

Forschen mit den FDM-Landesdiensten: Eine Vision

Frauke D. Meyer ist Informatikerin an einer Mitgliedshochschule der DH.NRW und erhebt in Kooperation mit einer Arbeitsgruppe aus den Sozialwissenschaften für ihre Simulationen große Datenmengen. Seit sie Zugriff auf die FDM-Landesdienste in NRW hat, kann sie damit ihre Forschungsdaten noch effizienter und transparenter verwalten.



Die Hochschulcloud **sciebo** nutzt Frauke D. Meyer, um Dokumente zu organisieren. Dazu zählen beispielsweise Konzepte zum Onboarding neuer Mitarbeitenden oder Präsentationen über die Arbeit ihrer Arbeitsgruppe. Sciebo verfügt über ein integriertes Online-Office, das gängige Bearbeitungen von Textdokumenten, Präsentation und tabellarischen Daten unterstützt, und ermöglicht außerdem die Bearbeitung mit lokalen Office-Programmen. So wird das gemeinsame Bearbeiten und Teilen von Daten mit Kooperationspartner:innen und Kolleg:innen unterstützt. Für ein neues Projekt beantragt Meyer eine sogenannten Projektbox mit bis zu 2 TB zusätzlichem Speicher. Diese vereinfacht das gemeinsame Teilen und Bearbeiten von Dateien unterschiedlicher Formate. Meyer hat die Projektbox zwar beantragt, kann ihre Verwaltung jedoch nach Bedarf unkompliziert auf andere Personen übertragen. Daten in der Projektbox können zudem einfach mit Projektbeteiligten geteilt werden. Durch die Nutzung des sciebo-Klienten kann Meyer jederzeit sowohl von ihrem Laptop als auch ihrem Smartphone auf die Daten zugreifen. Der Zugriff per Webbrowser auf dem Computer im Institut ihrer Kooperationspartner:innen ist ebenfalls möglich, so dass sie jederzeit die neueste Version der Daten zur Hand hat.

Aus den Ideen von Frauke D. Meyer und ihrer Kolleg:innen ist ein gefördertes Forschungsprojekt geworden. Für das Projekt sollen die Daten mit einem eigenen Software-Code analysiert werden. Dafür kann sich Meyer bei **git.nrw** anmelden und erhält somit Zugriff zur NRW-weiten GitLab Instanz. Dort kann sie für ihre Forschungsdaten ein neues Projekt anlegen, sowie auch ihre internationalen Projektpartner:innen aus den USA und Japan als Gäste hinzufügen. Meyer kann mit git.nrw ihre Forschungsdaten und den dazugehörigen Software-Code versionieren. Das bedeutet, dass sie immer wieder auf Zwischenstände oder ältere Versionen zurückgreifen kann, wenn eine Änderung nicht das erwünschte Ergebnis hervorbringt oder kleine Fehler passieren. GitLab verfügt über ein integriertes Continuous Integration / Continuous Delivery Feature. Durch die Funktion CI/CD ist es möglich, Prozesse zu automatisieren. So kann die Forschergruppe mögliche Updates am Code automatisch integrieren, ältere Versionen des Codes einsehen und im Zweifel sogar auf einen vorherigen Stand zurücksetzen. Für ihr persönliches Datenmanagement kann Meyer außerdem die Projektmanagement-Tools in GitLab nutzen. Hierzu zählt ein Wiki, das sie bspw. mit Textbausteinen, Definitionen oder vorgeschriebenen Codes füllen kann. So hat sie alles an einem Ort, was sie für die Auswertung ihrer Daten und die Entwicklung neuer Codes benötigt. Zusätzlich bietet GitLab ein Ticket-Tool an, mit welchem sie

sich einen Überblick über die Aufgaben verschaffen kann. Die Tickets lassen sich mit Tags versehen und auf einem Kanban-Board oder einem Meilensteinplan darstellen. So hat Meyer stets einen Überblick über die nächsten zu erledigenden Aufgaben und kann ihre Überlegungen zu bereits abgeschlossenen Aufgaben im zugehörigen Ticket nachlesen.

Ihre Simulationen führt Frauke D. Meyer mit **JupyterHub.nrw** in einem Jupyter-Notebook aus. In der Notebook-Oberfläche werden Programmcodes und Visualisierungen von Daten und Ergebnissen nebeneinander dargestellt, so dass Meyer ihre Simulationen interaktiv weiterentwickeln kann, ohne zwischen verschiedenen Programmen wechseln zu müssen. Sie arbeitet bequem im Webbrowser und führt ihre Programme auf leistungsstarken Rechnern in der Cloud aus, egal ob sie gerade im Büro, im Home Office oder unterwegs auf Dienstreise ist. Da das Jupyter-Notebook Format standardisiert und universell verbreitet ist, kann Meyer ihre Programmcodes und Ergebnisse leicht mit Kolleg:innen weltweit teilen, auch über GitLab und ähnliche Dienste. Ihre Kolleg:innen können im Jupyter-Notebook die Ergebnisse betrachten und den enthaltenen Programmcode nachvollziehen und selbst ausführen. Verwenden die Kolleg:innen selbst JupyterHub.nrw, dann können sie dank der vorinstallierten wissenschaftlichen Softwarepakete und Arbeitsumgebung sofort mit dem Notebook arbeiten und müssen selbst keine Software installieren.



Damit Frauke D. Meyer den Machine-Learning-Algorithmus gut trainieren kann, braucht sie viele Daten – in Summe über 50 TB. Zum Glück muss Meyer keinen eigenen Objektspeicher kaufen, sondern sie kann den **FDM:Storage.nrw** nutzen. Über Coscine kann sie auf den **DataStorage.nrw** zugreifen und auf dieser großen und stabilen Plattform ihr Datenmanagement und die Datenablage organisieren. Zum Abschluss ihres Projekts werden ihre Forschungsdaten auf **DataArchive.nrw** übertragen und können für mindestens 10 Jahre sicher aufbewahrt werden. Meyer legt daher ihr Projekt in **Coscine** an und lädt die anderen PIs als Projekteigentümer:innen sowie ihre Doktorand:innen als Projektmitglieder ein. Ihre internationalen Projektmitglieder aus den USA und Japan kann Meyer dank des Logins via ORCID ebenfalls zum Projekt einladen und ihnen Zugriff auf die Daten ermöglichen. Um den kostenlosen Speicherplatz zu erhalten, muss Meyer ein Metadatenprofil für Ihre Dateien auswählen – nur so können ihre Dateien gemäß der FAIR-Prinzipien gespeichert werden. Sie wendet sich zur Unterstützung bei der Erstellung des passenden Metadatenprofils an NFDI4DataScience und wird durch einen Data Steward unterstützt. Nach der Erstellung des Metadatenprofils wird ihr Coscine Speicherplatz-Antrag über 50 TB bewilligt und Meyer lädt über ein Python Skript ihre Dateien inklusive Metadaten automatisiert über Coscine auf den DataStorage hoch. Andere Projektmitglieder mit weniger Programmiererfahrung und kleineren Datenmengen nutzen dafür einfach den Upload über die Coscine-Weboberfläche. Der projektbezogene Software-Code aus GitLab wird über Coscine ebenfalls allen Projektmitgliedern zur Verfügung gestellt und mit Metadaten beschrieben. Die von Coscine vergebenen projekt- und ressourcenspezifischen PIDs nutzt Meyer

zum Referenzieren ihrer Forschungsdaten außerhalb von Coscine – z.B. in Preprints. Nach Abschluss ihres Projekts speichert Frauke D. Meyer ihre Daten in Coscine für 10 Jahre. Dank der angegebenen Metadaten findet Meyer im Rahmen eines nachfolgenden Projekts passende Daten über die Suchfunktion von Coscine wieder und kann sie nachnutzen.

Im Laufe des Projekts hat Frauke D. Meyer ein Set von 250 GB Daten in Coscine zusammengestellt, das im Rahmen eines ersten Papers inklusive DOI publiziert werden soll. Da Coscine speziell für die aktive Projektphase konzipiert wurde, kann Frau Meyer ihre Daten dort nicht veröffentlichen. Sie kann jedoch über eine Schnittstelle in Coscine ihren Datensatz samt Metadaten direkt an das Forschungsdatenrepositorium **DataPublication.nrw** weiterleiten, so dass sie keine doppelten Eingaben vornehmen muss. Über DataPublication.nrw kann Frau Meyer ihre Forschungsdaten qualitätsgesichert veröffentlichen und automatisch mit einer DOI versehen – so werden Ihre Daten nach langfristig zitierbar. Der Dienst bietet für jede Hochschule einen eigenen, individuell konfigurierbaren Bereich und ist eng an die bestehenden Landesdienste angebunden. Nach der Veröffentlichung wird Frau Meyers Datensatz über DataCite, OpenAIRE und weitere internationale Nachweissysteme indexiert. Dadurch sind ihre Forschungsdaten für andere Forschende weltweit gut auffindbar. Die gewählte Lizenz macht transparent, unter welchen Bedingungen die Daten nachgenutzt werden dürfen; Urheberschaft und Herkunft bleiben jederzeit eindeutig nachvollziehbar. Mit der Veröffentlichung über DataPublication.nrw erfüllt Frau Meyer außerdem die Vorgabe ihres Forschungsförderers, die im Projekt erzeugten Forschungsdaten entsprechend den FAIR-Prinzipien bereitzustellen.

Um sicherzustellen, dass ein Teil der Daten auch in mehr als 10 Jahren noch für wissenschaftliche Forschung verfügbar ist, werden sie in die **Langzeitverfügbarkeit** (LZV) überführt. Dafür nimmt Meyer ebenfalls Kontakt zu ihrer Bibliothek auf, um den Transfer zu veranlassen. Da die Hochschulbibliothek ein Kooperationspartner der Landesinitiative LZV.nrw ist und das Repositorium an die LZV-Infrastruktur angebunden hat, wurden die im Repositorium veröffentlichten Daten automatisch in die LZV überführt. Meyer bespricht nun mit ihrer Kuratorin, welche der bisher unveröffentlichten Forschungsdaten aus der Projektphase ebenfalls langzeitverfügbar gehalten werden sollen. Nachdem die relevanten Daten identifiziert worden sind, prüft die Kuratorin, ob bereits jetzt Risiken hinsichtlich der Dateiformate vorliegen und ob alle notwendigen Metadaten wie bspw. Nutzungsrechte, signifikante Eigenschaften der Forschungsdaten und Strukturinformationen vorhanden sind. Nachdem Meyer der Bibliothek die benötigten Rechte zur LZV eingeräumt hat, transferiert die Bibliothek auch die unveröffentlichten Forschungsdaten in die LZV-Infrastruktur.

Und damit ist wieder ein Forschungsprojekt erfolgreich abgeschlossen. Aber Frauke D. Meyer hat schon wieder viele neue Ideen mit ihre Kolleg:innen auf sciebo gesammelt.

