ULG Library and Information Studies, MSc Modul 9.3. Forschungsdatenmanagement

Termin: 28.10. 14.15 – 15.45 Uhr (2 UE)

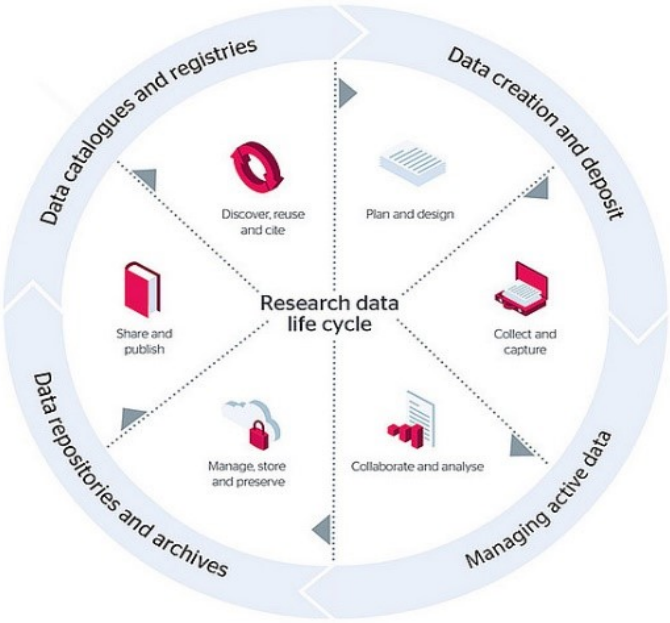
Vortragende: Sarah Stryeck (TU Graz; sarah.stryeck@tugraz.at)

Thema: Forschungsdatenmanagement in der Praxis … Lebens – und Naturwissenschaften

Begriffsklärungen

Forschungsdatenmanagement (FDM)

FDM bezieht sich auf den gesamten Lebenszyklus von Forschungsdaten - von der Planung, Generierung, Analyse, Verarbeitung und Sicherung von Forschungsdaten über ihre Dokumentation, Archivierung und Veröffentlichung bis zur allfälligen Nachnutzung durch Dritte. (https://ub.uni-graz.at/de/dienstleistungen/publikationsservices/forschungsdatenmanagement/)



Datenmanagementpläne

Ein Datenmanagementplan (DMP) strukturiert und beschreibt den Umgang mit Forschungsdaten eines wissenschaftlichen Projekts von der Erhebung bis zur Archivierung. Er hilft somit, den gesamten Forschungslebenszyklus zu planen und zu überblicken und Datenverlust sowie Zusatzaufwand zu vermeiden. In DMP ist beispielsweise anzugeben, welche Daten generiert werden sollen, wer für die Organisation, Ablage und Erhaltung der Daten verantwortlich ist und welche Daten wann und wie zugänglich gemacht werden sollen. (https://ub.uni-graz.at/de/dienstleistungen/publikationsservices/forschungsdatenmanagement/datenmanagementplaene/)

FAIR Data

Eine Verfügbarkeit und Zugänglichkeit von Forschungsdaten trägt dazu bei, die Nachnutzbarkeit einmal erhobener Daten zu steigern und die Nachvollziehbarkeit von Forschungsergebnissen zu sichern. Forschungsdaten sollten daher auffindbar, zugänglich, interoperabel und wiederverwendbar sein (FAIR Data – findable, accessible, interoperable, reusable). Insbesondere aussagekräftige Metadaten, persistente Identifikatoren und eindeutige Lizenzen sind erforderlich, um Daten für Menschen und Maschinen nachnutzbar und damit FAIR zu machen. FAIR Data sind nicht notwendigerweise Open Data: sie können auch eingeschränkt zugänglich sein, zentral ist aber ihre möglichst exakte Beschreibung. (https://ub.uni-graz.at/de/dienstleistungen/publikationsservices/forschungsdatenmanagement/fair-data-und-open-data/)

European Open Science Cloud

Die European Open Science Cloud (EOSC) ist ein Konzept einer wichtigen wissenschaftlichen Kerninfrastruktur für die europäische Wissenschaft. Sie soll eine virtuelle Umgebung mit offenen und nahtlosen Diensten für die Speicherung, Verwaltung, Analyse und Wiederverwendung von Forschungsdaten schaffen. Dazu sollen bestehende nationale, regionale und institutionelle Dateninfrastrukturen, die derzeit über die Disziplinen und die EU-Mitgliedstaaten verteilt sind, vernetzt werden. (https://ub.uni-graz.at/de/dienstleistungen/publikationsservices/forschungsdatenmanagement/european-open-science-cloud/)

Repositorium

Ein Repositorium besteht im Wesentlichen aus einer Repositoriumssoftware und einer Datenbank. Die datengebenden Personen können die Daten über eine webbasierte Benutzeroberfläche in das Repositorium überführen oder die Repositorienbetreibende sammeln diese automatisiert über entsprechende Protokolle und Schnittstellen von anderen Plattformen ein (harvesten). Für die Nachnutzung durch Dritte werden zusätzlich zu den eigentlichen Daten deren Metadaten benötigt. Diese kann die Datengeberin oder der Datengeber z. T. aus anderen Anwendungen übernehmen oder manuell hinzufügen. Metadaten beschreiben den Inhalt der Forschungsdaten und liefern Informationen über deren Entstehung, dabei verwendete Software bzw. Methoden sowie rechtliche Aspekte. In den Metadaten sollten außerdem Nutzungsbedingungen in Form von Lizenzen festgelegt werden, die u.a. den Zugang zu den Daten regeln (Registrierung, Embargo o. ä.). Damit die Daten dauerhaft referenzierbar und zitierfähig sind, vergeben die meisten Repositorien eindeutige persistente Identifikatoren. Sowohl über die persistenten Identifikatoren (oft DOIs oder URNs) als auch über entsprechende Schnittstellen werden die Inhalte vieler Repositorien in Suchmaschinen und Fachdatenbanken indiziert (z.B. Google Scholar). Des Weiteren verfügen Repositorien über eine Suchfunktion, mit der die Nutzerinnen und Nutzer die enthaltenen Daten finden, betrachten und herunterladen können. (https://www.forschungsdaten.info/themen/veroeffentlichen-und-archivieren/repositorien/)

FDM in den Naturwissenschaften

In den Naturwissenschaften versucht man Phänomene durch Beobachtungen, Analyse und Messungen zu erklären. Die Disziplinen in den Naturwissenschaften (z.B. Chemie, Physik, Astronomie) sind sehr unterschiedlich. Sie verwenden eine Vielzahl an Methoden um Forschungsdaten zu generieren. Demnach unterscheiden sich auch die Datensätze und es ist sehr schwer übergreifende, einheitliche Richtlinien im Bereich FDM zu etablieren. Zudem sind manche Messungen sehr teuer (z.B. Sequenzierungen, Arzneimittelentwicklung), inkludieren sensitive Daten (z.B. MRT Messungen von Patienten) und können manchmal nicht wiederholt werden (z.B. Beobachtung von Sternkonstellationen in der Astronomie). Ein weiterer Aspekt ist, dass ein Großteil der Forschungsdaten von öffentlichen Geldern finanziert wird und somit auch der Öffentlichkeit zur Verfügung stehen sollte. Es werden auch von Funding Organisationen, sowie von Publishern vermehrt Open und FAIR Data Management gefordert. Demnach ist es essentiell, dass die Forschungsdaten FAIR sind und somit eine nachhaltige Nutzung der Daten zu garantieren und die Anforderungen der Stakeholder zu erfüllen.

Entwicklungen in den Naturwissenschaften

Es entwickeln sich in den Naturwissenschaften viele Hochdurchsatz-Methoden (z.B. Massenspektrometrie, Synthese in anorganische Chemie, Screening in Pharmabranche). Diese Experimente generieren in einem kurzen Zeitraum eine enorme Menge an Daten. Diese rasant wachsenden Datenmengen werden oft als Big Data bezeichnet. Parallel dazu entwickeln sich auch immer mehr IT-Systeme und Lösungen, welche diese großen Datenmengen gut verarbeiten können (Big Data Analytics, beispielsweise Machine Learning Algorithmen). Diese neuen Datenmengen und Verarbeitungstools bilden die Chance für die effiziente Erzeugung von Wissen, aber bringen auch neue Herausforderungen für Datenmanagement, Verwaltung und Weiternutzung mit sich.

Eine zentrale Herausforderung ist die Bereitstellung von Daten für unterschiedliche Nutzergruppen und die Wiederverwendbarkeit. Deshalb werden sogenannte Datenplattformen in vielen Bereichen benötigt. Beispielsweise in der Physik/Materialwissenschaften ist das Teilen von Daten sehr wertvoll, da viele Stakeholder die Daten benötigen (Synthetiker, Forschende, Endnutzer). Ein Beispiel für eine solche Datenplattform ist die sogenannte Innovations-Plattform MaterialDigital.[1]

> *Wegen der Vielfalt der Materialien und der mit ihrer Herstellung und Nutzung verbunden Prozesse, der Komplexität der Lebenszyklen von Materialien und der Vielzahl von akademischen und industriellen Beteiligten gibt es viele zerstreute Aktivitäten, doch es fehlt eine Bündelung der Anstrengungen aller Stakeholder, die Redundanzen und mangelnde Akzeptanz beheben und eine gemeinsame Basis in der Digitalisierung der Materialien vorantreiben kann. Unser Ziel zur Bewältigung dieser Herausforderung ist die mit allen Stakeholdern konsistente Kontextualisierung der Materialdaten: Alle notwendigen Information zum Materialzustand einschließlich der fertigungs- und einsatzbedingten Veränderungen müssen über eine einheitliche, maschinenlesbare Beschreibung verfügbar gemacht werden. Die Basis dafür zu legen ist eine weitere zentrale Aufgabe: Aufbau und Transfer einer leistungsfähigen Ontologie für Materialien und der damit verbundenen Prozesse in die Anwendungen.*[2]

[1] https://www.materialdigital.de/
[2] https://www.materialdigital.de/

Eine weitere Möglichkeit, um die Wiederverwendbarkeit von Forschungsdaten zu erhöhen, welche auch von vielen Naturwissenschaftlern genutzt wird, ist die Publikationen von Forschungsdaten in Data Journals. In diesen Publikationen dokumentiert und beschreibt man Forschungsdaten im Bezug auf (i) Datenerhebung, (ii) Charakteristika, (iii) Funktionen und (iv) potentielle Nachnutzungsmöglichkeiten. Beispiele sind:

- Scientific data (https://www.nature.com/sdata/)
- Data in Brief (https://www.journals.elsevier.com/data-in-brief)
- Data (https://www.mdpi.com/journal/data)
- Data Science Journal (https://datascience.codata.org)

Wie bereits angeschnitten wird auch das FDM von Fördergebern und Publishern vermehrt gefordert. Parallel zu diesen Entwicklungen werden immer mehr Standards entwickelt, welche die Interoperabilität und die Wiederverwendbarkeit von den Forschungsdaten gewährleisten sollen. Gesammelte, international Standards findet man beispielsweise auf der Seite FAIRsharing[3]: A curated, informative and educational resource on data and metadata *standards*, inter-related to *databases* and data *policies*.

Außerdem gibt es immer mehr Portale, die die Suche von Forschungsdaten ermöglichen:

- B2FIND Datensuche (http://b2find.eudat.eu/)
- Creative Commons CC Search (https://search.creativecommons.org/)
- DataCite Metadata Search (https://search.datacite.org/)
- DataONE Datenkatalog (https://search.dataone.org/#data)

Tools in den Naturwissenschaften

Tools für FDM in den Naturwissenschaften umfassen Repositorien, aber auch Versionsverwaltungssysteme. Eng vernetzt mit FDM-Systemen sind auch die Data Analytics Plattformen, welche ein gutes Forschungsdatenmanagement voraussetzen. Nachfolgend finden Sie eine Auswahl an Tools für FDM und Analytics, welche Anwendung in den Naturwissenschaften finden. Natürlich gibt es noch eine Vielzahl weiterer Anwendungen. Eine gute Quelle für die Suche von passenden Repositorien ist https://www.re3data.org/.

PANGAEA (Publishing Network for Geoscientific and Environmental Data)[4]

Daten in PANGAEA werden über das System SYBASE von SAP verwaltet. Importiert werden Daten über SYBASE, der Export erfolgt über verschiedene webbasierte Clients. Die Plattform bietet den Forschenden eine Langzeitarchivierung und einen DOI. Es werden etablierte Standards für Metadaten und Vokabel verwendet. PangaVista ist die Websuchmaschine, mit der man auf die in PANGAEA hinterlegten Daten Zugriff bekommt.

[3] https://fairsharing.org/
[4] https://www.pangaea.de/

NCBI (National Center for Biotechnology Information)[5]

NCBI bietet einen Zugang zu DNA-, RNA- und Proteindatenbanken an, sowie eine Taxonomie-Suchfunktion zur Suche nach Daten zu bestimmten Spezies, eine Literaturdatenbank (PubMed) und Standardsoftware in der Bioinformatik (z.B. Blast). Es ist die bedeutendste Anlaufstelle für Forschende aus der Molekularbiologie und Bioinformatik.

ICSD (Inorganic Crystal Structure Database)[6]

ICSD enthält Datensätze von Kristallstrukturen aus Wissenschaft und Industrie, hauptsächlich aus dem Bereich Materialwissenschaft. Zusätzlich kommen auch organische Strukturen, sowie bibliografische Daten auf das Repositorium. ICSD bietet eine Weblösung an, aber auch eine Desktopversion sowie ein Intranet als In-house Solution.

Für die Speicherung von Forschungsdaten, für die kein geeignetes fachliches Repositorium existiert, bieten sich institutionelle Repositorien**,** angeboten von einer wachsenden Zahl von Universitäten und Forschungseinrichtungen, oder auch generische Repositorien (z.B. Zenodo, Open Science Framework)**,** oft bereitgestellt durch zentrale Einrichtungen oder gemeinnützige Organisationen, an.

GitHub[7]

Git ist eine freie Software zur verteilten Versionsverwaltung. Besonderheiten von Git inkludieren (i) dass man Branching und Merging Funktionen für Projekte verwenden kann (nicht lineare Entwicklung), (ii) dass es keinen zentralen Server gibt, sondern jeder User eine lokale Kopie des Repositories hat und, (iii) dass bei jedem Commit (Revision) allen im Repository vorhandenen Dateien eine neue, für alle Dateien dieselbe Revisionsnummer zugewiesen wird.

CyVerse Austria[8]

CyVerse Austria ist eine FDM und Data Analytics Plattform, welche in Graz an den BioTechMed Universitäten (TU Graz, MedUni Graz, KF Uni Graz) installiert wurde. Diese Plattform bietet verteilte Speicherung von Forschungsdaten basierend auf der IRODS (integrated Rule-Oriented Data System) Technologie an. Damit haben alle teilnehmenden Universitäten ihre eigenen Speicherressourcen und können somit ihr FDM entsprechend der institutionellen Policies durchführen. Jeder Nutzer kann Daten hochladen, mit Metadaten versehen und mit KollaborationspartnerInnen teilen. Falls die KollaborationspartnerInnen nicht an derselben Institution tätig sind, verbleiben die Daten dennoch an den Speicherressourcen der ursprünglichen Institution und werden für den/die KollaborationspartnerIn nur ‚sichtbar' gemacht. Zusätzlich bietet die Plattform auch eine Analytics-Ebene. Dort werden reproduzierbare Datenanalysen und Workflows durch die Docker Technologie ermöglicht. Daten für rechenintensive Auswertungen auf High Performance Computing Clustern werden aus den FDM System zu dem Rechencluster transferiert, gerechnet und der Output wieder im FDM System abgelegt. Somit werden diese Prozesse für Forschende automatisiert.

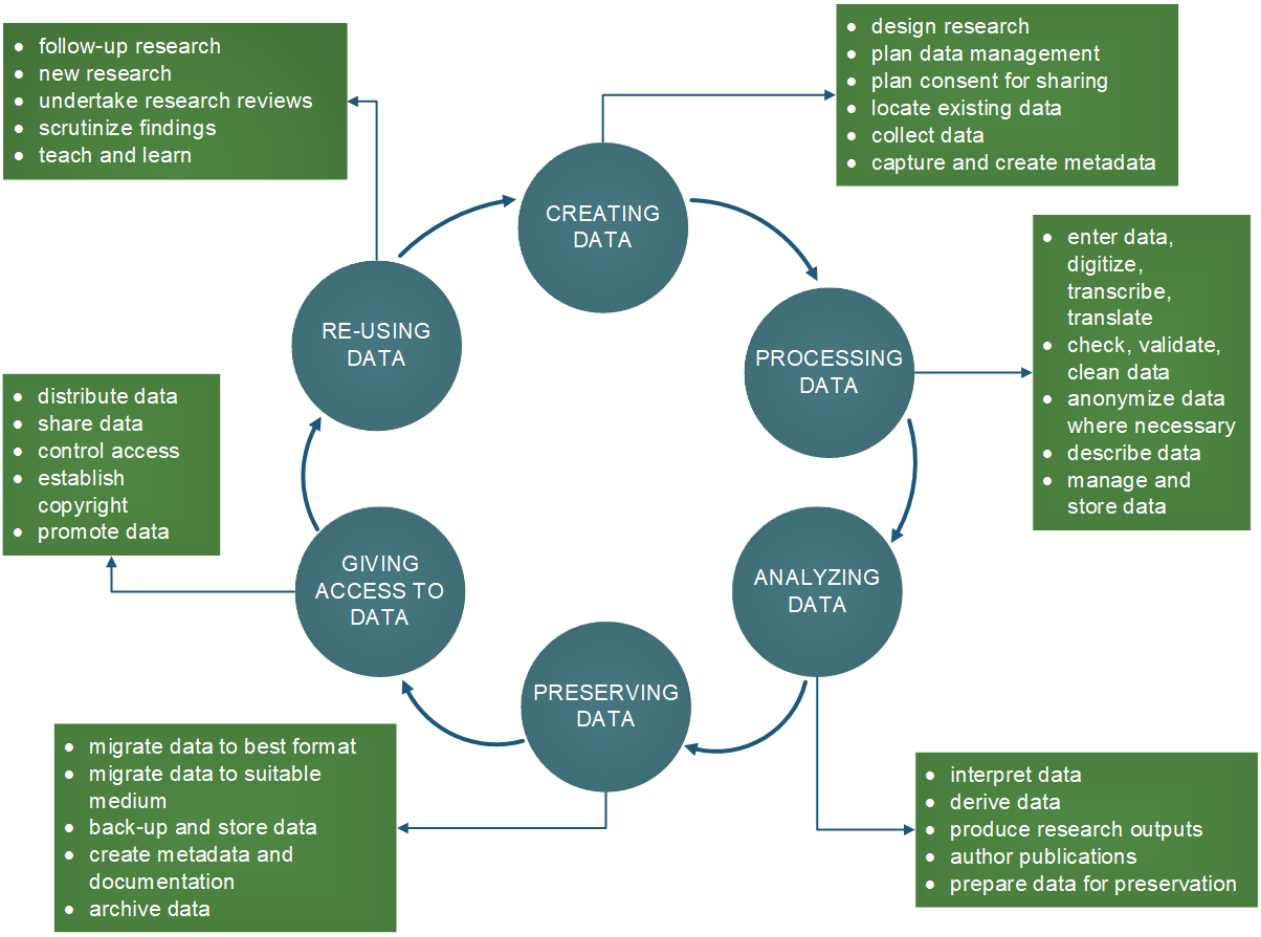
[5] https://www.ncbi.nlm.nih.gov/
[6] https://icsd.products.fiz-karlsruhe.de/
[7] https://github.com/
[8] https://cyverse.tugraz.at

Galaxy[9]

Galaxy ist eine Open Source Software für rechenintensive, bioinformatische Auswertungen. Die Plattform richtet sich vor allem an Forschende, die wenig bis keine Programmierkenntnisse haben. Sogenannte Workflows (aufeinanderfolgende Auswerteschritte) werden über eine grafische Oberfläche einfach und verständlich nähergebracht. Die Plattform bietet eine Vielzahl an fertigen Tools für Auswertungen im Bereich Genetik und Biomedizin. Außerdem können auf Galaxy eigene Daten hochgeladen, aber auch Daten anderer Quellen verwendet werden.

Data Stewardship – Modelle und Konzepte

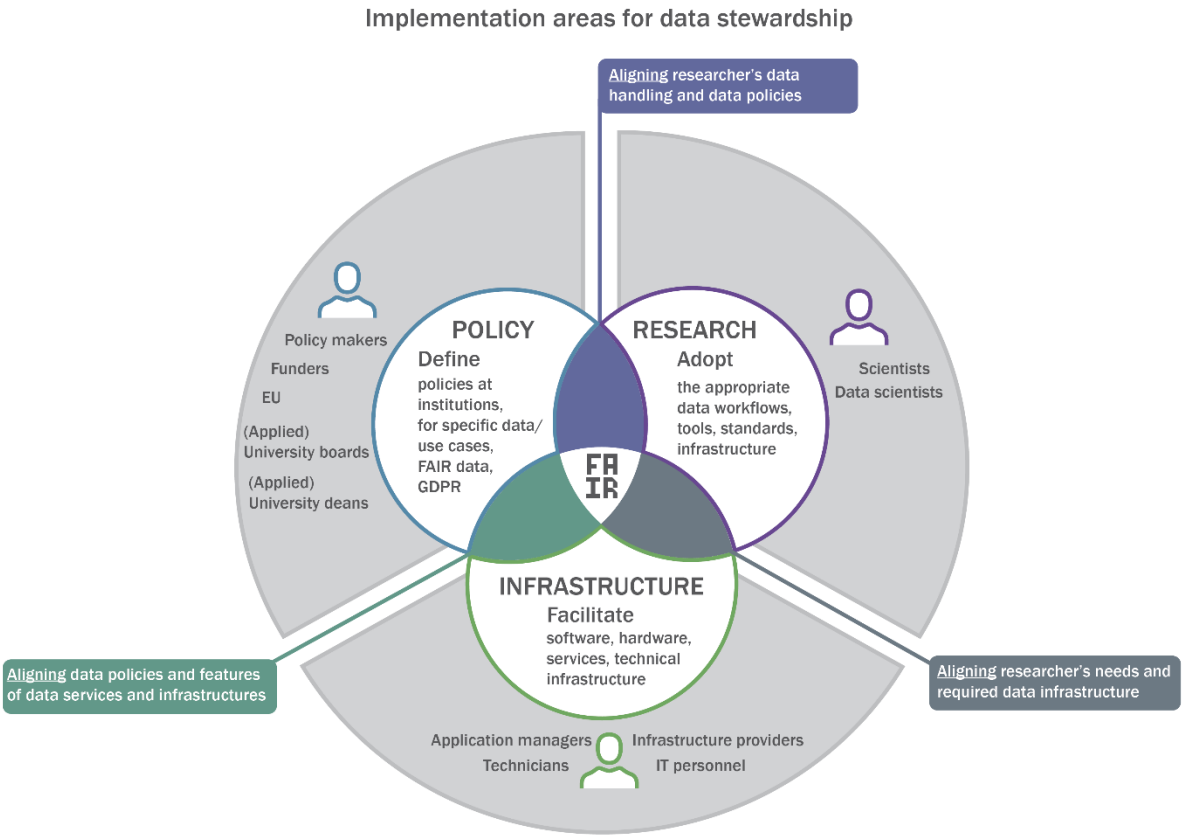
Data Stewards bilden eine neue Berufsgruppe. Sie sollen an Universitäten (und anderen Institutionen) die FDM Prozesse unterstützen. Da diese Berufsgruppe sich gerade erst entwickelt und noch nicht ganz klar definiert wurde, kann man die Aufgabenbereiche noch nicht ganz klar einordnen. Es ist aber angedacht, dass sich die Datastewards bei datengetriebener Forschung um die Aufgaben rund um den Lebenszyklus von Forschungsdaten kümmern werden. Jedoch ist das Ausmaß unklar (z.B. Data Analytics – eher beratende Funktion zu Tools und Softwarelösungen).



Quelle: https://libguides.mines.edu/c.php?g=949816&p=6850016

[9] https://galaxy.medunigraz.at/

Durch die Vielfalt der Aufgaben um den Lebenszyklus von Forschungsdaten, sowie übergreifende Aufgaben (z.B. Policies, Infrastruktur), ergeben sich verschiedene Profile von Data Stewards abhängig von der Institution und den vorhandenen Ressourcen/der vorhandenen Struktur. Diese Implementierungsbereiche der Datastewards sind zwischen den Bereichen Forschung, Infrastruktur und Policies angesiedelt.



Quelle: https://libguides.mines.edu/c.php?g=949816&p=6850016

Data Stewardship @ TU Graz

An der TU Graz ist es geplant ein Data Stewardship Programm einzuführen. Derzeit ist das FDM eingebettet in das Handlungsfeld Forschung im strategischen Projekt Digitale TU Graz. Im Rahmen dieses Projektes entwickeln wir Tools und Services um den gesamten Lebenszyklus von Forschungsdaten zu unterstützen. Es beginnt beim Niederschreiben der Forschungsidee (Proposal-Phase, DMP) über den aktiven Forschungsprozess (Data Management, Analytics) bis hin zur Publikation und finalen Archivierung.

In diesem Jahr haben wir Forschende mit einem Open Call motiviert ihre eigenen Projekte im Bereich FDM einzubringen. Ein paar Beispiele sind nachfolgend aufgelistet:

1. **Research Data Management system in Biomechanics from Experimental and Computational Aspects (Dr. Selda Sherifova)**

Forschende am Institut für Biomechanik generieren eine Vielzahl an experimentellen und computerbasierten Daten, um humane Gewebe zu modellieren. Dadurch können Krankheiten besser verstanden werden. In den letzten Jahren wurde klar, dass Forschungsdaten nicht mehr verwendet werden können, wenn ein Forschender das Institut verlässt. Grund dafür sind fehlende Dokumentation und Standards für die Archivierung. In dem Projekt werden (i) Bedürfnisse von Forschenden erhoben, (ii) Metadaten Standards definiert, (iii) Tools für FDM installiert und implementiert and (iv) Ergebnisse verbreitet.

2. **Data Sharing in der Mikrobiomforschung (Dr. Alexander Mahnert, Dr. Tomislav Cernava)**

In der Mikrobiomforschung ist Kollaboration der Schlüssel. Forschende profitieren sehr von Sequenzierdaten von anderen Forschungsgruppen. Forschende in Graz an der TU Graz und der MedUni Graz möchten ein kollaboratives Projekt mit gemeinsamer Datenanalyse von Sequenzierdaten durchführen, haben jedoch keine Möglichkeit für den sicheren Datentransfer.

In dem Projekt haben wir mit der CyVerse Austria Plattform einen Workflow definiert, bei dem die Forschenden ihre Daten mit ihrem Kollaborationspartner sicher teilen können. Mithilfe von Jupyter Notebooks haben wir einen automatisierten Data-Merging Prozess eingerichtet.

3. **Innovative Dokumentation mit elektronischen Laborbüchern (Dr. Werner Grogger)**

Das Institut für Elektronenmikroskopie an der TU Graz generiert täglich eine große Menge an Daten für unterschiedlichste Forschende. Üblicherweise werden Setup und Laborprozesse in Hardcopy Büchern notiert. Diese sind jedoch nicht automatisch durchsuchbar und können verloren gehen. Um die enorme Datenmenge in der Elektronenmikroskopie bestmöglich zu unterstützen wird eine Instanz für elektronische Laborbücher (elabFTW) installiert und Prozesse dafür definiert.