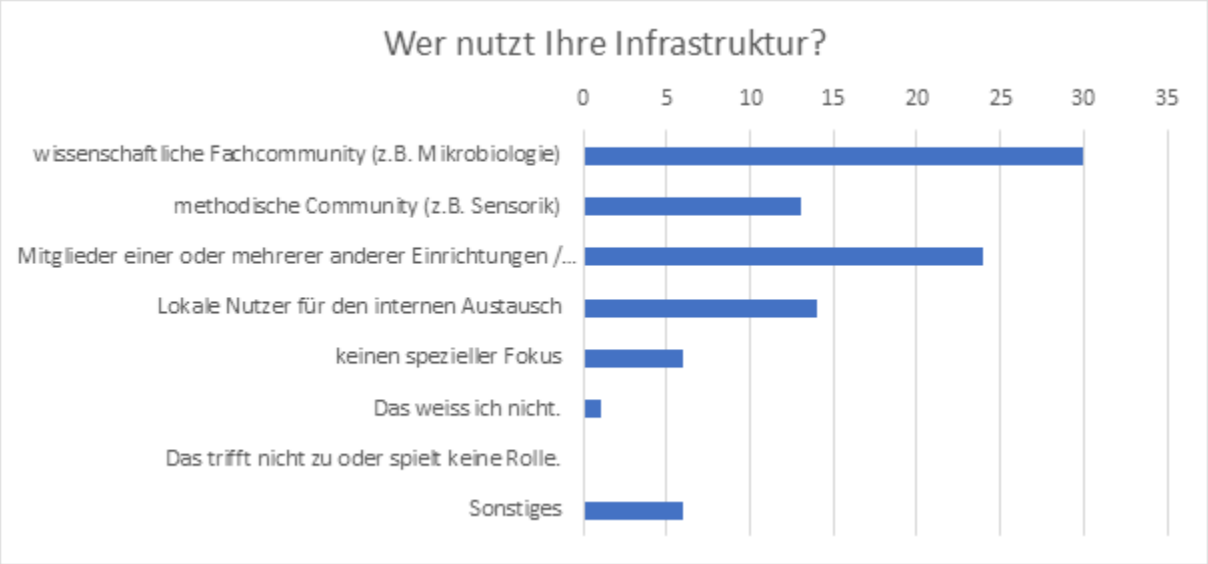

Abb. 5: *Wer nutzt Ihre Infrastruktur? (Mehrfachauswahl möglich)*

ebenfalls Daten aus Modellen und Simulationen. Acht der 37 DIS speichern auch wissenschaftliche Software. Gut 90 Prozent der Befragten geben an, dass neben den Metadaten auch die Daten der beschriebenen Ressource selbst innerhalb der DIS gespeichert werden.

Die **inhaltliche Ausrichtung** der meisten DIS orientiert sich primär an einer bestimmten wissenschaftlichen Fachcommunity (30), gefolgt vom Austausch von Datensätzen mit Mitarbeitenden anderer Zentren oder Einrichtungen (24). Vierzehn dienen dem zentrumsinternen Datenaustausch, dreizehn der DIS haben einen **methodischen** Fokus, siehe Abb. 5.

Als **Zielgruppen** für die DIS werden grundsätzlich Wissenschaffende bzw. Forschende adressiert. Viele DIS richten sich auch an Behörden und eine interessierte Öffentlichkeit, allerdings an nachrangiger Stelle. Kommerzielle Interessen und Entscheidungsträger werden nur selten ausgewählt. Innerhalb einer DIS wird als wichtigste Zielgruppe Schüler in Fokus genommen. Insgesamt werden als **Nutzer** vorwiegend die wissenschaftlichen Fachcommunities genannt.

Erfassung, Kuration und Qualitätssicherung

Dieser Abschnitt dient zur Erschließung von Angaben zur Datenerfassung, zum Kurationsprozess und zur Qualitätssicherung. Dazu gehört auch die Berücksichtigung von Umsetzungsempfehlungen von Forschungsdaten-Interessengruppen, die Nutzung von Datenmanagementplänen (DMPs) und ggf. Informationen zu einer Zertifizierung der Infrastruktur durch entsprechende Stellen.

Die **Erfassung der Daten und Metadaten** erfolgt überwiegend durch Wissenschaffende (30) und durch eigene Kuratoren (20), weiterhin durch Kuratoren anderer Einrichtungen (8). Neunzehn DIS erfassen Daten und Metadaten (teilweise) automatisiert. Weitere genannte Erfassende sind Bürger (Citizen Science), studentische Hilfskräfte und die Bibliothek.

Spezifische **Workflows** zur Erfassung der Metadaten werden von 17 der DIS vorgegeben. Von diesen Workflows ist einer manuell, vier sind vollautomatisiert und zwölf teilweise automatisiert, indem z.B. bestimmte Metadaten bei der Erfassung strukturiert erhoben oder abgefragt werden.

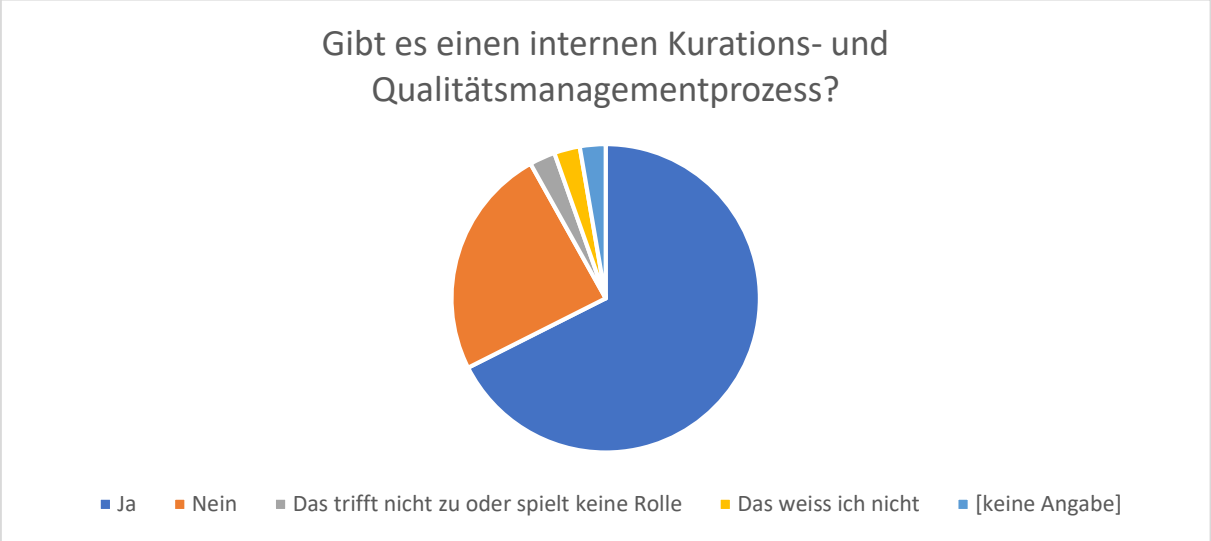

***Abb. 6**: Gibt es einen internen Kurations- und Qualitätsmanagementprozess? (Optionsfelder)*

Es zeigt sich, dass die meisten der DIS (25) einen **internen Kurationsprozess** implementiert haben. Rund ein Viertel der DIS (9) haben keinen solchen Prozess implementiert, siehe Abb. 6. Bei den DIS ohne Kurationsprozess werden die Daten überwiegend, aber nicht ausschließlich, automatisch erfasst.

Der **Umgang mit Fehlern** in Datensätzen und Metadaten ist unterschiedlich. Während einmal publizierte Datensätze nur in Ausnahmefällen (1) korrigiert werden, werden falsche oder fehlende Metadaten überwiegend nachgefordert und korrigiert (30), wo möglich, in Absprache mit dem jeweiligen Metadatenkontakt (16 von 30).

Definierte **Qualitätskriterien** an die Metadaten werden von 20 DIS angelegt. Davon verlangen fast alle DIS Mindestanforderungen an die Metadaten. Eine DIS bietet eine freiwillige Überprüfung der Metadatenqualität an. Die Qualität wird z.T. automatisch, z.B. über XML Schemata überprüft (6). Bei mindestens 5 DIS wird die Vollständigkeit durch die Kuratoren überprüft. Die Metadatenqualität dient im Wesentlichen der internen Kontrolle und wird nur bei einer DIS mit den Daten veröffentlicht.

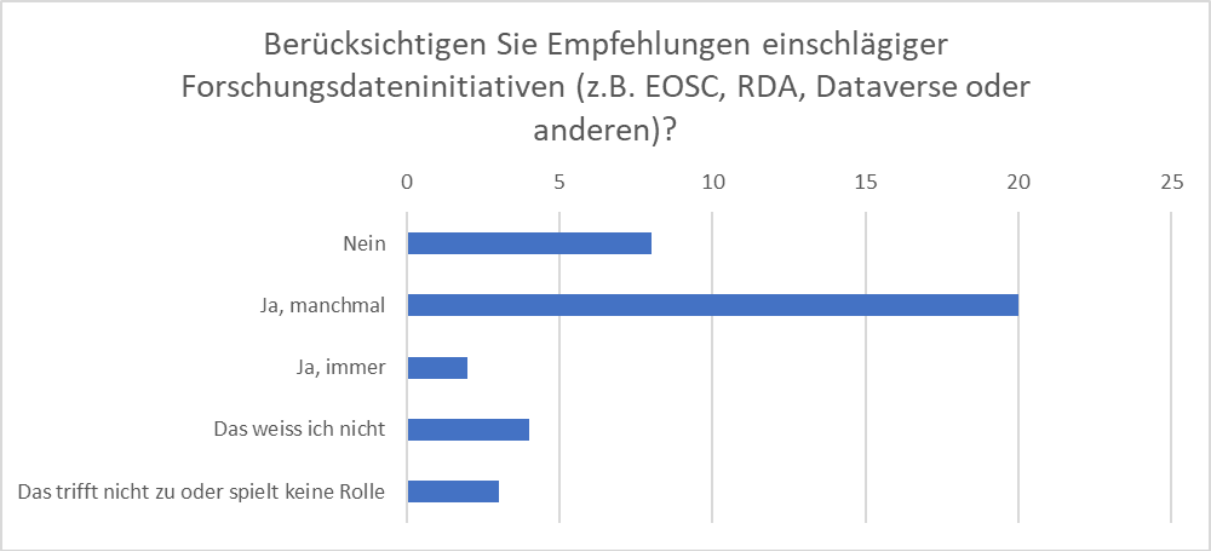

Abb. 7: *Berücksichtigen Sie Empfehlungen einschlägiger Forschungsdateninitiativen (z.B. EOSC, RDA, Dataverse oder anderen)? (Optionsfelder)*

Die **Empfehlungen einschlägiger Forschungsdaten-Interessensgruppen** werden von 22 DIS berücksichtigt, siehe Abb. 7. Die dabei überwiegend genannten Interessensgruppen sind die Research Data Alliance (RDA) (19) und die European Open Science Cloud (EOSC) (11). Alle übrigen Gruppen werden nur vereinzelt genannt. Der tatsächliche Mehrwert der Empfehlungen wird von neunzehn DIS als grundsätzlich nützlich bewertet. Die Abstufungen bewegen sich von „sehr oft“ (4), über „meistens“ (8) bis „manchmal“ (7) nützlich. Von 35 Befragten geben 11 für ihre DIS an, aktiv bei der Entwicklung von Empfehlungen beteiligt zu sein. Weitere 5 DIS beobachten die Entwicklungen ohne aktive Mitwirkung, um diese ggf. zügig umsetzen zu können.

Spezielle **Anreize zur Verbesserung der Datenqualität**, wie z.B. Preise oder Titel werden von 5 DIS angeboten. Einige DIS bieten bestimmte projektspezifische Mehrwerte, wie herausgehobene Datenpublikationen oder Kompilationen an. Eine DIS präsentiert „Featured Datasets“, Open Data Awards sind bei einer anderen DIS in Planung. Auf die Frage nach weiteren Verbesserungsvorschlägen werden vor allem verbindliche Vorgaben, die Nutzung von PIDs, eine automatische Erfassung, Schulung sowie Kuration - bis hin zur Empfehlung, Wissenschaftler zu involvieren - genannt.

Die **Erstellung von Datenmanagementplänen (DMP)** wird von vielen DIS (29) empfohlen. RDMO[3] wird von zwei DIS empfohlen, OpenAIRE ARGOS[4] von einer DIS. 6 DIS habe eigene DMP Lösungen, von denen eine die Integration von Informationen aus dem DMP in die DIS durch eine automatisierte Metadatenerfassung vorsieht.

Von den befragten DIS ist eine (1) durch das Core Trust Seal (CTS)[5] zertifiziert. Weitere sechs befinden sich im Prozess der CTS **Zertifizierung** oder in Vorbereitung dazu. Einige, wenige DIS enthalten Daten / Services von DIS, die an anderer Stelle zertifiziert sind. Abgesehen von CTS werden keine anderen Zertifikate genannt.

[3] Research Data Management Organizer (RDMO): https://rdmorganiser.github.io/

[4] OpenAIRE ARGOS: https://argos.openaire.eu/

[5] Core Trust Seal: https://www.coretrustseal.org/

Technologien und Praktiken

Dieser Bereich enthält Angaben zu verwendeten Technologien und Praktiken, z.B. Metadatenstandards, Persistenten Identifikatoren, Provenienz, Datenformaten und digitalen Objekten.

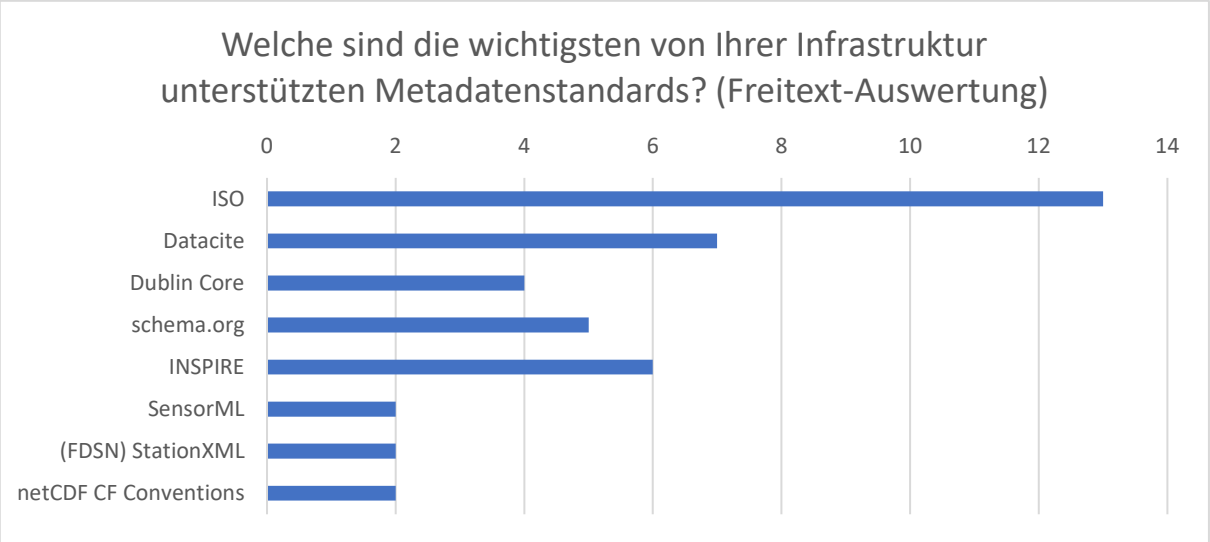

***Abb. 8:** Welche Metadatenstandards werden von den DIS unterstützt (Mehrfachnennungen möglich).*

Von 37 Antwortenden geben 22 der DIS an, international etablierte **Metadatenstandards** zu verwenden. Von den genannten Standards wird die Spezifikation ISO 19115 am häufigsten eingesetzt – gewertet durch 13 direkter Benennungen und 6 weiteren mit „INSPIRE Profil", welchem dieser ISO-Standard zugrunde liegt. Dieser wird gefolgt von DataCite (7), dem Dublin Core (4) Metadatenstandard und schema.org (5) (Abb 8). Weitere genannte Standards mit jeweils 2 Benennungen sind unter anderem FDSN Station XML (Seismik), NetCDF CF conventions, und SensorML.

Semantische Konzepte wie Vokabularien, Thesauri oder Ontologien sind bei 16 DIS implementiert. Die meisten davon unterstützen je nach Fachgebiet mehrere Vokabularien, z.B. NASA GCMD, NERC, SeadataNet. Nicht alle der befragten DIS haben die Art der Unterstützung konkret benannt, was darauf hindeuten könnte, dass ein semantisches Konzept beim Aufbau der DIS (noch) keine Rolle gespielt hat.

Die **Provenienz** der Daten wird von nur wenigen DIS systematisch erfasst. Zwei DIS erfassen nach dem PROV Standard, weitere 11 erfassen Provenienz mit eigenen Verfahren. Aus den eingetragenen Kommentaren könnte abgeleitet werden, dass von einigen Befragten eine Versionierung als Provenienz gewertet wird.

Persistente Identifikatoren (PIDs) werden von den meisten DIS verwendet. Von 37 DIS verwenden 26 PIDs, um **Datensätze** zu referenzieren. 22 davon nutzen DOIs[6]. Weitere genannte PIDs sind IGSN[7] (5, z.T. in Planung), EPIC[8] (2), HDL[9] (2), ORCID[10] (3) und andere. Persistente Identifikatoren, um Informationen in **Metadaten** zu referenzieren werden von 24 DIS verwendet, 20 verwenden DOIs, ORCID[10] (10), HDL (4), ROR[11] (2, z.T. in Planung).

[6] Digital Object Identifier: https://www.doi.org/

[7] International Geo Sample Number (IGSN): https://www.igsn.org/

[8] EPIC: https://www.pidconsortium.net/

[9] Handle Identifier: https://en.wikipedia.org/wiki/Handle_System

[10] Open Researcher and Contributor ID (ORCID): https://orcid.org/

[11] Research Organization Registry (ROR): https://ror.org/

Die meisten DIS sind offenbar auf bestimmte Datentypen zur Speicherung (Dateiformat) spezialisiert. Das könnte bei insgesamt 30 Antworten anhand der Häufigkeit (9) der ausgewählten Stückzahl „10“ für verschiedene Formate abgeleitet werden. Es können 17 DIS gezählt werden, welche Daten in 6 oder weniger unterschiedlichen Formaten speichern. Nur 4 DIS geben an, Datensätze in sehr vielen unterschiedlichen Formaten - mehr als 20 Stück - vorzuhalten. Die gebräuchlichsten Formate sind dabei netcdf (13), csv (10) sowie verschiedene Bildformate (7) wie „jpg“, „png“ und „tiff“. Weiterhin sind sowohl „txt“ (5) - und „ASCII“-basierte Formate (4) als auch Vektorformate wie „shp“ (3) in Nutzung. Nennenswert ist noch ein Format zum Austausch digitaler seismologischer Daten namens „miniseed“, welches dreifach genannt wird.

Datentypen im Stil von Typeregistry.org könnten in Zukunft als standardisiertes Element von FAIR Digital Objects (FDOs)[12] an Bedeutung gewinnen. Derzeit werden diese aber von keiner einzigen DIS registriert.

FAIR Digital (manchmal Data) Objects sind Bestandteil von Konzepten, deren Ziel es ist Daten maschinenlesbar verfügbar zu machen. Die Nutzung von FAIR Digital Objects wird von drei DIS unterstützt. Es bleibt zu klären, inwiefern ein einheitliches Verständnis der Definition von FDO unter den DIS gegeben ist und wie dieses gegebenenfalls etabliert werden kann.

Schnittstellen und Zugriff

Es wurde untersucht, welche Zugriffsmöglichkeiten und Schnittstellen auf die DIS, z.B. zu Workflows, Application Programming Interfaces (APIs) oder Protokollen existieren. Außerdem wurde zum Thema der Verwendung von Lizenzsystemen in DIS befragt.

Für den **Zugriff auf Metadaten** gibt es unterschiedliche Modelle. Die Metadaten von 29 DIS sind komplett frei verfügbar. Bei 5 DIS können die Nutzer die Freigabe der Metadaten selbst bestimmen, d.h. beispielsweise Embargoperioden einrichten oder den Zugriff verstecken oder nur auf Nachfrage erlauben. Eine Infrastruktur lässt einen Zugriff sowohl auf Daten, als auch auf Metadaten, nur auf Anfrage zu. Einen generell freien **Zugriff auf die Daten** selbst erlauben 26 DIS, die übrigen nur auf Anfrage.

Einen **Zugriff per API** haben 24 DIS implementiert, 12 unterstützen keine API. 19 der APIs sind öffentlich verfügbar, 5 nur von intern aus den Projekten, Zentren oder Arbeitsgruppen. 17 der APIs sind als REST Schnittstelle implementiert, gefolgt von OAI (6), CSW (5), SOAP (3) und OWS (2). Die APIs aller 24 DIS erlauben einen Zugriff auf die Metadaten. Bei 12 DIS ist es möglich auch auf die Daten selbst zuzugreifen, ggf. mit den zuvor genannten Einschränkungen.

Die Erfassung von **Lizenzen an Datensätzen** wird von 22 DIS unterstützt. Davon stellen sechs DIS ihre Daten unter eine gemeinsame Lizenz, 16 ermöglichen es, Lizenzen individuell mit Datensätzen zu verknüpfen. 10 der DIS stellen maschinenlesbare Lizenzen zur Verfügung. Nach den verwendeten Lizenztypen wurden nicht gefragt.

Nachnutzung

Dieser Abschnitt enthält Angaben zur Nachnutzung und zum Harvesting der Bestände der DIS, d.h. zu den Nutzergruppen und evtl. spezieller Software zur Auswertung.

Als **wichtigste Datenuploader** oder Datenquellen werden mit großer Mehrheit Kolleg:innen aus der internen DIS, aber auch externer Helmholtz-Zentren genannt. Weitere genannte nationale Einrichtungen sind Universitäten und das Bundesamt für Seeschifffahrt und Hydrographie (BSH). Basierend auf Web-Logs werden als Herkunft der Datenbereitstellenden auch Länder wie USA und China identifiziert. Die Angaben zu den **Datendownloadern** decken sich mit denen der Uploader. Daraufhin könnte rückgeschlossen werden, dass die Nutzergruppe ihre eigene DIS nutzen, um

[12] FAIR Digital Objects (FDO): https://fairdo.org/

untereinander Daten auszutauschen. Eine DIS erhebt mit Verweis auf die DSGVO dazu keine eigenen Daten.

Von 37 Antworten sagen 18 der DIS aus, Informationen zur Nachnutzung ihrer Daten zu haben. Diese Informationen beziehen sie überwiegend aus direkten Kontakten zu ihren Nutzern (10), eigenen Statistiken, Statistiken von Suchmaschinen oder Publishern, wie Scholix[13] oder Google Analytics sowie auch Statistiken durch DOI Abfragen. Als Portal, an welches Daten weitergereicht werden oder von welchem aus auf die DIS zugegriffen wird, wird überwiegend das Portal der Deutschen Allianz Meeresforschung DAM (marine-data.de) genannt (9). Weitere Mehrfachnennungen sind B2FIND[14] (3), EPOS[15] (2), DataCite[16] (2) und Google[17] (2). Ferner werden OpenAIRE[18] und eine Reihe fachspezifischer Portale genannt, die entsprechende Daten aggregieren.

Der Zugriff mit Hilfe von extern entwickelter Software auf die DIS erflogt bei 14 DIS. Dabei handelt es sich überwiegend um Eigenentwicklungen von wissenschaftlichen Gruppen, die z.B. per Python-Script oder anderen Scripten (z.B. Matlab, R) die Schnittstellen abfragen. Weitere Nennungen betrafen auch fachspezifische Software, z.B. im Bereich Seismik oder Laboranalytik.

16 der DIS bieten über den reinen Datendownload hinaus weitere Dienste zur Auswertung der Daten an. Diese umfassen z.B. die Visualisierung der Daten (9), die Darstellung auf Karten (5), statistische Funktionen (4), bestimmte Datenfilter (6) sowie das Plotten von Diagrammen oder Grafiken (3).

Bedarfe, Wünsche und Herausforderungen

Besonderes Augenmerk haben wir auf die Ermittlung von möglichen Bedarfen und Verbesserungswünschen der Infrastrukturen gelegt. Das umfasst unter anderem die Zufriedenheit mit dem Status quo der Datenerfassung und -haltung sowie Ideen für Verbesserungen, auftretende Hindernisse, weiterhin Vorschläge für geeignete Anreizsysteme zur Qualitätsverbesserung. Dabei haben wir insbesondere die Differenzierung von Metadaten- und Datenqualität betrachtet.

Von den 35 befragten DIS sind 21 mit der **Qualität ihrer Metadaten** insgesamt zufrieden. 10 sind nicht zufrieden und 4 machten keine Angaben oder es traf nicht auf sie zu. Auf die Frage, für welche Metadatentypen die Erfassung am dringendsten verbessert werden sollte, wurde am häufigsten die Dokumentation der Methodik genannt (16), gefolgt von Kurzfassungen / Abstracts (12), der Erfassung der Provenienz (12) und Parameternamen (11).

Zu **Verbesserung** der Metadatenqualität insgesamt werden eine ganze Reihe von konkreten Vorschlägen gemacht.

Ein großer Teil der Metadaten wird von den Wissenschaffenden selbst erfaßt. Als **Maßnahmen** für eine **Erhöhung der Metadatenqualität** werden überwiegend eine verbesserte Standardisierung der Metadaten und der Vokabularien vorgeschlagen (8). Probleme in der Genauigkeit bei der Dokumentation von Metadaten könnten auch aufgrund von Zeitdruck bei Forschenden (4) sowie durch bestehende Unsicherheiten von Forschenden bezüglich Lizenzsystemen und Datenpolicies (3) auftreten. Weitere Vorschläge betreffen die Umstände der Datenerhebung, bei der sich viele eine **stärkere Automatisierung der (Meta)Datenerfassung** wünschen. Auch eine bessere technische

[13] Scholix: http://www.scholix.org/.

[14] B2FIND: http://b2find.eudat.eu/.

[15] European Plate Observing System (EPOS): https://www.epos-eu.org/.

[16] DataCite: https://datacite.org/.

[17] Google Dataset Search: https://datasetsearch.research.google.com/.

[18] OpenAIRE: https://explore.openaire.eu/.

Unterstützung von Forschenden bei der Vorbereitung von Experimenten oder Messungen wurde genannt.

Es wurde angemerkt, dass der Fokus vieler Arbeitsgruppen auf der Sicherung der Daten an sich, weniger auf die Verständlichkeit der Beschreibungen liegt. Dies wurde auf **mangelndes Bewusstsein** für die Datendokumentation und deren Nutzen zurückgeführt sowie auf eine **mangelnde Ausbildung im Datenmanagement**. Dafür sind offenbar zu wenig praxisnahe, in Kuration qualifizierte Hilfsangebote vorhanden. Auch ein einfacherer Zugang zu **Schulungen** für Wissenschaffende wurde häufig genannt. Durch Schulungen könnten verbindliche Vorgaben vermittelt werden und durch Best-practice Beispiele der Mehrwert der Datendokumentation dargestellt werden. Als **Anreize** für Wissenschaffende könnten (nicht nur) in den DIS Scoringsysteme für gute Datendokumentationen implementiert werden.

Die meisten DIS würden sich die Möglichkeit wünschen, ihren **Kurationsprozess** zu optimieren. Das Kuratieren und Aktualisieren von Daten und Metadaten würde einen erhöhten Einsatz von Personal erfordern, welches zusätzliche Mittel erfordern würde. Einige DIS schlugen daher vor, mehr Ressourcen bereitzustellen (7). Auch ein Upgrade der DIS an neue Erfordernisse, z.B. die Implementierung von aktuellen PID Systemen, erfordert Ressourcen, die oft nicht vorhanden sind.

Weitere Maßnahmen zur Verbesserung der Metadatenqualität bezogen sich auf den **Datenabgleich** und eine verbesserte Kommunikation zwischen DIS um z.B. Daten aus derselben Quelle miteinander in Beziehung setzen zu können und auch die Provenienz von Datensätzen besser darstellen zu können. Dafür wäre eine Vernetzung auch mit internationalen Strukturen notwendig. Es wurde beklagt, dass diese **Abstimmungsprozesse** häufig wenig Priorität in den DIS haben, da der Nutzen und Aufwand für viele in keinem erkennbaren Verhältnis stehen. Genannt wurde auch eine verbesserte Benennung von Ansprechpartnern in den DIS um den Austausch zu verbessern.

Als **Herausforderung bei der Erfassung von Daten durch Forschende** wird vor allem ein fehlendes Bewusstsein bzw. Anreize für RDM bei den Forschenden (7), in Verbindung mit Zeitmangel bei den Forschenden (5) genannt. Als Konsequenz sollte die Erfassung von Metadaten z.B. durch Automatisierung vereinfacht werden (4). Einheitliche Vokabularien, Bezeichnungen und Standards (4) könnten die Eingabe erleichtern und die Qualität verbessern. Dem gegenüber steht eine oft mangelnde Fachkenntnis der Erfassenden und eine daraus resultierende sehr aufwendige Datenkuration (4). Eine DIS unterstützt zum Beispiel mit „standardisierten und gereviewten Datenbeschreibungen“ als "Beipackzettel“, welcher die wichtigen Informationen bereitstellen und den Erfassenden den Einstieg erleichtern sollen. Eine detaillierte Kuration der Daten und Metadaten ist dabei dennoch notwendig.

Um die **Nachnutzung der Daten durch Forschende** zu verbessern, wird vorgeschlagen, die Erfassung der Provenienz zu verbessern (3) und Referenzen auf andere (Meta-)Datensätze, Publikationen oder Methodiken besser zu erfassen (3). Im Zugriff sollten APIs vereinheitlicht sowie die Möglichkeit für größere Datenübertragungen geschaffen werden (3). Auch die Bestimmung und Darstellung der Datenqualität ist ein Hindernis (2), welches eine Harmonisierung von Datensätzen schwierig gestaltet. Die Bereitstellung von geeigneten Tools zur Datenprozessierung könnte die Qualität und damit die Nachnutzbarkeit deutlich verbessern (2), weil dadurch einheitliche Prozessierungsschritte etabliert und Fehler vermieden würden. Wie auch in den anderen Bereichen wurde auch hier die Bedeutung dokumentierter Lizenzen, semantische Harmonisierung und Maschinenlesbarkeit von einzelnen DIS hervorgehoben. Ein konstruktiver Austausch mit den Forschenden wäre wahrscheinlich ebenfalls hilfreich.

Von 31 geben 37 DIS an, Interesse daran zu haben, ihre Interoperabilität mit anderen DIS zu verbessern. Von den vorgeschlagenen Maßnahmen wurde eine intensivere Zusammenarbeit (17) sowie ein verbessertes Mapping von Metadatenstandards (17) priorisiert, gefolgt von einer Verbesserung semantischer Konzepte (10) und einem besseren Kontakt zur wissenschaftlichen Community (10).

Als weitere Hürde für die Implementierung von verbesserten interoperablen Verfahren wird ausgesagt, dass es einfacher sei, traditionelle, historisch gewachsene, „händische“ Methoden zu belassen, als die möglicherweise zeitlich eher aufwendigeren Veränderungen umzusetzen. Verbunden mit einer

mangelnden Standardisierung von Metadaten und heterogenen Datensätzen entstünden schnell komplexe Probleme bei der Datenbeschreibung, welche ohne Training oder geeignete Beratung nur schwer oder nicht zu lösen sind. Das führt dazu, dass „für eine bestimmte Nachnutzung meist einige Detailkenntnis erforderlich ist, d.h. man kann Interoperabilität eigentlich immer nur mit bestimmten Anwendern entwickeln“. Ein „Einigungsprozess in nationalem oder besser internationalem Kontext“ wäre in Bezug auf die Standardisierung von Verfahren zur Interoperabilität hilfreich. Eine DIS wurde im Wesentlichen von technischem Personal betrieben, das selbst im Bereich FDM keinen Support leisten kann. In solchen Fällen sei die Kommunikation zwischen Forschenden und technischem Personal häufig schwierig. Außerdem wurde angemerkt, dass durch proprietäre Datenformate und eine nötige regelmäßige Aktualisierung der Datenformate eine evtl. einmal vorhandene Interoperabilität sukzessive verlorengehen könnte.

Die DSGVO sehen nur 3 DIS für ihre Arbeit als problematisch an. Kritisiert werden inhaltliche Unklarheiten bei der Umsetzung, insbesondere im Umgang mit persönlichen, aber z.T. frei zugänglichen Daten in Zusammenhang mit Datensätzen. Eine Lockerung der DSGVO und vor allem klarere Regelungen wurden vorgeschlagen.

Eine Unterstützung bei ihrer Arbeit durch nationale Initiativen wie NFDI, HMC, EOSC oder ähnliche würden sich bei 35 Antworten 18 DIS wünschen. Geäußerte Wünsche umfassen z.B. Hilfen bei der Vernetzung der DIS und der Priorisierung von Aktivitäten. Vorgeschlagen werden auch die Vereinbarung und Umsetzung von Standards und Vokabularien und der gemeinsamen Organisation von Fortbildungsveranstaltungen für die Wissenschaft. Auch Hilfe beim Hervorheben der Bedeutung von RDM und damit verbundener nachhaltiger Finanzierung wird gewünscht.

6 Erste Erkenntnisse und Schlußfolgerungen

Auch wenn sich nicht alle angeschriebenen DIS in der Helmholtz-Gemeinschaft an der Erhebung beteiligt haben, sind die meisten und, vor allem, die größeren Repositorien aus dem Forschungsbereich Erde und Umwelt vertreten. Die Erhebung zeigt daher einen guten Überblick und einen aus unserer Sicht repräsentativen Querschnitt durch die Kapazitäten und den Entwicklungsstand der DIS der Helmholtz Gemeinschaft. Es sind eine Reihe unterschiedlicher Motivationen zur Einrichtung der DIS, fachlich-, technisch-, policy-getrieben erkennbar. Auch die inhaltliche Fokussierung auf unterschiedliche fachliche Zielgruppen und Wissenschaftscommunities sind klar erkennbar und **für uns eine hilfreiche Referenz bei der Planung zukünftiger Aktivitäten**.

Weiterhin ist festzustellen, dass der Betrieb der verschiedenen DIS sehr individuell umgesetzt wird. Viele DIS werden wenigstens teilweise dauerhaft aus ihren Zentren finanziert. Für fast alle spielen externe und zeitlich befristete Mittel eine wichtige Rolle. Ein Ziel der Befragung war es unter anderem auch, Informationen über die Nachhaltigkeit der DIS ableiten zu können. Ob zeitlich befristete Mittel für den Basisbetrieb notwendig sind oder nicht, wurde nicht abgefragt. Aus diesem Grund können wir nur vermuten, dass der Basisbetrieb der DIS in den meisten Zentren über zeitlich befristete Projektmittel querfinanziert wird.

Bislang sind innerhalb der Helmholtz-Gemeinschaft keine Festlegungen dafür getroffen worden, dass eine DIS dauerhaft in den jeweiligen zentralen Infrastrukturen einzurichten ist. Ob eine solche Vereinbarung umsetzbar ist, müsste mit den betreibenden Zentren im Einzelnen geklärt werden. Eine Finanzierung aus dem Zentrumshaushalt weist aber zunächst auf eine zentrumsseitige, nachhaltige, langfristige Perspektive der DIS hin. DIS, die überwiegend aus befristeten externen Mitteln betrieben werden, könnten ggf. ohne Finanzierung nicht mehr zuverlässig und dauerhaft verfügbar sein. Dafür wäre möglicherweise in Modell der Umlagefinanzierung durch Nutzende zu diskutieren. Die Abhängigkeiten der DIS vom jeweiligen **Finanzierungsmodell und deren Auswirkungen auf die Nachhaltigkeit** sollten genauer betrachtet werden.

Eine zentrale Fragestellung der Erhebung war: Wie FAIR sind die DIS in der Helmholtz-Gemeinschaft? Fast allen DIS sind die FAIR-Prinzipien bekannt, und viele bemühen sich, diesen im Rahmen ihrer Möglichkeiten gerecht zu werden. Dabei verfolgen die Beteiligten unterschiedliche Strategien und treffen im Zuge dessen auf unterschiedliche Hindernisse. Bei der Analyse der Antworten im Hinblick auf die FAIR-Prinzipien, wird sichtbar, dass die Herausforderung bei der Umsetzung mit der Tiefe und organisatorischen **Komplexität bei der Implementierung der FAIR-Prinzipien** zunehmen[19]. Während viele Infrastrukturen sich sehr bemühen, Kriterien zur Auffindbarkeit (Findability) und Zugriffsmöglichkeit (Accessibility) über DOIs und Metadatenstandards zu erfüllen, ist es offensichtlich schwieriger die Interoperabilität und Nachnutzbarkeit (Reusability) umzusetzen.

Die meisten DIS nutzen etablierte Metadatenstandards, einige unterstützen die Datendokumentation mit Hilfe von mehreren Standards, manche stellen Metadaten in proprietären Formaten zur Verfügung. Am häufigsten verbreitet ist dabei die ISO191XX und deren Derivate (z.B. INSPIRE). Auch sind bei vielen DIS verschiedene APIs für externe Zugriffe implementiert. Sowohl der Einsatz von Standards als auch die Nutzung von APIs könnten auf eine gut durchdachte, strukturierte Implementierung des jeweiligen Repositoriums rückschließen lassen. Dies ist jedoch keine zwingende Voraussetzung für eine Interoperabilität, weil Metadateninhalte desselben Standards in unterschiedlicher Struktur ausgeliefert werden können und weil es durchaus Spielräume bei deren Umsetzung gibt. Eine **intensivere Abstimmung der DIS untereinander** und eine Einigung auf gemeinsame Austausch-formate und Prozesse, könnte die Anschlussfähigkeiten der DIS untereinander deutlich erhöhen.

Die Umsetzung semantischer Konzepte in den Metadaten ist in der Praxis weitgehend nicht geschehen. Von einigen DIS wird hierfür ein Bedarf zur Verbesserung der Interoperabilität gesehen. Einige, wenige DIS unterstützen schema.org-kompatible Metadaten, z.B. in ihren APIs oder Webseiten. Datensätze in diesen DIS sind dadurch über verbreitete Suchmaschinen auffindbar. Einige DIS unterstützen die Erfassung von Vokabularien, eine konsequente Implementierung von Ontologien oder „leichteren" semantischen Konzepten wie Glossaren oder Thesauri wird jedoch nicht umgesetzt. Ein möglicher Grund dafür könnte sein, dass in den meisten DIS gegenwärtig die **nötige Expertise für die Umsetzung von Ontologien fehlt**. Darauf deuten einige Antworten auf Fragen in diesem Kontext hin. Ähnliches gilt für die Umsetzung von Provenienzinformationen in den meisten DIS, wie auch für das Verständnis von FAIR Digital Objects (FDO). Beide Konzepte werden in vielen DIS derzeit vermutlich nicht ausreichend verstanden, um diese praktisch umsetzen zu können.

Viele DIS sehen sich als Serviceeinrichtung für die Wissenschaft und sind daher bemüht, Anregungen und Bedarfe aus der Wissenschaft ggf. im Zusammenspiel mit den Förderinstitutionen umzusetzen. Sehr viele DIS nutzen DOIs, um Datensätze als Datenpublikationen persistent zu referenzieren. Der Einsatz von DOIs ist zum derzeitigen Zeitpunkt die beste etablierte Methode, um Datensätze referenzierbar zu machen. DOIs sind in der Bibliothekswelt für Publikationen unabdingbar. Allerdings liefern sie nur eine einzelne Referenz, und die Metadaten verschiedener **DOI Implementierungen sind nicht konsequent standardisiert**. Obwohl es Mapping-Tabellen (Crosswalks) zwischen verschiedenen DOI Implementierungen gibt, ist eine Vergleichbarkeit von Metadateninhalten über DIS hinweg nur schwer möglich. Über die Bereitstellung von DOIs allein ist daher eine Interoperabilität nur sehr begrenzt herstellbar. Ein Grund für die hohe Verbreitung von DOIs liegt möglicherweise darin begründet, daß diese von vielen Forschungsförderern als Datenreferenz in Projektberichten verpflichtend eingefordert werden. Noch dazu gelten DOIs inzwischen fast als Gütesiegel für einen gut dokumentierten Datensatz. Letztere Einschätzung ist allerdings mit Blick auf die mangelnde Vollständigkeit vieler Metadatensätze in Repositorien leider nicht haltbar[20].

Die wissenschaftlich ausgerichtete Nachnutzergemeinde von Forschungsdaten ist den DIS weitgehend unbekannt. Die Notwendigkeit zum Aufbau einer übergeordneten Struktur, welche die verschiedenen

---

[19] Söding Emanuel. 2022. „Wie weit sind wir noch von FAIR entfernt?" RDA-Deutschland Tagung 2022, https://indico.desy.de/event/31380/contributions/115818/attachments/71931/91924/RDA_DE_2022_DataPublishingBestPractices_WieWeitSindWirNochVonFAIREntfernt.pdf.

[20] Burger, Marleen, Anette Cordts, und Ted Habermann. 2021. „Wie FAIR Sind Unsere Metadaten? : Eine Analyse Der Metadaten in Den Repositorien Des TIB-DOI-Services". *Bausteine Forschungsdatenmanagement*, Nr. 3 (Dezember).1-13. DOI:10.17192/bfdm.2021.3.8351.

DIS verbindet und einen **interoperablen Datenraum** öffnet, wie z.B. in der EOSC angedacht[21], wird von den DIS zwar gesehen, offenbar aber aufgrund mangelnder Ressourcen zugunsten ihres direkten Serviceauftrages an die datenproduzierenden Wissenschaffenden nicht umgesetzt. Dadurch wird die Argumentation zur Umsetzung eines interoperablen Datenraumes zu einem abstrakten Ziel, das denjenigen, die über die Finanzmittel entscheiden, nur schwer zu vermitteln ist. Diese Entscheider erwarten, daß in der Regel mittelbar wissenschaftliche Ziele verfolgt werden. Der kurzfristige Nutzen eines Datenraumes ist, ohne den Kontext eines wissenschaftlichen Mehrwerts durch die in Zukunft gesicherte, mögliche Nachnutzung, nur schwer vermittelbar. Hierin liegt eine große Herausforderung für das Forschungsdatenmanagement im Allgemeinen, mehr Informationen zur Nutzung der Forschungsdaten zu sammeln, **Datensätze gegebenenfalls sinnvoll zu aggregieren und maschineninterpretierbar für geeignete Studien zur Verfügung zu stellen**. Diese Aufgabe kann aber nur gemeinsam mit Datennachnutzern und entsprechend ausgewählten Anwendungsfällen umgesetzt werden.

Die Befragten liefern eine Reihe von **Vorschlägen**, wie verschiedene Aspekte des FDM verbessert werden könnten. Diese reichen von Verbesserungen in ihrem eigenen Bereich, z.B. einen verbesserten Zugang zu finanziellen Ressourcen, aber auch z.B. Trainings für die Datenproduzierenden / Forschenden um die Dokumentation zu verbessern. Insgesamt beziehen sich die Vorschläge auf ein Rollenverständnis, in dem Forschende Daten produzieren und dokumentieren und die FDM Spezialisten / Data Stewards diese verwalten und in einen FAIRen Kontext bringen. Die Möglichkeit andere Gruppen in den Datenerfassungsprozess mit einzubeziehen und Verantwortung zuzuweisen, z.B. technisches Personal, Operateure von Maschinen, Labors oder Experimenten, Bibliotheken, Wissenschaftsmanagende der Zentren mit der Erfassung ihrer Metadaten zu befassen, wird nicht geäußert. Das deutet auf ein relativ einfaches Lieferanten (Wissenschaffende) – Empfänger (DIS)-basiertes **Rollenverständnis im Datendokumentationsprozess** hin. Dieses sollte gemeinsam mit den DIS hinterfragt werden, um die **Aufgaben besser zu verteilen** und dadurch die Qualität der Metadaten sowie die automatisierte Erfassung aus extern gepflegten Datenquellen, zu verbessern.

7 Weiteres Vorgehen

Wie erwartet, zeigt sich bei den befragten DIS ein sehr heterogenes Bild, von drittmittelfinanzierten Projektdatenbanken bis hin zu seit Jahren etablierten Repositorien, die maschinelle Zugriffe ermöglichen, gängige Metadatenstandards ausliefern und für jedermann leicht auffindbar sind. Welche Schlüsse kann man also aus der Erhebung ziehen, um die nächsten Schritte auf der „Road to FAIR“ beschreiten zu können? Wie können die DIS untereinander anschlussfähig gemacht werden, und wie können wir einen gemeinsamen „Helmholtz-Datenraum“ (siehe Einleitung) entwickeln, der international an andere Datenräume, z.B. der EOSC, anschlussfähig sein wird?

Die Prinzipien von FAIR – Auffindbarkeit (F), Zugriffsmöglichkeit (A), Interoperabilität (I) und Nachnutzbarkeit (R) bauen inhaltlich und strukturell in vielen Bereichen aufeinander auf und bringen daher bei der Implementierung besondere Herausforderungen mit sich. Konzepte für F und A werden weitgehend in den DIS selbst, mit jeweils unterschiedlichen Methoden, umgesetzt. Der nächste Schritt auf der „Road to FAIR“, die Umsetzung von I - Interoperabilität und R - Nachnutzbarkeit, kann nur in enger Abstimmung innerhalb der DIS und der wissenschaftlichen Community erreicht werden. Dazu zu gehören in vielen Bereichen auch Prozesse, die für F und A durch DIS bereits unterschiedlich vorliegen. **Der Abstimmungsprozess, in dem gemeinsame Prozesse und Verfahren zur Datendokumentation innerhalb der technischen und wissenschaftlichen Community evaluiert und umgesetzt werden, muss unbedingt gestartet und koordiniert werden.** Der Prozeß des Forschungsdatenmanagements als eine wissenschaftsgeleitete, breit aufgestellte und abgestimmte

[21] Federated FAIR Data Space - F2DS: https://www.eosc-pillar.eu/federated-fair-data-space-f2ds.

Aktivität ist ein wichtiger Faktor des oft beschworenen „Cultural Change“ und ein wesentlicher Unterschied zur gegenwärtigen Praxis.

Für die Planung eines Abstimmungsprozesses, die Priorisierung von Aktivitäten und die Umsetzung in kleinen Schritten, kann diese Bedarfserhebung zu Grunde gelegt werden. Zum einen können Anwendungsfälle (Use Cases) bzw. Implementierungsbeispiele identifiziert werden, zum anderen auch bestehende Lücken aufgezeigt werden. Die am dringlichst erscheinenden Handlungsfelder sind demnach folgende:

1. Eine konsequente Nutzung abgestimmter und besser aufgelöster Metadaten, z.B. auf Basis von verschiedenen bestehenden, bereits etablierten, ausgewählten **PID Systemen** sollte konsequent implementiert werden. Derzeit gibt es in den Helmholtz DIS keine gute Abstimmung, wo bestimmte übergreifend einheitliche Referenzinformationen zu finden sind. Um Datenbanken mit Referenzinformationen nutzbar machen zu können, sollen ggf. neue PID Systeme entwickelt werden. Die alleinige Nutzung von DOIs, die von vielen DIS verfolgt wird, ist hier aus unserer Sicht nicht ausreichend. Für diese ist zwar eine gemeinsame Auflösungsverfahren verfügbar. Die unterschiedlichen Implementierungen machen eine interoperable Nutzung aber schwierig bis unmöglich. **Stattdessen sollte ein PID-Graph entwickelt werden**, aus dem hervorgeht, an welcher Stelle Referenzinformationen vorgehalten werden und wie diese abrufbar sind. Außerdem muß in dem PID-Graph vereinbart werden, welche Informationen ggf. redundant auftauchen und **welche Informationen als belastbare Referenz dienen können**.

2. Geeignete **semantische Konzepte** sollten gemeinsam mit Wissenschaffenden entwickelt werden und in sämtliche Metadaten der Datenprodukte der Helmholtz-Gemeinschaft integriert werden. Das ist die Voraussetzung, um einen **Helmholtz-Knowledge-Graphen** zu erstellen, welcher eine Helmholtz-interne semantische Interoperabilität gestattet. Erste Ansätze sind z.B. mit der Implementierung von schema.org kompatiblen Metadaten in einigen DIS sichtbar und sollten mit Unterstützung von HMC weiterverfolgt werden. Die Ergebnisse der Erhebung weisen darauf hin, dass eine fehlende Expertise bei den DIS die Implementierung in diesem Bereich behindert. Daher wären sowohl **Weiterbildungen** durch HMC als auch die Anwerbung geeigneten Personals wichtige Aktivitäten, um entsprechende Kenntnisse zu vermitteln und dieses Thema voranzutreiben.

3. Die **Maschinenlesbarkeit** sollte durch einheitliche, harmonisierte Schnittstellen mit abgestimmten APIs verbessert werden. Die breite Anwendung von REST-Schnittstellen, die viele DIS bereitstellen, ist ein erster Schritt zu einer technisch einheitlichen Lösung. Die Sicherstellung eines inhaltlich konsistenten, automatisierten Austauschs könnte z.B. mittels Implementierung von gemeinsam getragenen FDO Konzepten in den DIS geschehen. Auch hier könnte HMC mit Weiterbildungen ansetzen, um diese Konzepte zu vermitteln und die Umsetzung voranzutreiben.

4. Die **Provenienz** der Daten ist nur selten beschrieben und sollte im Dokumentationsprozess nachvollziehbar abgebildet werden. Auch in diesem Bereich wären **Weiterbildungen** durch HMC und die Anwerbung geeigneten Personals wichtig, um dieses Thema voranzutreiben.

5. In den Bereichen **Semantische Systeme, FDOs und Provenienz** ist es ebenfalls entscheidend, gute Use Cases zu entwickeln und umzusetzen. Hier könnte HMC unterstützend wirken und gezielt Projekte mit geeigneten Partnern in den DIS initiieren, die als Pilot-Anwendungsbeispiele bzw. Use-Cases dienen können.

6. Die oft gewünschte, automatische Erfassung von Metadaten kann nur über eine **breitere Einbindung von verschiedenen, für den jeweiligen Schwerpunkt zuständigen Stakeholder-Gruppen** umgesetzt werden, welche bestimmte Referenzdaten selbständig pflegen. Das könnten z.B. Daten zu Organisationen, Personen, Instrumenten, Methoden, Lizenzen usw. sein. Diese Referenzdaten sollten als Handle- / PID-Systeme von entsprechenden Gruppen gepflegt werden und könnten bei Bedarf **automatisiert eingebunden** werden. Hier sollte innerhalb des FDM überlegt werden, wie extern gepflegte Datenquellen nutzbar gemacht werden können, damit

Gruppen jenseits der Wissenschaft, z.B. aus dem technisch-administrativen Bereich, verantwortlich in den Datendokumentationsprozess eingebunden werden können.

7. Unbedingt notwendig ist die verbesserte **Einbindung von datennachnutzenden Wissenschaffenden**, um deren Bedarfe bei der Planung des Datenraumes berücksichtigen zu können. Nur so können die Stellschrauben für die Erhebung und Bereitstellung von Metadaten entsprechend justiert werden, damit die Zielgruppe erfolgreich und nutzerfreundlich aus den verschiedenen Datenquellen der Helmholtz-Gemeinschaft schöpfen kann.

8 Umsetzungsplan im Co-Design Verfahren

Die vorliegende Erhebung liefert eine Fülle an Informationen darüber, **wie Helmholtz-DIS die genannten Handlungsfelder bereits adressieren**. Diese Informationen sind wichtig für **die Entwicklung eines schrittweisen Planes**, wie F und A verbessert und I und R umgesetzt werden können. Aufgrund der oft unterschiedlichen Lösungsansätze in einigen Bereichen kann rückgeschlossen werden, daß nur begrenzt verbindliche Absprachen im Umgang mit Forschungsdaten zwischen den DIS getroffen werden. Die Umsetzung von FAIR-Prinzipien zur Erstellung eines maschineninterpretierbaren, interoperablen Datenraumes ist aber eine **Community-Aufgabe** und kann nicht an einzelne Personengruppen delegiert werden. Viele Gruppen werden dabei bisher weder ausreichend berücksichtigt noch in die Verantwortung genommen. Die Entwicklung und Erweiterung des bestehenden Rollenmodells könnte eine veränderte Verteilung von Aufgaben zur Folge haben. Dafür könnten auch **angepasste Policies** an den Zentren hilfreich sein, um diese Aufgaben entsprechend zuweisen zu können.

Wir schlagen daher über die oben skizzierten Maßnahmen hinaus vor, in einen **offenen, breit angelegten Community-Dialog im Co-Design Verfahren einzutreten**, um

- ein gemeinsames Verständnis von FAIR zu entwickeln
- gemeinsame Implementierungsstrategien für FAIR zu planen und umzusetzen
- die Verschiedenen beteiligten Stakeholdergruppen zu identifizieren, zu adressieren und am Dialog zu beteiligen
- für die geäußerten Bedarfe der wissenschaffenden Datenproduzenten und Datennachnutzern gemeinsam Lösungen zu entwickeln

Das Team HMC „Erde und Umwelt“ wird für diesen Dialog einen geeigneten Vorschlag entwickeln.

Danksagung

Diese Datenerhebung wäre nicht möglich gewesen, ohne die engagierte Unterstützung unserer Kollegen in HMC und an den Zentren. Wir danken unseren Kollegen der HMC Taskforce Survey - Silke Gerlich, Markus Kubin, Lucas Kulla, Christine Lemster, Katharina Rink, Jan Schweikert und Sangeetha Shankar für viele hilfreiche Hinweise bei der Entwicklung des Fragebogens und technische Unterstützung mit dem Umfragetool. Für die gründliche Durchsicht, die praktischen Tests der Fragen und die konstruktiven Rückmeldungen danken wir insbesondere den Kollegen der FDM Gruppen von GEOMAR Pina Springer und Hela Mehrtens sowie dem GFZ Kirsten Elger und Roland Bertelmann. Wir danken ebenfalls all jenen Kolleginnen und Kollegen, die die den Fragebogen so gewissenhaft ausgefüllt haben.