NFDI4Earth

NFDI4Earth Report

RDM Education and Data Science Training

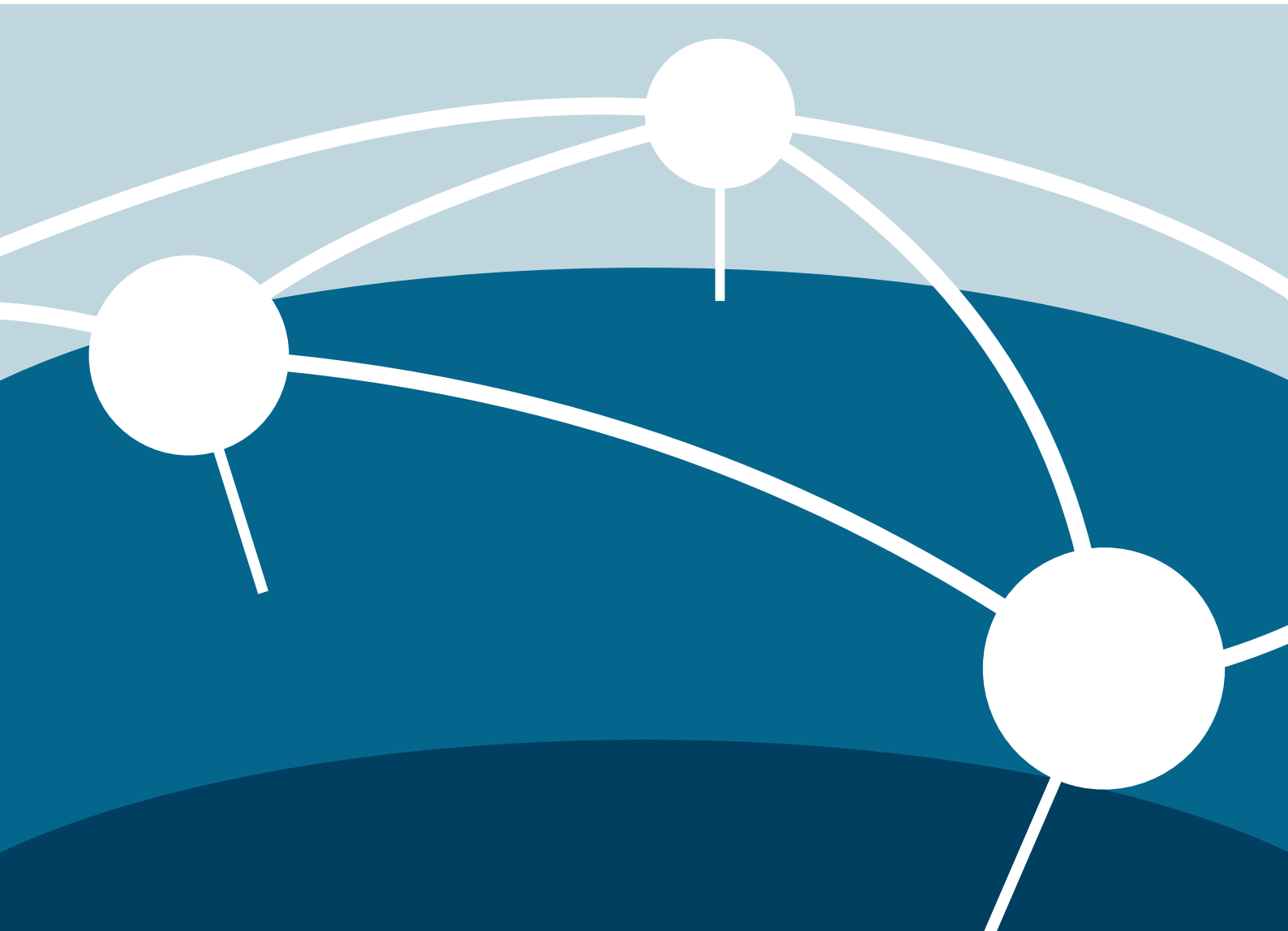
Joined report by NFDI4Earth Academy and NFDI4Earth EduTrain

Carsten Keßler (carsten.kessler@hs-bochum.de), Farzaneh Sadeghi, Hildegard Gödde (hgoedde@gfz-potsdam.de), Jonas Kuppler

2024-09

DOI: 10.5281/zenodo.13837959

nfdi4earth.de



Citation

Carsten Keßler, Farzaneh Sadeghi, Hildegard Gödde, Jonas Kuppler. 2024. *RDM Education and Data Science Training (NFDI4Earth Report)*. NFDI4Earth Community on Zenodo. https://doi.org/10.5281/zenodo.13837959

Acknowledgement

This work has been funded by the German Research Foundation (DFG) through the project NFDI4Earth (DFG project no. 460036893, https://www.nfdi4earth.de/) within the German National Research Data Infrastructure (NFDI, https://www.nfdi.de/).

Executive summary

NFDI4Earth supports training in the Earth System Sciences (ESS) by providing educational and training materials (NFDI4Earth EduTrain) and implementing systematic training approaches at the intersection of ESS and data science (NFDI4Earth Academy). This report briefly introduces work done by the EduTrain and Academy teams within NFDI4Earth as of August 2024, and presents an outlook on planned activities.

Contents

1 NFDI4Earth EduTrain

1.1 Overall Goals

The EduTrain (educational and training materials and services) team aims to provide NFDI4Earth users with access to high-quality educational materials and tools to practice open and reproducible research methods. EduTrain supports the Earth System Sciences (ESS) community by providing easy access to a curated collection of existing high-quality Open Educational Resources (OER) and developing OER requested by the community. We implemented a systematic approach to align services with the community's training needs.

In order to better match EduTrain's outputs with the community's needs, a comprehensive analysis of gaps in the current resources and services was carried out. This included identifying existing resources, mapping them to the required resources, and conducting a focus group interview with experts in ESS data education and training[1]. The OER identified during the gap analysis underwent an initial assessment to validate the credibility of the publisher, as well as the relevance and quality of the content delivery. The gap analysis identified five key areas requiring attention: findability, flexible curriculum, personalization, absence of mid-level-difficulty OER, and reusability.

1.2 Collection of Open Educational Resources

To enhance the findability of ESS OER, we set a minimum standard for educational resources published through NFDI4Earth's services, which includes at least seven mandatory metadata fields and nine optional fields[2]. The resources initially harvested for identifying the missing topics in ESS education were annotated with both the required and optional metadata using established vocabularies and standards, such as DFG subject areas[3] and Schema.org[4]. This process led to the creation of the initial version of NFDI4Earth's catalog of educational resources[5], accessible via the SPARQL endpoint of the NFDI4Earth Knowledge Hub[6].

The initial catalog comprised 90 OER sourced from 32 providers, focusing on Earth system data analysis and management. We have identified two categories of providers. The first category

[1] The findings are documented in the D1.3.1 report: https://zenodo.org/doi/10.5281/zenodo.7940194

[2] Further details can be found in D1.3.12: https://zenodo.org/doi/10.5281/zenodo.8345022

[3] https://www.dfg.de/en/about-us/head-office/programme-contacts#NaturalSciences

[4] https://schema.org/LearningResource

[5] https://onestop4all.nfdi4earth.de/search?resourcetype=Learning+Resource&sort=

[6] https://knowledgehub.nfdi4earth.de/

includes individual high-caliber content released under open licenses on platforms like GitHub, YouTube, or educators' personal websites. The second category covers entities consistently offering substantial amounts of quality-controlled content. Platforms such as EarthLab[7], MIT OpenCourseWare[8], and TIB[9] are listed for regular automatic metadata harvesting to ensure the catalog is continuously updated and expanded. The current collection offers access to approximately 500 high-quality open OER in ESS subjects and is consistently growing.

In order to effectively reuse resources from the OER catalog and tailor them to the educational needs of NFDI4Earth users, we follow the structure and learning objectives defined by the EduTrain curriculum. The curriculum aims to offer a clear framework for conducting open and reproducible data analysis and management. The format of the NFDI4Earth curriculum, originally intended to be a fixed document, has been transformed into a knowledge graph in response to the need for a curriculum that provides modularity and flexibility. This new format allows the curriculum to be broken down into smaller, interconnected parts that can be arranged in different ways. This transformation maintains the existing goals for the NFDI4Earth curriculum and introduces a modular design that facilitates the integration of new topics and technologies as they arise and makes it easier for users to explore the content in the order that best suits their needs rather than following a strict path. The curriculum covers a wide range of subjects, from the fundamentals of spatial data analysis and management to advanced data analysis techniques using open and reproducible methods and tools[10]. The NFDI4Earth curriculum comprises essential knowledge points and learning pathways for performing data analysis and management in the ESS using reproducible research practices. It forms the foundational framework for the NFDI4Earth Educational Portal (EduPortal)[11].

1.3 NFDI4Earth EduPortal

Building upon our OER collection and the EduTrain curriculum, we have developed the EduPortal to provide courses offering original content that fills gaps in current resources or provides unique learning paths not readily available elsewhere. While around 500 OER are accessible through the NFDI4Earth KnowledgeHub, the EduPortal currently features a curated selection of about ten courses. Each of these courses consists of multiple modules and integrates several resources. The courses on the EduPortal collectively reuse numerous OER, creating comprehensive learning resources that go beyond individual resources.

[7] https://www.earthdatascience.org/

[8] https://ocw.mit.edu/

[9] https://av.tib.eu/

[10] More information can be found in MS1.3.9: https://doi.org/10.5281/zenodo.13768718

[11] https://edutrain.nfdi4earth.de/

The EduPortal, which runs on the Open edX platform, offers modular courses that typically feature a module or a set of modules from the NFDI4Earth curriculum. These courses may also incorporate content utilized in EduTrain events or EduPilots. Using a Learning Management System (LMS) such as Open edX for course publication facilitates the transition toward personalized learning experiences. The process for choosing a LMS is outlined in D1.3.12[12].

EduPortal does not mandate users to complete any compulsory prerequisites to enroll in a course. However, recommended prerequisites are provided in the metadata displayed on the course home pages. These prerequisites are associated with courses that can offer the required knowledge; if such courses are not available, the subject becomes a priority in the development of new EduTrain courses. Courses typically consist of modules, and completion of a module is not obligatory before progressing to the next one. This allows users to plan their own learning paths to achieve their educational objectives, whether through the assistance of the NFDI4Earth curriculum or independently.

The courses are set to be self-paced, enabling users to learn at their own speed. We provide interactive Jupyter notebooks, essential for working with data, allowing students to engage in hands-on learning without facing installation barriers or dealing with insufficient computational power. Tab. 1 is the list of courses, along with their respective sources. Additional information can be accessed in MS1.3.4[13].

Table 1: List of courses developed or curated by EduTrain and published under an open license

Title	Material source	Status
Image processing and analysis	EduPilot	Public
How to create publishable netcdf-data	Educational event	Public
Introduction to Python	NFDI4Earth catalog	Public
Python for spatial data analysis	Educational event	Public
Introduction into chunking for large gridded datasets	EduPilot	Public
Fundamentals of spatial data mining and machine learning	EduPilot	Public
Fundamentals of spatial data formats for Earth data science	NFDI4Earth catalog	Under development
Introduction to open reproducible science	NFDI4Earth catalog	Public
Fundamentals of big spatial data	NFDI4Earth catalog	Under development

The courses available on EduPortal are published under open licenses to facilitate reusability. The default license for all course modules is the Creative Commons Attribution (CC BY) license[14].

[12] https://zenodo.org/doi/10.5281/zenodo.8345022

[13] https://doi.org/10.5281/zenodo.13769914

[14] https://creativecommons.org/share-your-work/cclicenses/

However, other Creative Commons licenses are employed when specified by the creators of the materials. Each course on the EduPortal has its own GitLab[15] repository linked to the course homepage, allowing for convenient download and broad community reuse. These repositories house material, course curriculum, and course metadata.

The guidelines for preparing course material for the EduPortal are available on NFDI4Earth GitLab. This living document is continuously updated to align with the evolving portal. It covers various topics such as the EduTrain metadata schema, templates for different CC licenses, course settings, modular course design recommendations, content formats, and adding various types of content to the EduPortal. While the portal and its content are public, publishing content on the portal is restricted. Educators interested in accessing the guidelines and the course authoring page must agree to basic terms and conditions, committing to providing proper credits, metadata for FAIR publication through NFDI4Earth channels, and using open licenses.

As of now, we have gathered comprehensive evaluations and feedback on the content of the following courses: ”Image Processing and Analysis,” ”Publishable NetCDF-Data Creation,” ”Fundamentals of Spatial Data Mining and Machine Learning,” and ”Python for Spatial Data Analysis.” Additionally, we are actively soliciting feedback on EduPortal through NFDI4Earth partners and promoting the portal at events such as the annual NFDI4Earth plenary and the EGU conference.

The feedback has been largely positive, but there have been some constructive criticisms that have proven valuable in identifying errors and outlining community-driven feature developments. Examples include requests for interactive notebooks and the inclusion of an anonymous feedback section at the conclusion of each course. All notable issues that could not be addressed immediately have been documented as GitLab issues to ensure nothing gets overlooked and to track and manage them until they are resolved.

1.4 NFDI4Earth EduPilots

The educational content used to cover curricular concepts is sourced from contributions made by NFDI4Earth partners, the reuse of OER in the EduTrain catalog, and funded Educational Pilots (EduPilots). The first call for EduPilots opened from June 9th to August 15th in 2022 and has since transitioned to a rolling format for better community engagement. This change has resulted in additional strong proposals from community members previously uninvolved with NFDI4Earth[16].

Each proposal undergoes a two-step review process. First, the EduTrain Leads review the format and completeness of the proposals. If mandatory information is missing, the applicant

[15] https://git.rwth-aachen.de/nfdi4earth/edutrain/content/extern

[16] https://www.nfdi4earth.de/?view=article&id=375&catid=15

is contacted. Complete proposals then move to the scientific review stage, where they are evaluated by NFDI4Earth reviewers who are familiar with the proposal topic. The current status of submitted EduPilots is detailed in Tab. 2. Accepted EduPilots that are not covered thematically with the NFDI4Earth curriculum are broken down into smaller modules. The relationship of each module to existing topics is then assessed, and learning pathways for each new topic are defined as part of our efforts to expand our curricula.

Table 2: list of EduPilot proposals

Title	Year	Status	Available on EduPortal
Introduction into chunking for large gridded datasets	2022	OER submitted	Yes
Teaching lead isotope geochemistry and application in archaeometry	2022	Incomplete submission	No
The future is urban, the data is smart	2022	OER submitted	In progress
An open-access and interactive coding platform to facilitate E-Teaching in Geospatial Data Analysis	2022	OER submitted	Yes
Climatematch: Computational Tools for Climate Science	2023	OER submitted	Yes
Image pre-processing, feature generation and classification in remote sensing	2023	OER submitted	In progress
Streamlining Atmospheric Science, Oceanography, Climate, and Water Research with For tran-accelerated R	2023	OER submitted	In progress
Artificial intelligence – basics and geographical applications	2024	In progress	No
Classification and change detection in remote sensing	2024	Accepted	No

2 NFDI4Earth Academy

2.1 Overall Goals

The NFDI4Earth Academy is dedicated to accomplishing three key goals. First, it aims to equip early career scientists with expertise in integrating Earth system and data science methods. Given the rapid evolution of computational methods and artificial intelligence, such interdisciplinary abilities are essential for addressing global challenges and generating novel insights.

Second, the Academy serves as a platform for networking, specialised training, and collaborative research. It fosters the mindset and skills crucial for conducting open and data-driven science among early career scientists and cultivates a collaborative and inclusive community. In this way, the Academy establishes a research culture grounded in openness, collaboration, and innovation.

Third, the Academy facilitates the integration of robust RDM practices into open and data-driven scientific workflows. Academy Fellows benefit from access to the full spectrum of NFDI4Earth innovations and services, enabling them to incorporate cutting-edge tools and technologies into their research projects. This not only aids in adopting FAIR and open RDM practices but also validates the utility of NFDI4Earth developments, which are tailored to the changing needs of active researchers.

2.2 Learning Environment and Training Approach

To achieve these goals, the NFDI4Earth Academy creates an agile and open environment, where Fellows play a pivotal role in shaping the content and direction of their educational and research pursuits. These bottom-up processes empower our Fellows to pursue their passions, fostering autonomy, curiosity, and self-confidence. We understand that moving away from traditional top-down structures may be challenging for those accustomed to more prescriptive educational environments. That's why our programme offers a balance between providing guidance and facilitating open exploration.

Through collaborative projects, peer mentoring, and specialised training courses, we create opportunities for interdisciplinary collaboration, enabling Fellows to leverage diverse perspectives and expertise to tackle complex scientific problems. As part of the NFDI4Earth community, Fellows have access to cutting-edge tools and technologies, ensuring their research remains at the forefront of innovation.

2.3 Recruitment and Selection Process

The Academy is looking for Earth system scientists with a background in data science or statistics, as well as data scientists or statisticians with a deep interest in ESS. Both doctoral and postdoctoral researchers are eligible, provided that they have a minimum one-year contract at a German institution from the Academy's starting point. It is desirable for candidates to have experience in interdisciplinary research, strong ESS skills, enthusiasm for working with data, Research Data Management (RDM), data analysis, mathematics, coding (Matlab, R, Python, etc.), as well as independent thinking, intrinsic motivation, and clear research goals. A further priority is the motivation to actively contribute to and participate in the Academy's activities.

So far, the NFDI4Earth Academy published two open calls for the NFDI4Earth Academy. For the first call, published in July 2022[17], we received 81 applications and selected 39 Fellows[18]. For the second call, published in October 2023[19], 42 applications and selected 34 Fellows[18]. In both cohorts, we selected Fellows from a broad range of research fields (Tab. 3).

Table 3: Research fields (orientated to the DFG research classification) of selected Fellows per cohort

Research field	1st cohort # Fellows	2nd cohort # of Fellows
Atmospheric Science	7	2
Geodesy, Photogrammetry, Remote Sensing, Geoinformatics, Cartography	6	7
Geology & Paleontology	3	1
Geophysics	4	6
Geotechnology	1	0
Human Geography	1	1
Mineralogy, Petrology and Geochemistry	2	1
Oceanography	0	4
Paleoclimatology	1	0
Physical Geography	2	1
Water Research	5	3
Other research fields	7	8
Civil engineering	0	1
Computer Science	1	0
Ecology	0	3
Environmental Science	1	0
Environmental toxicology	1	0

[17] NFDI4Earth Deliverable D1.4.1. https://doi.org/10.5281/zenodo.7908328

[18] NFDI4Earth Deliverable D1.4.2. https://doi.org/10.5281/zenodo.7908364

[19] NFDI4Earth Deliverable D1.4.6. https://doi.org/10.5281/zenodo.10402422

Research field	1st cohort # Fellows	2nd cohort # of Fellows
Forest Ecology	0	1
Information Technologies	0	1
Landscape ecology & Agriculture Sciences	1	2
Mathematics/Physics	1	0
Neural- and Behavioural Sciences	1	0
Software Engineering	1	0

2.4 Curriculum and Training Events

The Academy programme deviates from traditional graduate school curricula in two ways: First, we emphasise peer mentoring over teacher-based learning. Secondly, we provide a core event structure with three events per year, determined by the Fellows in an agile bottom-up process to ensure that it meets the Fellows' training needs. In addition to the Academy core events, Fellows have the opportunity to participate in additional workshops, meetings, and other events by the NFDI4Earth and others (non-core events). There is also a special focus on connecting Fellows to existing training opportunities, especially regarding RDM topics and the strong integration and linkage to all NFDI4Earth developments.

Each Academy core event comprises a blend of scientific input, collaborative exchange, and practical implementation, with the specific balance of content varying based on the topic of the respective event[20].

So far, the NFDI4Earth Academy has organized twelve Academy core events for cohort 1 and cohort 2 (Tab. 4). In addition to these core events, we provided regular online Coffee Lectures (first and fourth quarter of each year) and organized parts of the NFDI4Earth Plenary based on the Fellows needs (Tab. 5).

Table 4: NFDI4Academy core events

Event	Date	Format	Content	Community	Particpants
Year 1/cohort 1					
Kick-off	30.11 -02.12.2022	Retreat	Group building scientific exchange Academy planning[21]	Academy Fellows	34

[20] NFDI4Earth Deliverable D1.4.7. https://doi.org/10.5281/zenodo.12806503

[21] The findings are documented in the D1.3.1 report: https://zenodo.org/doi/10.5281/zenodo.7940194

Event	Date	Format	Content	Community	Particpants
Cohort Calls	10.02.2023	Online Workshop	NFDI4Earth Pilots Introduction RDM[17]	Academy Fellows	28
Cohort Calls	24.02.2023	Online Workshop	Jupypter Notebooks Quarto[17]	Academy Fellows	26
Cohort Calls	10.03.2023	Online Workshop	Research software development Docker for Open Science[17]	Academy Fellows	26
Cohort Calls	06.04.2023	Online Workshop	Research Data & MetaData[17]	Academy Fellows	25
Cohort Calls	21.04.2023	Online Workshop	Data management plans Git & GitLab[17]	Academy Fellows	25
Summer School	12.-14.06.2023	Summer School	Introduction Machine Learning Spatial & Temporal data analysis Physics-informed ML[17]	Academy Fellows	29
DataXplorers cross-community hackathon	23.10.-06.12.2023	Online Hackathon in-person final event	Various challenges from the NFDI4Earth, NFDI4Biodiversity and NFDI4Microbiota[17]	open	32
Think Thank	06.-07.12.2023	Workshop	Idea generation, Thematic discussions, Shared interest[17]	Academy Fellows	15
Year 1/cohort 2					
Kick-off	10.-12.06.2024	Retreat	Group building, scientific exchange, Academy planning[18]	Academy Fellows	26
Cohort Calls	03.07.2024	Online Workshop	Introduction RDM[18]	Academy Fellows	15
Cohort Calls	10.-11.07.2024	Online Workshop	Version Control Git & GitLab[18]	Academy Fellows	20

Table 5: Additional NFDI4Earth Academy events

Event	Date	Format	Content	Community	Participants
Coffee lecture	19.12.2022	Online lecture	Introduction DataTrain[22]	Academy Fellows	28
Coffee lecture	24.08.2023	Online lecture	Technical issue ML[17]	Academy Fellows	12
Coffee lecture	08.09.2023	Online lecture	NFDI4Earth IG High-performance computing IG Long-term Archiving[17]	Academy Fellows	7
Coffee lecture	22.09.2023	Online lecture	Software Licensing[17]	open	16
Coffee lecture	20.10.2023	Online lecture	PANGAEA[23]	open	37
Coffee lecture	22.11.2023	Online lecture	Planning Think Tank[17]	Academy Fellows	13
Coffee lecture	12.01.2024	Online lecture	GFZ Data Services[24]	open	80
Coffee lecture	21.02.2024	Online lecture	GEO-INQUIRE[25]	open	24
Coffee lecture	08.03.2024	Online lecture	Causal inference[17]	open	52
Coffee lecture	22.03.2024	Online lecture	Explainable ML[17]	open	105
Coffee lecture	12.04.2024	Online lecture	Helmholtz DataHuB[26]	open	35
Coffee lecture	26.04.2024	Online lecture	FID Karten[27]	open	26
NFDI4Earth Plenary	01.06.2023	Conference	Academy Get-together[17]	open	NA
NFDI4Earth Plenary	22.-24.05.2024	Conference Workshop	Academy Get-together Speeddating & Poster session ChatGPT/FAIR GPT[17]	open	NA

[22] https://www.bremen-research.de/data-train/

[23] https://www.pangaea.de/

[24] https://dataservices.gfz-potsdam.de/portal/

[25] https://www.geo-inquire.eu/

[26] https://helmholtz.software/projects/datahub

[27] https://kartographie.staatsbibliothek-berlin.de/en/

3 Outlook

Our strategic direction for EduTrain and the NFDI4Earth Academy is to continuously develop both structures in mutual alignment to maximize potential synergies and meet the evolving needs of the Earth system science and the NFDI communities.

In **EduTrain,** the continuous maintenance and growth of the EduPortal will be critical to ensure its relevance and usability. This will involve regular updates to the content and platform, addressing technical issues, and improving the user experience. Additionally, periodic reviews of the curriculum and the technological infrastructure will be essential to keep pace with the latest advancements in Earth System Sciences and educational technology. Enhancements will be driven by data analytics and user feedback, targeting areas such as user interface design, content delivery formats, and the incorporation of emerging topics and technologies that reflect the current and future needs of the ESS community and, in particular, Academy fellows.

The integration of Large Language Models (LLMs) for metadata management is a promising direction. These advanced models will be used to automate the extraction and generation of metadata, enhancing the accuracy and richness of metadata associated with OERs. By leveraging LLMs, we anticipate a significant improvement in metadata standardization and the facilitation of semantic searches, ultimately increasing the accessibility and findability of educational resources.

The development of automatic learning pathways represents a transformative direction for EduTrain. Employing AI and machine learning algorithms, the system will recommend personalized learning trajectories to users based on their prior knowledge, learning preferences, and learning goals. This adaptive learning approach aims to streamline the educational experience, making learning more efficient and tailored to individual needs.

Just like data and other research objects, Open Educational Resources need to be Findable, Accessible, Interoperable, and Reusable to unleash their full potential. The EduTrain team is therefore focusing on FAIR OERs and concentrates its future efforts on improving the interoperability and reusability of OERs by expanding integrations with other OER platforms and services.

The **NFDI4Earth Academy** will increasingly integrate EduTrain's course offerings and automated learning paths into its programme. The ongoing dialogue between the Academy and EduTrain ensures that Fellows can benefit from various courses tailored to their specific needs. Conversely, EduTrain will receive the necessary feedback to ensure that its programmes are practicable and meet the needs of early career academics.

In addition, relevant Academy events will be added to the EduPortal as Open Educational Resources, thus broadening the range of educational content and providing fellows with a blended learning experience. The Academy will significantly benefit from self-learning courses

on FAIR and open research data management while maintaining its strong focus on Data Science.

Building on the success of the Academy concept within NDFI4Earth, we have recently proposed a model for an open and innovative training structure across all NFDI consortia in a white paper: *NFDI4Earth Academy – A model for a joint NFDI Academy*[28]. The envisioned NFDI Academy aims to empower the next scientific generation for interdisciplinary research and promote a collaborative research culture to drive impactful discoveries and address complex challenges at the intersection of different scientific disciplines and data science.

[28] See https://doi.org/10.5281/zenodo.13684199