Training concepts in research data management and data science with the focus on health research

AUTHOR LIST	CONTACT
Jens Dierkes Universitäts- und Stadtbibliothek Köln	https://orcid.org/0000-0002-0121-9261
Julia Fürst ZB MED Information Center Life Sciences	https://orcid.org/0000-0003-2547-933X
Tanja Hörner University of Bremen	https://orcid.org/0000-0003-3280-6941
Sebastian Klammt KKS-Netzwerk e.V.	https://orcid.org/0000-0001-7852-4769
Birte Lindstädt ZB MED Information Center Life Sciences	https://orcid.org/0000-0002-8251-1597
Iris Pigeot Leibniz Institute for Prevention Research and Epidemiology – BIPS	https://orcid.org/0000-0001-7483-0726
Katja Restel Universitäts- und Stadtbibliothek Köln	https://orcid.org/0009-0003-7583-2303
Carsten Oliver Schmidt Institute for Community Medicine, University Hospital Greifswald	https://orcid.org/0000-0001-5266-9396
Dagmar Waltemath Medical Informatics Laboratory, University Medicine Greifswald	https://orcid.org/0000-0002-5886-5563
Atinkut Zeleke Medical Informatics Laboratory, University Medicine Greifswald	https://orcid.org/0000-0001-7838-9050

for NFDI4Health

NFDI4Health has received funding from the Deutsche Forschungsgemeinschaft (DFG) under Grant Agreement no. 442326535. Data Train is supported by the U Bremen Research Alliance and the Federal State of Bremen.

CONTACT	
NFDI4Health	www.nfdi4health.de
Jens Dierkes	dierkes@ub.uni-koeln.de

CONTACT	
Design, setting, layout	Alina Esken / www.alinaesken.de
Language Editing	Andrew Rennison / andrewrennison@gmail.com
Date of publication	TBD
DOI	10.4126/FRL01-006441348

Contents

List of tables

List of figures

Abbreviations

CRF	Case report form
CTIS	Clinical Trials Information System
de.NBI	German Network for Bioinformatics Infrastructure
DINI	Deutsche Initiative für Netzwerkinformation e.V., German Initiative for Network Information
DMP	Data Management Plan
DRKS	German Clinical Trials Register
DS	Data science
ECRIN	European Clinical Research Infrastructure Network
EDC	Electronic Data Capture
ELIXIR	European life science infrastructure for biological information
EOSC	European Open Science Cloud
ESFRI	European Strategy Forum on Research Infrastructures
EUCTR	European Union Clinical Trials Register
EUDAMED	European database on medical devices
GCP	Good Clinical Practice
GDPR	General Data Protection Regulation
GEP	Good Epidemiological Practice
GPD	Good Practice Data Linkage
eHR	Electronic health record
EMR	Electronic medical record
FAIR	Findable, Accessible, Interoperable, Reusable
FHIR	Fast Healthcare Interoperability Resources
ICH	International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use
ICMJE	International Committee of Medical Journal Editors
IEC/ IRB	Independent ethics committee/ Institutional review board
JSON	JavaScript Object Notation
KKS/ZKS	Coordination Centres for Clinical Trials/ Centre for Clinical Trials
(Academic)	
LOINC	Logical Observation Identifiers Names and Codes
MedDRA	Medical Dictionary for Regulatory Activities
NCA	National competent authority

Abbreviations

nestor	German competence network for long-term archiving and long-term availability of digital resources
NISO	National Information Standards Organization
RBQM	Risk-based quality management
RDF	Resource Description Framework
RDM	Research Data Management
SAE	Serious adverse event
SAP	Statistical analysis plan
SNOMED	Systematized Nomenclature of Medicine
SOP	Standard operating procedure
UAC	Use and Access Committee
XML	Extended markup language
WHO	World Health Organization

1 Introduction

The digital transformation taking place across all sectors and the associated increase in the importance of research data in gaining knowledge or innovating new technologies poses enormous challenges for the workforce, policy makers, funders, educational institutions, etc. In particular, it is essential to equip everyone with the skills or skill sets that are required (i) to handle the many facets of data, (ii) to understand their various constraints, (iii) to be able to make informed decisions about the appropriate use of available technologies, and (iv) to gain "wisdom" from the data. It is therefore crucial to start teaching basic skills while at school and to deepen them as early as possible, e.g. in academic training programmes. The field of data literacy tackles these issues and is an attempt to systematise tasks and corresponding skills by way of one or more frameworks. According to Ridsdale et al. (2015), *"Data literacy is the ability to collect, manage, evaluate, and apply data, in a critical manner."*

TABLE 1: DATA LITERACY COMPETENCIES ACCORDING TO RIDSDALE ET AL. (2015).

KNOWLEDGE DOMAIN	COMPETENCIES
Conceptual Framework	Introduction to Data
Data Collection	Data Discovery and Collection Evaluating and Ensuring Quality of Data and Sources
Data Management	Data Organisation Data Manipulation Data Conversion (from format to format) Metadata Creation and Use Data Curation, Security, and Reuse Data Preservation
Data Evaluation	Data Tools Basic Data Analysis Data Interpretation Identifying Problems using Data Data Visualisation Presenting Data Data-driven Decision-making
Data Application	Critical Thinking Data Culture Data Ethics Data Citation Data Sharing Evaluating Decisions based on Data

Table 1 lists the main knowledge domains and key skills in Ridsdale et al. (2015). This broad approach is targeted at the general population of the 21st century. In the context of research, certain aspects naturally require a deeper understanding and application expertise (knowledge dimensions) along with domain knowledge. These can be identified by looking at the *research data life cycle.*

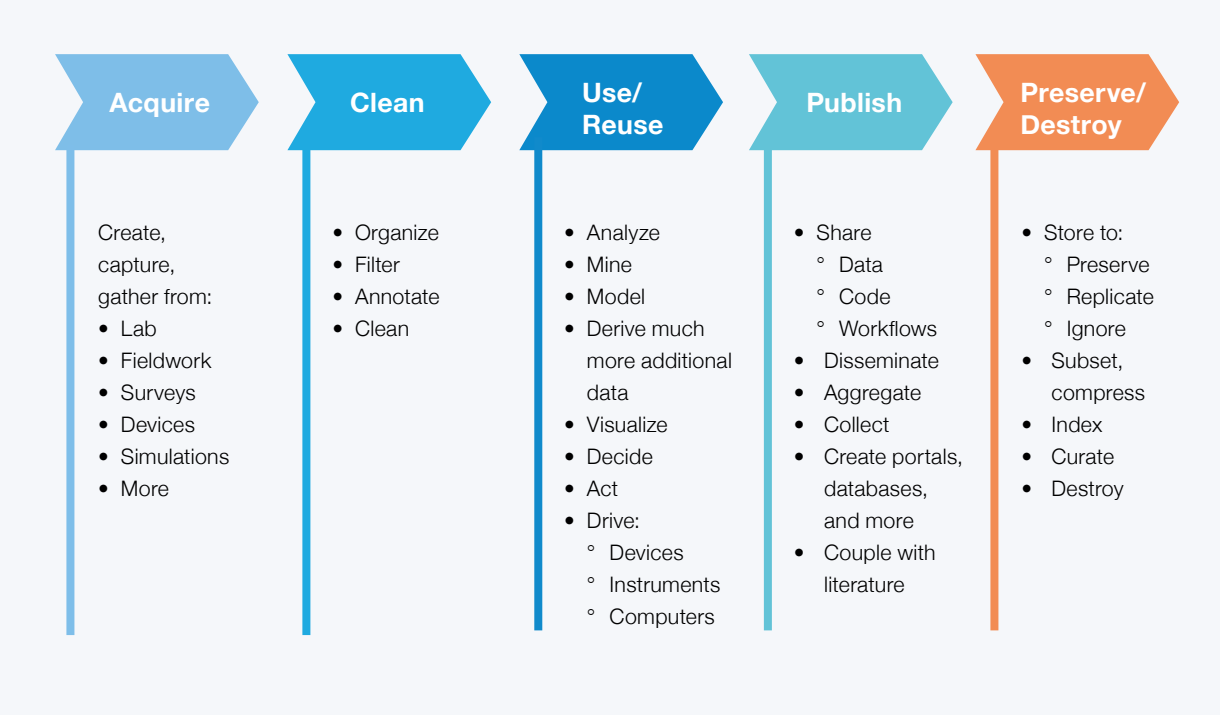

Figure 1: Data life cycle and data ecosystem (Berman et al. 2018)

Figure 1 depicts the main phases: *Acquire, Clean, Use/Reuse, Publish, and Preserve/Destroy.* All of these phases include overarching aspects such as management, organisation, and policies as well as law and ethics, financing, funding, metadata, and identifiers (e.g. Michener 2015). The basic data management concepts apply to all disciplines. The importance of certain aspects is weighted differently in individual disciplines. For instance, in clinical trials, epidemiological, and public health research, person-related data, data protection, and ethical issues play a crucial role.

The above skills and knowledge domains vary across disciplines, as do the roles of the people involved in the data value chain. In particular, the research process is increasingly becoming a collaborative enterprise, and there are several actors involved in the production of knowledge: professors, senior researchers, young researchers, doctoral, bachelor/master students, data stewards, laboratory technicians, IT engineers, librarians, etc. Depending on good practices of the discipline as well as the local environment, the actors share tasks and responsibilities throughout the data life cycle. Such networks have expertise and varying levels of competence located in different places and with different degrees of sustainability.

Given the importance of data in the production of knowledge based on what we described above, it comes as no surprise that data are considered as the new currency of the 21st century. It is therefore imperative that data are shared with the communities to realise their full potential.

A major challenge arising in connection with the aspects outlined above is to build up an infrastructure that allows research data to be shared both within and across disciplines. This not only requires findability of and access to data, but also data interoperability to ensure that data are reusable. Put simply, research data have to be provided to a wider community according to the FAIR principles (F = findable; A = accessible; I = interoperable; R = reusable; Wilkinson et al. 2016). Building up such a research infrastructure is a major endeavour requiring collaboration by researchers, data managers and computer scientists from all disciplines. In Germany, this endeavour has been taken up by an initiative of the German Federal and State governments (German Joint Science Conference) based on recommendations of the

German Council for Scientific Information Infrastructures (RfII) with the aim of establishing a National Research Data Infrastructure (NFDI). The aims of the NFDI as summarised by the DFG[1] are (1) to systematically manage scientific and research data, (2) to provide long-term data storage, backup and accessibility, (3) to network data both nationally and internationally, and (4) to bring multiple stakeholders together via a coordinated network of consortia, and provide science-driven data services to research communities where the long-term establishment of the NFDI as the German contribution to the European Open Science Cloud (EOSC) is envisaged as the ultimate goal.

To achieve the four aims listed above, the decision was taken to fund a total of 27 consortia in three rounds, initially for five years with the option to apply for another five years. Most of these consortia are discipline-specific so as to establish a research infrastructure that considers the specific needs of their respective research and user community. For instance, sharing sensitive person-related health data has to comply with privacy regulations and ethical requirements which have to be carefully considered by the National Research Data Infrastructure for Personal Health Data (NFDI4Health) consortium, but which do not play a major role for most of the other consortia. However, despite their different foci, the 27 consortia work together closely under the auspices of the registered association NFDI[2]. All of the consortia are members of the NFDI, which consists of various sections devoted to overarching topics, e.g. the Edu-Train section on training and education.

NFDI4Health was among the first nine consortia to receive funding during the first round of funding in October 2020. NFDI4Health aims (1) to enable findability of and access to structured health data from registries, administrative health databases, clinical trials, epidemiological studies and public health surveillance; (2) to implement a health data framework for a centralised search for and access to existing decentralised epidemiological and clinical data infrastructures; (3) to facilitate data sharing, record linkage, harmonised data quality assessments, federated analyses of personal health data, while complying with privacy regulations and ethical requirements; (4) to enable the development and deployment of new machine-processable consent mechanisms and innovative data access services by operationalising the FAIR principles for scientific data management and stewardship; (5) to support cooperation between clinical trial research, epidemiological and public health communities; and (6) to foster interoperability of currently fragmented IT solutions related to metadata repositories, cohort browsing, data quality and harmonisation. In addition, NFDI4Health places major emphasis on training the next generation of researchers and data stewards in research data management (RDM) as well as specific aspects of data science (DS). NFDI4Health already pursued this goal during the planning phase of the consortium and reached out to other potential consortia.

Here, NFDI4Health made use of an existing research network in Bremen, the U Bremen Research Alliance (UBRA) which consists of the University of Bremen and twelve federal and state financed non-university research institutes of the four major German science organisations as well as the German Research Center for Artificial Intelligence, where several institutes are involved in at least one NFDI consortium. Thus, UBRA offered a unique framework to meet the massive demand in science and economy by developing a cross-discipline training programme called “Data Train – Training in Research Data Management and Data Science” that teaches competencies in data literacy, research data management, and data science for doctoral researchers. The development of Data Train was supported by NFDI consortia involving UBRA. After a short pilot phase in 2020, Data Train was launched in 2021 and started its second round in 2022.

[1] https://www.dfg.de/foerderung/programme/nfdi
[2] https://www.nfdi.de/

As Data Train (for more details, refer to Chapter 2 of this handbook) may be considered a kind of blueprint for other comparable activities (e.g. Garbuglia et al. 2021), it will be presented here as a starting point for developing a training programme that focuses on health research. For this reason, this handbook describes the basic idea and modular structure of Data Train, including illustrations of the different components. Since Data Train is a generic concept, it must be adapted where the focus is on the specific requirements of health data. Such adaptation should consider the learning objectives as provided, e.g., by the DINI/nestor working group for research data[3], but with specifications for health research as addressed in Chapter 3.

1.1 Scope

After having briefly introduced the objectives and target groups as well as the structure and scope of the handbook, Chapter 2 provides an in-depth description of the "Data Train - Training in Research Data Management and Data Science" model, including the basic concept of the programme, training profiles, schedule, and evaluation modalities. Data Train aims to enhance the skills of young scientists in data literacy for RDM and DS, and to enable doctoral researchers to network across disciplines and institutions. Chapter 2 is the core section of the handbook, covering most of the generic RDM and DS content. Chapter 3 mainly focuses on the RDM learning objectives of the DINI/nestor working group and includes remarks on competencies requiring special attention for the specific use cases of NFDI4Health. Chapter 4 compiles further specific NFDI4Health RDM training courses designed to introduce RDM concepts and NFDI4Health services to the user community. Finally, the last chapter provides a summary and outlook for the handbook.

This handbook delivers an overview of RDM and FAIR principles based on practical implementation of generic competencies and lesson plans. The training is aimed at master students, doctoral researchers and professionals (postdocs and above) in the field of biomedical sciences and may also be of interest to trainers and coordinators looking to implement RDM teaching or training in line with FAIR principles, both in generic and healthcare contexts. As mentioned, the book focuses initially on domain independent, generic RDM and DS concepts that are applicable across disciplines. Once these concepts have been addressed, special attention is given to the specific needs of the NFDI4Health domain.

It is important to ensure that the content of the Data Train model and other NFDI4Health training courses covers the necessary knowledge and skill sets for the relevant profiles of the target group. There may be constraints in implementing such courses, particularly when it comes to addressing the specific RDM needs of the healthcare domains. Gathering user requirements iteratively will help identify these needs and develop appropriate competencies and training courses tailored to the specific needs of the healthcare domain while providing the necessary skills and knowledge to support successful research.

Consequently, this is planned to be a living document that should be updated in line with the needs and requirements of domain experts and users of the handbook. We anticipate making iterative changes at regular intervals to improve the document.

[3] Deutsche Initiative für Netzwerkinformation (DINI, German Initiative for Network Information) and the network of expertise in long-term storage of digital resources (nestor); https://dini.de/ag/dininestor-ag-forschungsdaten/

2 Data Train – a model for a training programme in research data management and data science

2.1 Basic concept of the programme

As already mentioned above, data-driven science is becoming increasingly important in answering the pressing research questions of our time. There is a clear lack of persons in Research Data Management (RDM) and Data Science (DS) who are qualified to foster innovative "Big Data" technologies for academic science and the private sector worldwide.

The Data Train training programme is tailored to the demands of doctoral researchers across the various institutes and disciplines, but open to staff interested in strengthening their data skills if they can free up the time resources needed to do so. Data Train therefore pursues the mission of strengthening basic competencies in data literacy, RDM, and DS, while simultaneously offering doctoral researchers a platform to build an interdisciplinary and cross-institutional network (see Hörner et al. 2021). In 2021, the programme was launched with a total of 40 lecturers, 29 lectures and 222 doctoral researchers at the U Bremen Research Alliance (UBRA) completing the course. Moreover, the programme was offered virtually, which enabled a broader audience to participate and resulted in more than 1,600 module participations. From 2022 onwards, the programme will be offered annually.

Data Train covers the entire data-value chain in a cross-institutional and interdisciplinary training course. Here, researchers not only learn about innovative approaches and technologies, but also get the opportunity to build their own networks at an early stage in their career. This exchange imparts new perspectives and methodological approaches while increasing creativity and inventiveness to maximise the potential of advanced research. The target group will transfer the acquired data competencies into the research sector as well as into the economic landscape. Therefore, the programme significantly contributes to data literacy training far beyond the borders of Bremen. Spreading the concept via the various National Research Data Infrastructure (NFDI) consortia, thus implementing such training in other locations, will help drive the cultural change towards *Open Science and Open Data.*

2.2 Development phase

The development of the basic concept and the curriculum was monitored by the Research Data Working Group of UBRA, which includes representation by UBRA member institutions, NFDI consortia along with RDM and DS initiatives. After an intensive review of existing training concepts and curricula (e.g. position paper on DS learning content by the Society of Computer Science (Society of Computer Science, 2019), the FAIRsFAIR teaching and training handbook for higher education institutions (Engelhardt et al. 2022), and the EDISON DS competence framework (Cuadrado-Gallego & Demchenko 2020)), a qualitative survey was carried out in 2020 to assess the status quo and training requirements among participating institutions and initiatives. We conducted more than 70 guided interviews to assess the competence requirements in RDM and DS from different status groups and sectors (management, infrastructure, and research). The focus was on (senior) scientists from all scientific disciplines at UBRA, representatives of the research data repositories located in Bremen (PANGAEA, QUALISERVICE), and coordinators of graduate programmes.

Based on this qualitative survey, a quantitative survey was developed to complement the more narrative information from the qualitative interviews with respect to the extent and range of the required teaching components. For this purpose, more than 6,000 persons including researchers (doctoral researchers, postdocs and professors) and infrastructure personnel were approached and more than 400 questionnaires were filled in.

2.3 Data Train model

Requirements (R): The training should provide researchers (R1) with basic knowledge covering both competency areas, RDM and DS. It is offered in addition to structured doctoral programmes on a voluntary basis. Therefore, the limited time resources of doctoral researchers (R2), their strongly varying thematic orientation (R3), and the different levels of prior knowledge (R4) need to be considered. The dynamic development in RDM and DS also requires high flexibility with respect to the courses offered and their content (R5). Moreover, early career researchers benefit from contact with peers and senior researchers across multiple disciplines and institutes as well as with partners like companies or NFDI consortia (R6), and from a written confirmation of participation (R7). In addition, specific support as well as training on demand are essential if participants are to successfully implement novel RDM concepts and DS methods in their daily research workflows (R8).

Operative implementation: In terms of content, the training covers the entire data-value chain, i.e. every step of the data life cycle (see Figure 1), including DS methods and the required technical infrastructure, as shown by Stodden (2020) (R1, see Figure 2). This approach allows synergy effects and innovation potentials of both fields, RDM and DS, to be exploited. Training modules follow a thematic structure in line with the data life cycle. However, each module is a stand-alone component (R5), allowing easy modifications to the curriculum and offering doctoral researchers the option to select courses to suit their individual needs (“course picking principle”) (R2, R3). In addition, modules are structured by difficulty level and topics, RDM and DS (R3, R4). Following the completion of the respective course series, participants receive a ‘Certificate of Participation’ (R7).

Various training formats offer a networking and socialising platform for participants, lecturers and invited speakers (R6). These also facilitate individual support as well as training on demand (e.g. self-learning, informal meetings for individual support, summer schools, and inspiring talks by partners with networking opportunities) (R8).

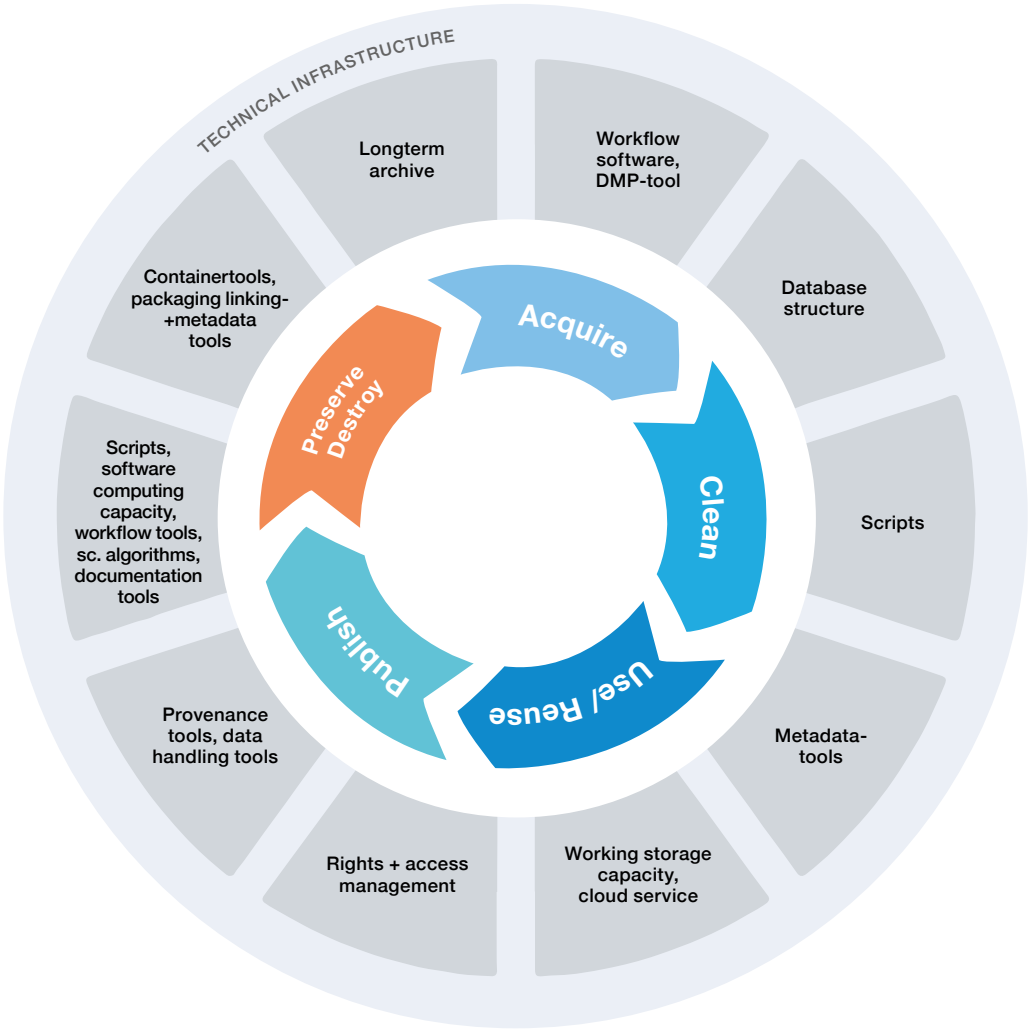

Figure 2: Schematic representation of the data life cycle, modified according to GFBio e.V. and Stodden (2020). The inner circle (highlighted in white) shows the individual steps of the data (science) life cycle, including the work steps of a data scientist. The outer circle (grey background) shows the technical infrastructure required to perform RDM and DS. Based on the individual work steps and required technical infrastructure, the competencies required can be derived synergistically.

In total, the programme is composed of:

1. The curriculum consisting of three individual course series, known as tracks *(Starter Track* and two *Operator Tracks),*
2. Additional modules for flexibly serving current topics and requirements *(Highlight Modules),*
3. Inspiring talks from the private sector and academia, accompanied by networking opportunities *(Data Stories)*,
4. A compilation of existing digital training materials for self-learning *(Digital Toolkit),*
5. Informal meetings to offer a platform for topic-specific consulting by experts and to foster networking and exchange across disciplines *(Hacky Hours),*
6. A summer school on up-to-date data handling concepts and analysis methods for specific data types *(Data Factory).*

2.4 Programme schedule

The programme and all its components are offered annually. The Starter Track is held in the first half of the year, with the two Operator Tracks, i.e. Data Steward Track and Data Scientist Track, alternating every other year (see Figure 3). Highlight Modules are added to a respective track according to their difficulty level and topic. Data Stories, Hacky Hours and Data Factories are held separately from the tracks to suit demand and the available resources. The Digital Toolkit is available at any time.

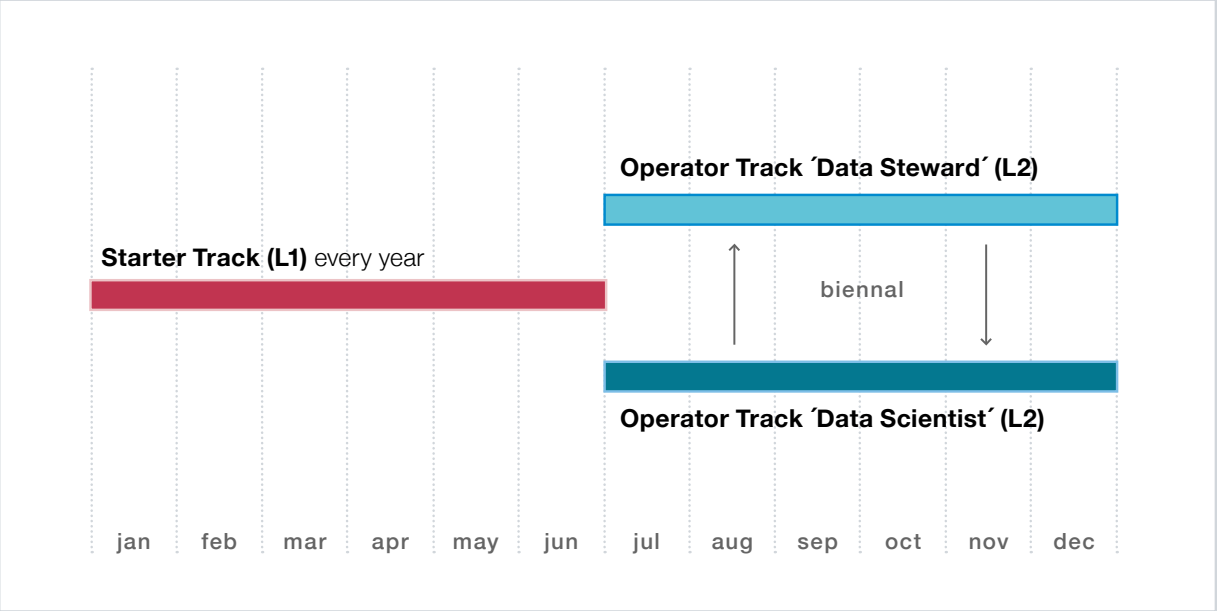

Figure 3: The Data Train curriculum schedule. Additional components are held independently of the tracks.

2.5 Training component profiles

This section describes the different training components in detail along organisational aspects (for details, see Tables 6-9 in Appendix 6.1). The component profiles should serve as implementation instructions for other locations and be adjusted according to local needs and conditions such as (1) the involved partners (e.g. universities, non-university research institutions, private companies), (2) research foci, (3) the targeted learning outcomes (competence levels), (4) trainer availability (internal, external), (5) the target group (career stage, time resources, scope), and (6) the type of offer (free of charge vs. paid, voluntary vs. mandatory).

All the organisational aspects described below rely on defined processes and workflows, and require appropriate (technical) infrastructure (i.e. a website, modular administrative system (registration, certification, participant management, etc.), a video conferencing system, online tools for collaborative training, technical equipment for hybrid formats, lecture rooms, etc.).

2.5.1 Curriculum

The curriculum is the main driver of the training programme. There are three tracks consisting of individual courses that doctoral students can follow: a track for beginners (*Starter Track*, level 1: “understand”) with lectures providing overviews of both RDM and DS, and two tracks (*Operator Tracks*, level 2 “apply”) with hands-on workshops, divided by subject into the *Data Steward Track* (mainly RDM competencies) and the *Data Scientist Track* (mainly DS competencies) (see Figure 4).

Descriptions of the individual modules in all three tracks are provided in Appendix 6.2 and 6.3. All of the modules are contributed and developed voluntarily by lecturers (mostly) from UBRA. The curriculum was piloted in 2021 and is continuously revised according to lessons learned, participant feedback, and new developments in the given fields. Consequently, learning content should be further developed to offer future-oriented training, which is essential for the research sector. Please note that this version of the handbook only provides the current state of the modules.

Each iteration of the curriculum starts with a *Kick-off event* to introduce the programme (scope, concept, and ties with the NFDI) and organisational aspects to the new cohort of participants. Moreover, introductory talks on DS applications and the importance and benefits of RDM should offer first insights and generate interest in data competencies among the participants.

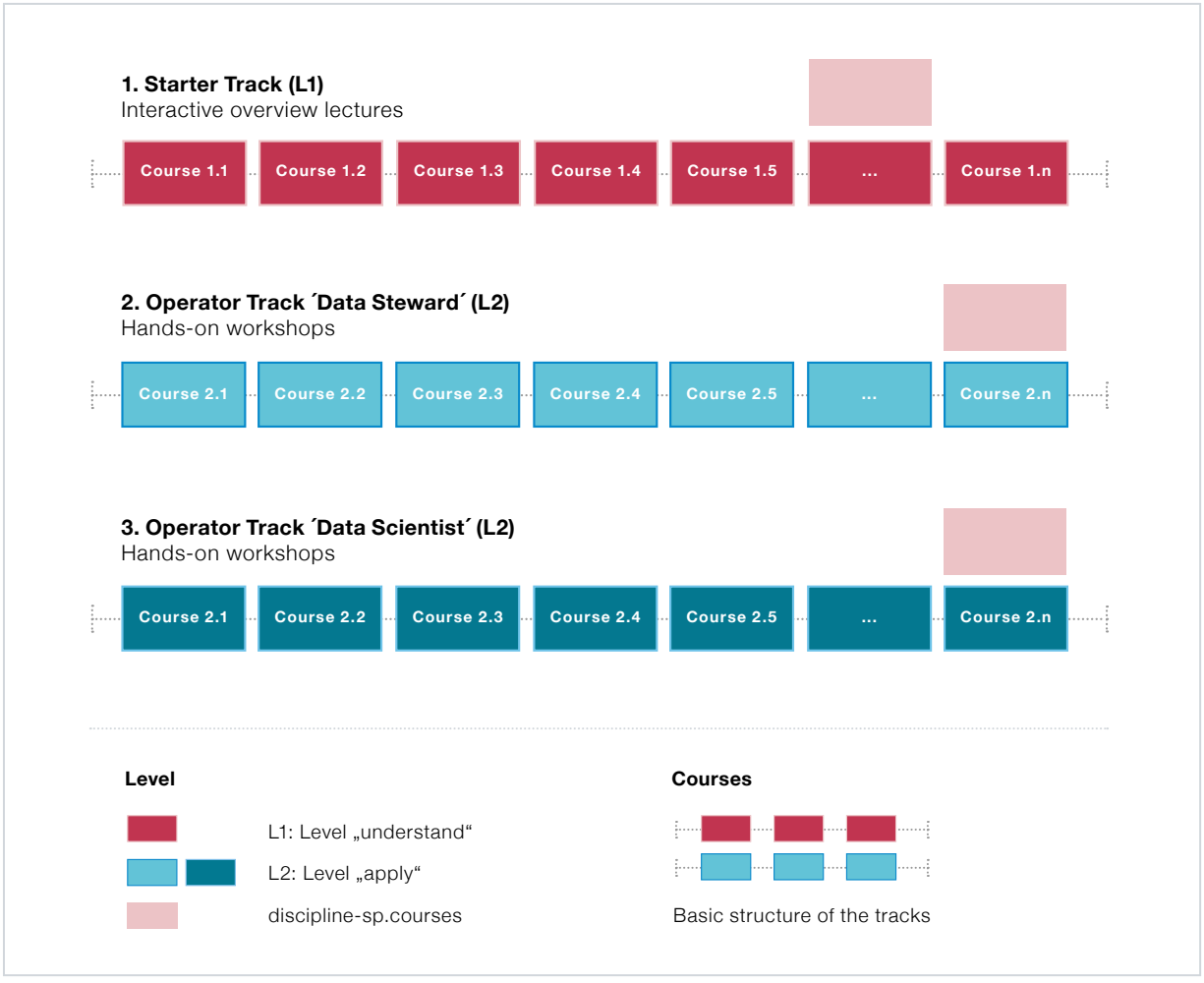

Figure 4: Schematic illustration of the curriculum’s concept. The boxes symbolise the individual courses. All of the courses within one track are connected by a black line which indicates the thematic structure of each track. Overview lectures in the *Starter Track* (level 1) are red, hands-on workshops in the *Operator Tracks* (level 2) are blue. The light red boxes are discipline-specific courses that could be offered in scientific domains or within the NFDI consortia, and can build on the basics taught in the Data Train programme. Modified from Hörner et al. (2021).

The *Starter Track* serves to introduce doctoral researchers to the generic topics of both fields of competence (RDM and DS) and to convey basic knowledge in an overview. Moreover, it is aimed at anyone who wants to get started in RDM and DS and, due to its online or hybrid format which allows high capacities, is open to anyone. This track serves as a guide for subsequently selecting the more time-consuming workshops in the *Operator Tracks* and therefore takes place at the beginning of each run. The *Starter Track* covers the basics of the following topics: Data Science, Big Data, Statistics, Computer Science (programming languages, cryptography, privacy and system security), RDM (FAIR principles, data protection and copyright, management of sensitive data (i.e. industrial and personal data), management of qualitative data).

The *Operator Tracks* are tailored to the specific needs of *Data Stewards* and *Data Scientists.*

The tasks of a *Data Steward* differ depending on the discipline, the area of application in science, and the local conditions. There is agreement that Data Stewards should implement the FAIR principles where a *FAIR Data Culture,* according to Scholten et al. (2019), needs to be implemented in different sectors of science: (1) infrastructure, (2) management (at a higher-level policy), and (3) research. Following this approach, the Data Steward Track focuses on the research sector and the essential competencies that researchers need for the FAIR handling of data: programming skills (R, Python, MATLAB), workflows in RDM (data management plans, reproducibility, data preparation/processing), innovative software and IT applications (database skills, data provision and extraction from online platforms) – especially for collaborative cooperation (Git/GitHub). This track is not a substitute for professional training for discipline-specific Data Stewards with an advisory, supportive and coordinating role at scientific institutions (see Figure 5).

Data Scientists typically focus on how data are prepared, processed and analysed, and on how results are transformed into decisions (Society of Computer Science, 2019). They need competencies in statistics, mathematics, computer science, as well as corresponding domain expertise (see Figure 5). The Data Scientist Track consists of the following content: quantitative analysis methods, machine learning (ML) methods, deep learning methods, methods to evaluate techniques, algorithms and results, data visualisation and visual analytics.

In addition, the *Critical Thinking* component is deemed necessary to make appropriate and informed decisions about the processing, sharing, and use of data. Moreover, a common language should be established between disciplines that is aware of potential limitations. This leads to better acceptance and greater openness towards novel working paths, research approaches and technologies. The ethical and legal frameworks of both RDM and Data Science should be discussed and critically reflected upon. Therefore, *Critical Thinking* is a key component of the entire curriculum addressed by all three tracks (e.g. by courses offering philosophical reflection on data science, meaningfulness of data, digital ethics, asking the right research questions in data science, evaluating AI/ML algorithms, and philosophical issues related to data visualisation.

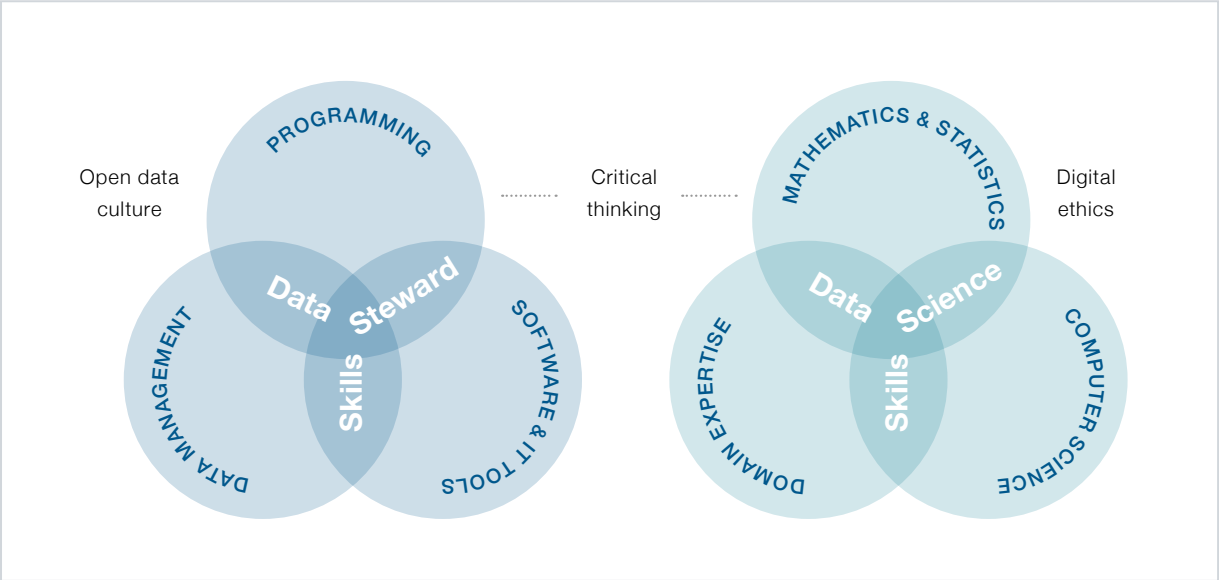

Figure 5: Competencies of Data Stewards and Data Scientists (Hörner et al. 2021).

2.5.2 Data Stories

In the *Data Stories* component, participants listen to inspiring stories about data handling, data management, and data science applied in the private sector or in academia. The speakers shed light on the importance of data competencies in regard to their individual working fields and discuss current challenges. These stories raise awareness, provide insights into different working fields, techniques, and applications while fostering the development of new ideas that participants can use in their research questions. Each story ends with an opportunity for discussion and a get-together so participants can build up a cross-discipline and cross-institutional network that stretches beyond the realms of science.

2.5.3 Digital Toolkit

The *Digital Toolkit* is a catalogue of selected existing digital self-learning materials that is openly accessible (to all career levels) to provide training on demand. Participants can select training content from this catalogue to suit their needs. Such training material catalogues are currently also developed by other programmes or projects (e.g. the *Training* DINI/nestor sub-working group and the *DALIA*[4] project) which are coordinated in the NFDI section *EduTrain*[5] . A cooperation with the *EduTrain* section is therefore recommended to avoid developing the same things separately.

[4] https://www.fst.tu-darmstadt.de/forschung_fst/zusammenarbeit_in_der_forschung/dalia/dalia_ueberblick.de.jsp (04.05.2023)

[5] https://www.nfdi.de/section-edutrain/?lang=en (04.05.2023)

2.5.4 Hacky Hours

Complementary to the generic and basic content of the curriculum modules, regular informal meetings known as *Hacky Hours* involving experts and users of data science applications serve as a platform for application-specific support, an exchange of experiences, and networking. Such meetings provide participants with an opportunity to discuss specific questions related to data science methods applied in their research projects. Each meeting may focus on a specific topic, e.g. a certain data science method or a specific process within the data science life cycle).

2.5.5 Highlight Modules

In order to keep pace with technological and methodological innovations, and to exploit the potential of research, the tracks may also include what are known as *Highlight Modules*. These modules focus on novel methods and exciting developments within the rapidly developing fields of RDM and DS. If such *Highlight Modules* are not part of the curriculum in the long term, it may make more sense to record training sessions and provide them on demand.

2.5.6 Data Factory

Data Factory is a summer school for doctoral researchers and post-doctoral researchers. This format offers participants the opportunity to learn and directly apply modern techniques and methods to their own data under the guidance of experts. Such intensive training with experts provides practical, individual and domain-specific support.

2.6 Evaluation of the programme

Internally, a steering committee monitors the programme in terms of scientific, content-related and organisational aspects. In addition, an interdisciplinary and cross-institutional working group of experts has been established from member institutions and relevant initiatives from UBRA with the aim of evaluating the concept and curriculum on a regular basis. As the programme is associated with the NFDI initiative, representatives of national consortia, in which the member institutions of UBRA are involved, participate in the development and training of the programme. Moreover, the Data Train developers are in exchange with further training programmes on a national level, for instance via the DINI/nestor *Training* sub-working group and the NFDI section *EduTrain.* In 2022, Data Train contributed to DINI/nestor's competence framework for RDM skills (Petersen et al. 2023) that was jointly developed with German RDM training programmes and authors of the FAIRsFAIR teaching and training handbook (Engelhardt et al. 2022). As a result, all of the initiatives involved took another step towards a coordinated competence framework for RDM.

In addition, participants evaluate each module in online surveys, with questions related to the learning content, difficulty level, speed of teaching, and organisational aspects. Such evaluation by participants is essential for the optimisation of the programme (to meet their training needs) and to improve the curriculum content.

3 RDM competencies related to health data

Relevant teaching content for Research Data management (RDM) and associated learning objectives from a number of national and international projects and training concepts on the subject of RDM have been identified by the "training courses" working group of the joint research data working group of the *Deutsche Initiative für Netzwerkinformation* (DINI, German Initiative for Network Information) and the network of expertise in long-term storage of digital resources (nestor), and published in a learning objective matrix (Petersen et al. 2023).

This chapter covers aspects and requirements specific to NFDI4Health with healthcare data and data obtained in prospectively planned epidemiological or clinical studies, along with a discussion of public health research projects and recommendations provided by DINI/nestor.

3.1 Overview

The DINI/nestor table below organises competencies into subject clusters, including „Basics and overarching concepts," „Working with data", „Documentation and metadata", „Long-term archiving, publication, and secondary use", „Law and ethics" and „Support structures." Each cluster contains several subject areas that have been grouped together for a better overview of the relevant subjects (see Table 2).

TABLE 2: SUBJECT CLUSTERS OF RESEARCH DATA MANAGEMENT (PETERSEN ET AL. 2023).

INDEX OF SUBJECT AREAS	For a better overview, the relevant subject areas have been grouped into clusters. Neither the clustering nor the order in which the subject areas are listed follow an evaluation or weighting of the listed aspects.
Basics and overarching concepts	General principles and concepts of research data management Research data policies Data Management Plans (DMPs) FAIR principles Open X (Open Data, Open Source, Open Science, etc.)
Working with data	Order and structure, versioning Data, data types, data formats Data storage and backup Data safety Data quality Tools (Research) software and coding
Documentation and metadata	Data documentation Metadata and metadata standards Persistent identifiers Ontologies and controlled vocabularies
Long-term archiving, publication, secondary use	Long-term archiving of data Publication channels for data Repositories Reuse of data
Law and ethics	General legal aspects Data protection and personal data Informed consent Anonymisation and pseudonymisation Ethical Aspects and Good Scientific Practice (GSP)
Support structures	Roles in data management / data stewardship Relevant Infrastructures Didactics Consulting

In addition to assigning relevant topics to six clusters, the DINI/nestor working group has published a more detailed matrix for different qualification levels as a guide (Petersen et al 2023).

This matrix provides a well-structured list of competencies, including learning goals, profiles, and levels for bachelor's, master's, PhD, and data steward students. The scope of this chapter is for doctoral students, but the competencies listed are generic and domain-independent by design. However, the core concept of NFDI4Health is focused on healthcare data, specifically clinical trials, epidemiological studies and public health research. While we embrace most of the generic recommendations provided by DINI, we also recommend special, domain-agnostic topics that may be important to adapt the competencies to healthcare-specific examples.

3.2 Basics and overarching concepts

The first cluster summarises the basics and overarching concepts in RDM (Table 3), including the research data life cycle (see Chapter 1) and its current developments, as well as important players and policies/guidelines. It also covers topics such as data management plans (DMPs) and principles that reusable research data must meet, such as the FAIR data principles, to ensure sustainable research data management. In addition, terms related to open science, such as open science content, open access, open data, open source, open data platforms, and the degree of openness are also part of this thematic area.

The above RDM concepts play a critical role in ensuring that research data are collected, managed, and used ethically and responsibly to support the advancement of medical knowledge and to improve patient care. The following examples are worth mentioning when introducing the concepts in healthcare contexts.

General principles and concepts of healthcare RDM

In healthcare research, most of the collected data are personal data and specific **personal data** like **genomic or health-related data,** which require more stringent handling and processing according to the EU **General Data Protection Regulation (GDPR).** Furthermore, data in an NFDI4Health context are typically collected in research projects and studies. Ethical principles for medical research involving human subjects, including research on identifiable human material and data, are laid down in the Declaration of Helsinki. Rights to self-determination, privacy, and confidentiality of personal information of research subjects are some of the basics of healthcare research and must be safeguarded by the person responsible for the research project.

In addition, for clinical interventional trials, scientific and technical aspects of pharmaceuticals are discussed by different regulatory authorities and the pharmaceutical industry, and appropriate guidelines are published by the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH). Good Clinical Practice (GCP) is one of the general and overarching guidelines and presents internationally recognised ethical and scientific quality standards for the design, conduct, monitoring, and reporting of clinical trials. GCP ensure that the rights, safety, and well-being of human subjects are protected, and that the data generated from the trial are reliable and accurate.

Some basic definition/procedures of GCP relevant to RDM

Sponsor is an individual, company, institution, or organisation which takes **responsibility** for the initiation, management, and/or financing of a clinical trial. This includes, for example, providing direct access to source documents, quality assurance, appropriate analysis and result reporting as well as archiving/retaining of essential documents.

Institutional review board (IRB) or Independent ethics committee (IEC) is an independent body whose responsibility is to ensure the protection of the rights, safety and well-being of human subjects involved in a trial or study, and to provide public assurance of that protection by, among other things, reviewing and approving/providing favourable opinion on, the trial protocol and the suitability of the investigator(s) and facilities.

Informed consent is a process by which a subject voluntarily confirms his or her willingness to participate in a particular trial, after having been informed of all aspects of the trial that are relevant to the subject‘s decision to participate. Informed consent is documented by means of a written, signed and dated informed consent form. Specific consent for data processing is often obtained simultaneously.

Source data are the original data collected during the course of a study. They are collected by the study team and recorded in source documents, such as case report forms (CRFs), electronic data capture (EDC) systems, or audio recordings. Source data should be attributable, legible, contemporaneous, original, accurate, and complete. This must be verified by appropriate quality assurance systems (see below). Changes to data should be traceable, should not obscure the original entry, and should be explained if necessary (e.g. via an audit trail).

Audit trail is a documentation that allows reconstruction of the course of events. It is used to track changes made to the study data, such as the date and time of the change, the user who made the change, along with a description of the change. Audit trails are important for maintaining the integrity of the study data and for ensuring compliance with GCP.

Quality assurances are planned and systematic actions established by the sponsor to ensure that the trial is performed, and the data are generated, documented (recorded), and reported in compliance with GCP and the applicable regulatory requirement(s). These actions might be, for instance, a.) **Standard Operating Procedures (SOPs),** a detailed set of written instructions that outline the procedures to be followed when conducting a specific task or activity; b.) **monitoring**, an act of overseeing the progress of a clinical trial, and of ensuring that it is conducted, recorded, and reported in accordance with the protocol, SOPs and the regulatory requirement(s), and c) **audit** as an examination of trial-related activities and documents to determine whether, e.g. the data were recorded, analysed, and accurately reported.

Risk-based quality management (RBQM) is a quality management approach used in the conduct of clinical trials. It involves the identification and assessment of risks to the quality of the trial or study, and the implementation of measures to mitigate those risks.

Essential documents are documents which individually and collectively permit evaluation of the conduct of a trial and the quality of the data produced. They have to be stored and readily available, and directly accessible upon request for at least 25 years as required in Article 58 of EU Regulation 536/2014.

Use and access committees: For German epidemiological and clinical studies and public health projects, the use of and access to data are typically governed by use and access committees (UACs). These committees are responsible for ensuring that the use and access of data follow legal and ethical requirements, as well as the regulations of the study sponsor or the funding agency.

Reporting/mandatory reporting (DSUR, SAE, request from EC/NCA, etc.): The following reports are important for ensuring the safety and well-being of trial participants and for ensuring the integrity of the trial data. Serious adverse events (SAEs) are any unexpected or serious adverse events that occur during the conduct of a trial. They must be reported to the sponsor by the investigator. According to regulatory requirements and GCP guidelines, the sponsor has different reporting obligations when it comes to SAE and other safety reports to be submitted to national regulatory authorities and/or the IEC.

Besides GCP Guideline ICH-E6 (R2), several other ICH guidelines also apply to data management in clinical studies, e.g. E2a (Clinical Safety Data Management: Definitions and Standards for expedited Reporting), E8 (General Considerations for Clinical Trials), E9 (Statistical Considerations in the Design of Clinical Trials, and E1 (The Extent of Population Exposure to Assess Clinical Safety for Drug Intended for Long-term Treatment of Non-Life-Threatening Conditions).

The Medical Dictionary for Regulatory Activities (MedDRA) was developed under the auspices of the ICH, providing an international medical dictionary applicable to all phases of biopharmaceutical and medical product development.

A clinical data management plan used in clinical trials describes which clinical data will be acquired and how they will be handled, stored, checked for consistency and plausibility at each stage of the individual study, while a research data management plan in the broader sense contains information about the data management life cycle for the data to be collected, processed and/or generated, as well as making those data findable, accessible, interoperable and reusable (FAIR). Open access to the project results or publication of the obtained data according to FAIR principles is often required by funding programmes for clinical and epidemiological studies or research projects.

Before clinical trials will be considered for publication, the International Committee of Medical Journal Editors (ICMJE) requires registration of clinical trials in a public trial registry at or before the time of first patient enrolment, and manuscripts submitted to ICMJE journals to contain a data sharing statement.

Considering the ethical principles of the Declaration of Helsinki and ICH guidelines, several other specific guidelines and standards in human research have been developed. To ensure Good Epidemiological Practice (GEP), a guideline has been developed by the German Society for Epidemiology which also considers recommendations for secondary data analysis (Hoffmann et al. 2019).

Good Practice Data Linkage (GPD) is a set of guidelines for the responsible and ethical linkage of data from different sources for research and other purposes (March et al. 2020).

For clinical investigations, GCP principles are elaborated in DIN ISO 14155 to assess the clinical performance or effectiveness and safety of medical devices.

The above examples illustrate how the overarching concepts of general principles and concepts of research data management, research data policies, data management plans (DMPs), FAIR principles, and open science (open data, open source) can be applied in healthcare research data management to ensure that data are collected, managed, and used in an ethical and responsible manner to support the advancement of medical knowledge and to improve patient care.

3.3 Working with data

Working with data (Table 3) in the context of healthcare requires careful attention to order, structure, versioning, data types and formats, storage, backup, safety, and quality.

Order and structure refer to the organisation of the data in a logical and consistent manner. In healthcare, examples of order and structure include organising patient data by way of demographics, medical history, and treatment.

Versioning refers to the practice of maintaining multiple versions of the data over time. In healthcare, examples of versioning include maintaining different versions of electronic medical records/electronic health records (EMR/eHR) or clinical trial protocols.

Data types and formats refer to the types of data and the ways in which they are stored. In healthcare, examples of data types include numerical data (such as lab results), text data (such as patient notes), and image data (such as x-rays). Examples of formats include spreadsheets, databases, and image files.

Data storage and backup refer to the methods used to store and protect the data. This includes using secure storage methods, such as cloud storage or local servers, and regularly backing up the data to remote servers to safeguard against data loss.

Data safety refers to the measures taken to protect the data from unauthorised access or breaches. This includes using encryption, secure communication methods, and access controls. In healthcare, examples of data safety include implementing secure login protocols and using firewalls to protect patient data.

Data quality refers to the accuracy and completeness of the data. In healthcare, examples of data quality include verifying that patient data are accurate and up to date and ensuring that data are collected using consistent and reliable methods.

According to GCP, the sponsor is responsible for data handling, data verification, conducting statistical analyses, and preparing trial reports by appropriately qualified individuals. When using electronic data capturing (EDC) systems (see below), the sponsor must ensure that the EDC systems comply with the sponsor's established requirements for completeness, accuracy, reliability and consistent intended performance (i.e. validation), and maintains specific SOPs for using these systems. These SOPs should cover system setup, installation, and use (system validation and functionality testing, data collection and handling, system maintenance, system security measures, change control/versioning, data backup, recovery, contingency planning).

Furthermore, the sponsor is responsible for implementing and maintaining quality assurance and quality control systems with written SOPs to ensure that data are generated, documented (recorded), and reported in compliance with the protocol, GCP, and the applicable regulatory requirement(s).

Some of the available EDC systems are validated to fulfil the requirements of FDA (21 CFR Part 11), GCP, GDPR and ISO 14155, while others are not. If clinical trial data are deemed part of a marketing authorisation, such systems should be given precedence over others that have no validated compatibility with the requirements mentioned above. Clinical studies without the scope of market authorisation or CE certification and epidemiological studies can use other appropriate EDC systems which may be less expensive. Analysis of data of an individual clinical trial or epidemiological study must be planned and documented in a statistical analysis plan (SAP). A detailed SAP is mandatory in GCP compliant trials and should consider appropriate ICH guidelines, e.g. ICH-E9: Statistical

Principles for Clinical trials. As these data are gathered by the study team(s) for this pre-defined analysis, these data are considered to be primary data. Results of the individual trial should be made publicly available either as a study results report (aggregated data) or – if possible – as encoded or anonymised individual data. These published data can be used by other researchers for further analyses. Such analyses of data from multiple studies can be performed either as federated/decentralised or distributed data analyses, or as a centralised process where all the data from individual studies are transferred to and processed at a central data repository. NFDI4Health is aimed at offering advanced distributed data analysis services.

Anonymisation and pseudonymisation are techniques used to protect the privacy of participants by removing or obscuring identifying information from the data. This can include removing names, addresses, and other personal data and replacing them with pseudonyms.

In summary, working with data in the context of healthcare requires careful attention to regulatory conditions, data protection regulations and ethical concerns. The use of tools, software, and coding can facilitate the collection, management, analysis, protection, and sharing of data.

3.4 Documentation and metadata

Data documentation and metadata play an important role in ensuring that health research data are well-organised, easily accessible, and usable for future research.

Data documentation refers to the process of providing detailed information about the data, including descriptions of the variables, data collection methods, and any limitations or uncertainties associated with the data. This is important for ensuring that the data can be understood and used by others, and for supporting reproducibility and transparency in research.

As an example, case report forms (CRFs)/annotated CRF and data dictionaries can be mentioned. These contain information on the data collected, the format and structure of the data, definitions of the variables, and any data cleaning or processing that has been performed. An annotated CRF is a version of the CRF that includes additional information or annotations to provide context or explanations for the data that have been collected. This additional information can include field-specific instructions, definitions, and examples, as well as information about the data validation rules and any logic checks that are built into the CRF.

Metadata

According to the National Information Standards Organization (NISO) definition, metadata are defined as structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource. The types of metadata can be categorised as follows:

- descriptive metadata describing a resource for discovery and identification purposes,
- structural metadata describing the schema, data models, and reference data, and
- administrative metadata providing information about the management of a resource.

Metadata are used to provide information about the context, content, and structure of the data, and to make it more easily discoverable and usable. In healthcare, examples of metadata include information about the data collection methods, sample size, and inclusion/exclusion criteria.

Metadata standards are guidelines for the structure, format, and content of metadata that help to ensure consistency and compatibility across data sets.

The following metadata standards are commonly mentioned in literature:

- Structure standards (e.g. ISO 11179, ISO 15926, openEHR, CDISC ODM, OMOP, DCAT and W3C PROV)
- Technical standards (e.g. XML, RDF, OWL, JSON-LD and ClaML)
- Semantic standards (e.g. ICD-11, UMLS, SNOMED CT, LOINC, MedDRA, RxNorm).

Moreover, FHIR (Fast Healthcare Interoperability Resources) is becoming popular. It is a set of standards for exchanging healthcare data, and defines metadata schema for describing various entities relevant to healthcare, such as patients, procedures, and clinical reasoning, including protocols for exchanging data and metadata records between systems. The information may be serialised as XML, JSON, ND-JSON, or RDF/Turtle.

NFDI4Health has set itself the goal to evaluate and develop standards for FAIR data. This concerns data management and publication policies, (meta-)data standard harmonisation, data quality assessment as well as standardisation of health data access. Furthermore, NFDI4Health will deliver services, service-enabling tools, and the necessary technical support. These include a central search service, tools supporting data quality assessments, and metadata annotations.

3.5 Long-term archiving, publication, secondary use

In the context of healthcare data management, long-term archiving, publication, and the secondary use of research data, can play a critical role in advancing medical knowledge and improving patient care.

Archiving and access for regulators are important considerations in the conduct of clinical studies. Archiving refers to the process of preserving and storing study-related documents and data for a specified period. Access refers to the ability of authorised individuals or organisations to review and use archived data.

For example, according to the EU Regulation 536/2014 for clinical trials with medicinal products, archiving of essential documents and the Trial Master File is required for a period of at least 25 years for review by regulatory authorities, as well as for further research and analysis.

Besides the specific requirements for clinical trials, electronic health records must be stored in Germany for at least 10 years (Section 630f of the German Civil Code (BGB)) unless a prolonged retention period is stipulated elsewhere.

Prospective registration of a clinical research project in publicly available registries is recommended or required by several international institutions (e.g. WHO, ICMJE), European guidelines and regulations (e.g. EU Regulation 536/2014) or by funding bodies (e.g. BMBF, DFG). Such primary registries in the WHO registry network, which meet specific criteria for content, quality and validity, accessibility, unique identification, technical capacity and administration, include the German Clinical Trial Register (DRKS) for all categories of clinical and epidemiological studies, or public health research projects or the EU Clinical Trial Register (EUCTR) for clinical trials with medicinal products only. Although the widely used study registry clinicaltrials.gov does not fulfil the WHO criteria for a primary registry, it has been accepted as an appropriate registry by the ICMJE for study registration as a prerequisite for publication in a journal. It is worth mentioning that for registration of clinical trials with drugs, the EUCTR will be replaced by the European Clinical Trial Information System (CTIS) starting in 2022 and an appropriate European registry for trials with medical products as a part of the European database on medical devices (EUDAMED) will be established.

Publication channels for information on research projects and clinical and epidemiological studies as well as their results include journals, databases, and repositories that are specifically designed for sharing research data. Examples of such channels in healthcare include, in addition to the registries mentioned above, the National Institutes of Health‘s National Library of Medicine‘s database PubMed, and the National Center for Biotechnology Information‘s database GenBank.

Repositories are specialised data storage and sharing platforms for specific research domains on an institutional, national, or international level. For example, the US National Cancer Institute‘s Cancer Data Access System and the International Cancer Genome Consortium Data Portal are examples of repositories specific to cancer research.

NFDI4Health develops a German Central Health Study Hub as an inventory of German health studies. The aim is to enable findability of studies and access to structured health data to improve the management of public health data.[6]

[6] https://csh.nfdi4health.de/ (04.05.2023)

The European Open Science Cloud EOSC-Life brings together the 13 life science research infrastructures according to ESFRI (European Strategy Forum on Research Infrastructures) to create an open, digital and collaborative space for biological and medical research. The project will publish 'FAIR' data and a catalogue of services provided by participating research infrastructures for the management, storage and reuse of data in the EOSC. An overview of relevant results of the project can be found at https://www.eosc-life.eu/resources/project-deliverables. Further information about German public database inventorying the national health databases and registries can be found as part of EOSC-Life deliverables 4.5 (Panagiotopoulou et al. 2022).

ELIXIR (European Life Science Infrastructure for Biological Information) is an intergovernmental organisation that coordinates bioinformatics resources and services across Europe. These resources include databases, software tools, training materials, cloud storage and supercomputers.

ECRIN (European Clinical Research Infrastructure Network) has developed a Clinical Research Metadata Repository, an online tool to help scientific researchers find documents and data linked to a clinical research study, and to obtain information on the accessibility of those results.[8]

de.NBI (German Network for Bioinformatics Infrastructure) is a German research infrastructure that provides bioinformatics services and resources to researchers in the life sciences. It aims to support the development of new methods and tools to analyse biological data and make these resources widely available to the scientific community.[9]

[7] https://elixir-europe.org (04.05.2023)

[8] https://ecrin.org/clinical-research-metadata-repository (04.05.2023)

[9] https://www.denbi.de/services (04.05.2023)

3.6 Law and ethics

Law and ethics play a crucial role in the management of healthcare research data. They help to ensure that the data are collected, stored, and used in a way that is both legal and ethically responsible. Some aspects have already been discussed in preceding sections of this chapter and are thus not repeated here.

General legal aspects in healthcare research data management include compliance with regulations such as the General Data Protection Regulation (GDPR) in the EU and the Health Insurance Portability and Accountability Act (HIPAA) in the US. These laws set standards for the protection of personal health information and ensure that the data are collected, used, and shared in a way that respects individuals‘ rights and privacy.

In the European Union, clinical trials are regulated by the Clinical Trials Regulation (CTR, EU Regulation 536/2014) on medicinal products and the Medical Device Regulation (MDR, EU Regulation 2017/745). In vitro diagnostic medical devices are governed by EU Regulation 2017/746. The CTR came into effect in 2022 and lays down rules for the conduct of clinical trials on medicinal products for human use in the European Union. The CTR aims to simplify the process of application, and conducting clinical trials in the EU, while maintaining a high level of protection for trial participants. The Clinical Trials Information System (CTIS) is now the single-entry point for sponsors and regulators of clinical trials for the submission and assessment of clinical trial data. The MDR came into effect in 2021 and lays down rules for the design, manufacture, and sale of medical devices in the European Union. The MDR also aims to improve the safety and performance of medical devices, while increasing the transparency of regulatory requirements. Investigators must hold appropriate qualification. In Germany, training and qualification requirements are laid down by the German Medical Association.[10]

EU Regulations 536/2014 and 2017/745 are implemented into German law by way of specific provisions in the German Medicinal Products Act (Arzneimittelgesetz, AMG)[11] and the German Medical Device Law Implementation Act (Medizinprodukte-Durchführungsgesetz, MPDG)[12].

In general, clinical trials in Germany involving medicinal products or medical devices require approval by an IEC and by the national regulatory authority (BfArM or PEI, depending on the characteristics of the interventional product). Both can be applied by a single-entry point, either CTIS or the German Medical Devices Information and Database System (DMIDS).

Clinical and epidemiological studies as well as public health research projects beyond the scope of AMG and MPDG are less regulated in Germany. However, GCP principles have to be considered in these studies as well, and a consultation with a local IEC is required for physicians planning biomedical research on humans according to Section 15 of the professional code for physicians. This approach is also recommended for other researchers planning a study with human subjects.

[10] https://www.bundesaerztekammer.de/fileadmin/user_upload/BAEK/Themen/Medizin_und_Ethik/Ethikkommissionen_LAEK/2022-01-14_Empfehlungen.pdf (04.05.2023)

[11] https://www.gesetze-im-internet.de/englisch_amg/ (04.05.2023)

[12] https://www.gesetze-im-internet.de/mpdg/ (04.05.2023)

3.7 Support structures

The preceding sections covered a variety of different institutions, organisations and projects which provide biomedical research support and training. Due to rapid developments, the diversity of health-related research projects, and the very large number of different and specific support offers, only a limited selection can be provided in this handbook.

It is important to define the roles of the staff involved in data management to ensure a smooth workflow such that data are collected, managed, and analysed in support of the scientific study objectives while also protecting the privacy and confidentiality of study participants. Common roles in clinical studies include a (principal) investigator, project manager, data manager, data steward, data entry clerk, data analyst, database administrator, IT staff and monitor personnel for source data verification.

Advice can be provided by a national competent authority (NCA) for specific questions regarding the development process (Scientific Advice), before admission of a clinical trial (Pre-Admission Advice) and, upon request, prior to grant approval (Pre-Grant Approval Advice). Simultaneous National Scientific Advice (SNSA) offers consultation of different NCAs at the same time.

Local academic support infrastructures for clinical studies (Koordinierungszentren für Klinische Studien, KKS and Zentren für Klinische Studien, ZKS) offer user-friendly consultation and advice and training programmes.[13]

For example, the 54 medical IECs in Germany can be consulted if any regulatory or data protection questions arise during preparation of a research project.[14]

[13] Local KKS/ZKS can be found at https://www.kks-netzwerk.de/en/network/about-us/members/. (04.05.2023)

[14] Local IECs can be found at https://www.akek.de/ethik-kommissionen/ (04.05.2023)

4 Further NFDI4Health training courses covering RDM aspects

While Chapter 2 dealt with the data train programme, which is a broader approach to enhance RDM skills and competencies in a more diverse cross-disciplinary audience, this chapter focuses on training offerings tailored more specifically to the needs of the NFDI4Health community. This is done according to prior knowledge and the competences sought. For each of the following offerings, reference will be made to the competency matrix in Chapter 3.

The level of detail about training offerings will increase throughout this chapter, ranging from broader RDM in the life sciences to offerings tailored to the needs of certain groups and to increase familiarity with tools and services designed by NFDI4Health. Section 4.1 provides the concept for training (bio-)medical researchers at all qualification levels on RDM best practices, while Section 4.2 outlines a domain-specific adaptation (epidemiological research). This still rather broad approach is followed by training courses on data quality assessment and courses aimed at facilitating NFDI4Health specific products such as the German Central Health Study Hub.

4.1 Training courses on RDM in biomedicine

These courses are conducted with local cooperation partners, which may be university research data centres (Section 4.1.1), but also individual institutes or a research group for individual epidemiological or clinical studies (Section 4.1.2). In addition to the different cooperation partners, the training approaches also differ in their practical relevance: the more concrete the target group, the more NFDI4Health services can be brought into focus, in turn increasing the hands-on mentality.

4.1.1 Best practices on RDM in biomedicine – generic level

4.1.1.1 Basic concept

This course is intended for researchers from (bio-)medicine disciplines at all levels who are interested in an introduction to best practices in RDM. Its basic concept was developed by ZB MED Information Centre Life Science as one of the NFDI4Health partners and will be adapted to suit the needs of the local research data centre. Thus, the individual course content is the result of a collaboration between ZB MED and the respective local research data centre. The respective centre usually presents the generic aspects of RDM and introduces the local services and infrastructure, while ZB MED imparts (bio-)medicine-specific aspects of RDM (e.g. privacy issues) and introduces its services (e.g. RDMO-4Life, the ELN guide[15]) as well as NFDI4Health services.

The course runs for 7 to 8 hours online or in-person, and includes interactive elements. In practice, it is recommended to split the course into smaller parts.

[15] RDMO4Life represents the instance of the research data management organiser RDMO (https://rdmorganiser.github.io/) for the life sciences, https://rdmo.publisso.de/. For the electronic laboratory notebook guide, see ZB MED (Ed.) 2021.

4.1.1.2 Course outline and agenda

The course outline comprises all aspects of RDM, including fundamental concepts like the FAIR data principles and all of the management steps in the research data life cycle, specifically:

FUNDAMENTAL CONCEPTS	STEPS IN THE RESEARCH DATA LIFE CYCLE
Research data	Planning (e.g. data management plan)
RDM	Data collection and documentation (e.g. using electronic lab notebooks ELN)
Metadata & metadata standards	Data processing (e.g. data standardisation)
FAIR data principles	Data publishing and sharing (e.g. research data repositories like the German Central Health Study Hub)
Good scientific practice	Data preservation (e.g. appropriate data formats)
Policies and guidelines on managing research data (e.g. NFDI4Health publication guidelines)	Data reuse and search (e.g. search portals for research data like the German Central Health Study Hub)
Legal issues (privacy and licences)	

The following agenda is an example of a training course divided into four two-hour time slots over four days.

DAY 1

- 09:00 - 09:25 Welcome
- 09:25 - 09:50 Fundamental concepts (1)
- 09:50 - 10:00 Break
- 10:00 - 10:40 Fundamental concepts (2)
- 10:40 - 10:55 Q&A
- 10:55 - 11:00 Feedback

DAY 2

- 09:00 - 09:40 Fundamental concepts (3)
- 09:40 - 09:55 Planning
- 09:55 - 10:05 Break
- 10:05 - 10:40 Data collection (1)
- 10:40 - 10:55 Q&A
- 10:55 - 11:00 Feedback

DAY 3

- 09:00 - 09:20 Data collection (2)
- 09:20 - 09:40 Data processing
- 09:40 - 09:50 Data publishing & sharing (1)
- 09:50 - 10:00 Break
- 10:00 - 10:25 Data publishing & sharing (2)
- 10:25 - 10:40 Data preservation
- 10:40 - 10:55 Q&A
- 10:55 - 11:00 Feedback

DAY 4

- 09:00 - 09:15 Data reuse & search
- 09:15 - 10:00 Legal issues
- 10:00 - 10:10 Break
- 10:10 - 10:35 Best practice example – NFDI
- 10:35 - 10:40 Closing
- 10:40 - 10:55 Q&A
- 10:55 - 11:00 Feedback

All slides with links to further information are available online.[16]

[16] https://doi.org/10.4126/FRL01-006452660 (07.06.2023)

4.1.1.3 Learning objectives

The learning objectives are based on the fundamental concepts and the steps in the research data life cycle (Chapter 4.1.1.2)[17]. Participants are therefore able to:

FUNDAMENTAL CONCEPTS	STEPS IN THE RESEARCH DATA LIFE CYCLE
Research data • Define research data • Identify the data type(s) of their own research project	**Planning** (e.g. data management plan) • Describe what a DMP is • Explain why DMPs are a step towards FAIR • Explain what should be covered in a DMP • Sketch a DMP for their own project
RDM • Define RDM • Appraise the usefulness of RDM • Apply the steps of the research data life cycle to their own project	**Data collection and documentation** (e.g. using electronic lab notebooks ELN) • Identify different types of documentation • Explain the purpose of the documentation • Define ELNs • Appraise the usefulness of ELNs • Electronic Lab Notebooks (ELNs) (2/2) • Describe the types of ELNs and their respective pros and cons • Explain why ELNs foster good scientific practices and the FAIR data principles • Identify the role of ELN in research data management • Use appropriate resources and tools to choose an ELN
Metadata & metadata standards • Describe types of metadata • Identify metadata standards • Use metadata standards to describe resources • Search and find metadata standards in registries • Appraise the usefulness of metadata standards to describe a resource	**Data processing** (e.g. data standardisation) • Appraise the usefulness of data standardisation • Define code-based data processing and analysis • Appraise the usefulness of code-based data processing and analysis • Identify best practices when writing code • Identify tools and techniques for computational reproducibility • Identify the different types of version control systems and their pros and cons
FAIR data principles • Paraphrase the FAIR data principles • Explain why the FAIR principles were developed • Apply the FAIR data principles to their own work • Evaluate the FAIRness of their own work or the work of others • Differentiate FAIR and Open	**Data publishing and sharing** (e.g. research data repositories like the German Central Health Study Hub) • Identify the types of data that can be shared • Identify collaboration tools • Find and choose an appropriate repository to publish their data • Identify persistent identifiers and explain their different use cases • Appraise the usefulness of PIDs

[17] Furthermore, the learning objectives are based on the outcomes of the FAIRsFAIR teaching and training handbook (Engelhardt et al. 2022) and customised by ZB MED.

FUNDAMENTAL CONCEPTS	STEPS IN THE RESEARCH DATA LIFE CYCLE
Good scientific practice • Identify examples of GSP and scientific misconduct • Policies & guidelines • Identify research data management-related and discipline-specific policies and guidelines to follow	**Data preservation** (e.g. appropriate data formats) • Identify the difference between backup, archive and publication of research data • Appraise the usefulness of digital preservation • Identify the requirements in terms of digital preservation • Decide what data to keep • Identify appropriate data formats for digital preservation
Policies and guidelines on managing research data (e.g. NFDI4Health publication guidelines) • Identify research data management-related and discipline-specific policies and guidelines to follow	**Data reuse and search** (e.g. search portals for research data like the German Central Health Study Hub) • Find published datasets in their discipline • Successful case of data reuse • Appraise the usefulness of data discovery and reuse
Legal issues (privacy) • Define data privacy and related terms • Appraise the usefulness of data de-identification • Privacy issues (2/2) • Identify the differences between pseudonymisation and anonymisation • Explain reasons for data protection • Identify tools to manage and analyse sensitive data • Define informed consent • Find a template for consent forms	
Legal issues (licences) • Define licences • Identify properties of recommendable licences • Differentiate and combine Creative Commons copyright licences • Identify when they need explicit permission to use a resource	

4.1.1.4 Practical guidelines

Number of participants (for personal training): The ideal number of participants for such an interactive course is 10 to 30. If there are more than 30 registrations, the organisers should choose a format they deem appropriate, and if there are fewer than ten registrations, the workshop should be postponed.

Registrations: The local research data centre is responsible for registrations, with a registration template provided to determine the specific needs of the participants.

Platform for online training: The local research data centre is responsible for setting up the meeting(s).

Content: All documents (outline, examples, timetable, slides incl. template, list of interactive activities, poll templates, feedback form) needed to prepare the workshop should be made available in a shared folder via a cloud service. Ideally, the slides should be ready two weeks before the workshop and made publicly available with a CC BY licence (unless stated otherwise).

Feedback: Feedback should be collected via appropriate instruments (feedback round, interactive pad, survey, etc). Timeline: The timeline should be negotiated in a timely manner to arrange all aspects of organising the training workshop.

After the workshop: The poll reports should be anonymised and uploaded to the shared folder. All additional material like online whiteboard figures should be merged with the course material, and the CC BY licence should be added and made publicly available.

4.1.1.5 Past experiences

In the more than 20 training courses conducted since 2019, the following key experiences have emerged:

- The more precisely the target group is known, the more targeted the training can be. For example, if it is known in advance that the participants work with study data, the NFDI4Health services can be included.
- Interactivity plays an important role in keeping the attention of the participants. Conducting exercises or surveys or using a digital board (for online training) will liven up the event.
- Regarding the topics to be trained, it is advisable to divide the concept into individual modules so that one delimited topic is covered per time slot, e.g., data management planning or electronic lab books. However, the basic concepts remain part of each training, at least in summarised form.

4.1.2 Best practices on RDM in biomedicine – hands-on level

4.1.2.1 Basic concept

This course was given to an audience of Ph.D. or higher-level researchers from the Institute of Nutritional Epidemiology (IEL) at the University of Bonn[18] who were interested in being introduced to best practices in Research Data Management (RDM). The IEL training covered general aspects of RDM and introduced their local services and infrastructure. ZB MED complemented (public) health-specific aspects of RDM (e.g. privacy issues) and introduced NFDI4Health's services (e.g. German Central Health Study Hub[19], the Publication Policy[20], the Metadata Schema[21]) ZB MED is involved in.

The course ran for three hours online (via Zoom) and included interactive elements and polls. It was split into two 90-minute sessions on two separate days, with a one-week gap between the sessions.

4.1.2.2 Course outline and agenda

The course outline comprised all aspects of RDM, including fundamental concepts like the FAIR data principles and some of the management steps in the research data life cycle adapted to the NFDI4Health's offer. It is mostly in line with the course outline illustrated in Table 2 in Section 4.1.1.2, supplemented by the specific NFDI4Health services.

TABLE 3: COURSE OUTLINE FOR THE IEL TRAINING

FUNDAMENTAL CONCEPTS	STEPS IN THE RESEARCH DATA LIFE CYCLE
Research data	Planning (following the publication guidelines of the NFDI4Health and the NFDI4Health Task Force COVID-19)
Research data management	Data documentation (using the NFDI4Health Task Force COVID-19 Metadata Schema and other recommended standards)
Metadata and metadata standards	Data publication (in the NFDI4Health's data portal German Central Health Study Hub using the corresponding entry form)
Ontologies and controlled vocabularies	Data reuse and search (using the NFDI4Health German Central Health Study Hub as a search portal)
FAIR data principles	
Good scientific practice	
Policies and guidelines on managing research data (NFDI4Health publication guidelines)	
Legal issues (CC licences)	

[18] https://www.ernaehrungsepidemiologie.uni-bonn.de/ (07.06.2023)
[19] https://csh.nfdi4health.de/ (07.06.2023)
[20] https://doi.org/10.4126/FRL01-006431467 (07.06.2023)
[21] https://doi.org/10.4126/FRL01-006439110 (07.06.2023)

The agenda for the training course was divided into two 90-minute time slots over two days. It was adapted from the agenda presented in Table 3 in Section 4.1.1.2.

TABLE 4: AGENDA OF THE IEL TRAINING.

DAY 1	DAY 2
INTRODUCTION TO RESEARCH DATA MANAGEMENT (RDM)	**A STEP FORWARD TO MAKING DATA FAIR WITH NFDI4HEALTH**
• 13:00 - 13:25 Welcome • 13:25 - 13:30 Research data • 13:30 - 13:45 Research data life cycle • 13:45 - 13:55 FAIR data principles • 14:05 - 14:10 Guidelines on RDM • 14:10 - 14:25 Q&A • 14:25 - 14:30 Feedback	• 13:00 - 13:10 Introduction • 13:10 - 13:35 Publication Policy • 13:35 - 13:50 German Central Health Study Hub • 13:50 - 14:05 Metadata schema • 14:05 - 14:10 Recommended standards • 14:10 - 14:25 Q&A • 14:25 - 14:30 Feedback

4.1.2.3 Learning objectives

The learning objectives for this training are a subset of the learning objectives described in Table 4 in Section 4.1.1.3. The differences to the aforementioned table are due to shortening the training to less than half of the recommended time and adapting it to the NFDI4Health services.

TABLE 5: LEARNING OBJECTIVES FOR THE IEL TRAINING.

FUNDAMENTAL CONCEPTS	STEPS IN THE RESEARCH DATA LIFE CYCLE
Research data • Define research data	**Planning** (following the NFDI4Health publication guidelines) • Differentiate between different study documents • Know all the required steps to prepare different study documents for publication
RDM • Define RDM • Appraise the usefulness of RDM • Apply the steps of the research data life cycle to their own project	**Data documentation** (using the NFDI4Health Metadata Schema) • Identify different types of documentation • Explain the purpose of the documentation • Define ELNs • Appraise the usefulness of ELNs • Electronic Lab Notebooks (ELNs) (2/2) • Describe the types of ELNs and their respective pros and cons • Explain why ELNs foster good scientific practices and the FAIR data principles • Identify the role of ELN in research data management • Use appropriate resources and tools to choose an ELN
Metadata & metadata standards • Describe types of metadata • Identify metadata standards	**Data publishing and sharing** (publishing in the NFDI4Health German Central Health Study Hub) • Identify the types of data that can be shared • Know how to publish their data FAIR in the German Central Health Study Hub using the entry form
FAIR data principles • Paraphrase the FAIR data principles • Apply the FAIR data principles to their own work • Differentiate FAIR and Open	**Data reuse and search** (using the NFDI4Health German Central Health Study Hub as a search portal) • Find published studies and datasets in their discipline • Know how to use the provided filters for optimised search results

TABLE 5: LEARNING OBJECTIVES FOR THE IEL TRAINING.

FUNDAMENTAL CONCEPTS	STEPS IN THE RESEARCH DATA LIFE CYCLE
Good scientific practice • Identify examples of GSP and scientific misconduct	
Policies and guidelines on managing research data (e.g. NFDI4Health publication guidelines) • Identify research data management-related and discipline-specific policies and guidelines to follow	
Legal issues (licences) • Differentiate and combine CC licences	

4.1.2.4 Practical implementation

Number of participants: There should be no more than 30 participants. On the occasion described, there were 13 participants, with 8 taking part on both days.

Registrations: The registration process should actually be managed by the local cooperation partner. In this case, the participants were part of an existing research group that held regular meetings at the same time as the training, so registering for the training was not necessary as the time slot could be used to attend the training.

Platform for online training: The local cooperation partner should be responsible for hosting the training. The IEL hosted the training online on Zoom and granted host rights to the ZB MED trainers.

Content: All documents (outline, examples, timetable, slides incl. template, list of interactive activities, poll templates, feedback form) used to prepare and host the workshop should be made available in a shared folder via a cloud service. For this training, they were made available to all trainers in a shared folder and have been published subsequently.[22] The slides were ready one week before the training and shared with the participants after the training.

Feedback: Feedback should be collected using appropriate instruments. Feedback about the training was collected via an online survey after the second day of training. The IEL sent two reminders in the weeks following the workshop, prompting participants to fill in the survey.

Timeline: The timeline should be negotiated in a timely manner to arrange all aspects of organising the training workshop. For this training, the timeline was negotiated several weeks before the workshop, the specific content and focus of the workshop determined, and the trainers agreeing on the presenter of each topic. Dates for both days were set and communicated to the potential participants.

After the workshop: The poll reports should be anonymised and uploaded to the shared folder. All additional material should be merged with the course material, and a CC BY licence should be added and made publicly available. The poll reports for this training were supposed to be saved. Due to technical difficulties, however, it was not possible to gain access to and save the results after the training. The slides used in the training were shared under a CC BY licence with the participants after the training.

[22] https://doi.org/10.4126/FRL01-006452640 (05.06.2023)

4.1.2.5 Past experiences

After evaluating the workshop, it became clear that some adjustments would be needed for future workshops. The evaluation raised the following points:

- The **learning objectives** should be clearly stated. An additional slide should be used to list what participants can expect from this workshop and what they should learn on each day.
- The **time allocated to each topic** should depend on the focus of the workshop. It might be helpful to discuss certain areas of focus in advance with the trainer from the local cooperation partner, perhaps after collecting participants' expectations. In the event of a general workshop providing an overview of the topic, additional, more specialised workshops could be offered subsequent to the workshop.
- There should be (at least) one **practical session** on each day / in each slot of the workshop. The examples chosen to illustrate single topics should also be practical and transferable to the participants' scientific work.

There should be more emphasis on how to integrate the knowledge gleaned from the workshop into the participants' scientific work. Research needs, such as guidelines or requirements, should be given more attention and accompanied by tips on how to **implement** them on a practical level.

4.2 Training course on data quality assessment

Duration: 3 hours | Group size: 10-30 participants

Data quality should be assessed within and across studies in comparable and transparent ways, yet it is not always clear how to do this successfully. It is important to obtain the required skills to do so, for example to implement a data monitoring pipeline during an ongoing data collection, or perform the final assessment of the obtained data quality once data collection has been completed.

4.2.1 Target audience

This course is suitable for bachelor's, master's and PhD students, senior researchers in biomedical research, as well as data managers and data stewards who deal with data quality assessments and preparatory steps to enable them to carry out tasks such as metadata management.

4.2.2 Learning objectives

- Understand ways to conceptualise data quality
- Understand requirements on metadata to conduct data quality assessments
- Understand how such assessments can be performed with R, e.g. the dataquieR package

4.2.3 Course outline and agenda

The course is structured as follows:

1. The first level imparts definitions of data quality and their implications. This primarily involves data quality dimensions relevant to typical assessments and single data quality indicators. The initial focus is on a framework for observational studies1, which distinguishes four dimensions: Integrity, Completeness, Consistency, and Accuracy. This is contrasted with a data-quality approach for electronic health record data2, which follows a distinct structure by differentiating three data-quality categories: Conformance, Completeness, and Plausibility. Knowing these concepts provides guidance on what to look out for during assessments.
2. The second level refers to adequate information management to enable structured data quality assessments. A key component is the handling and formats of metadata to describe properties of the collected study data and expectations about them, such as variable and value labels, value lists, admissibility ranges, and assumptions about distributions or associations. How to assemble this information for reuse remains a challenge yet to be overcome.
3. The third level covers data quality assessments, with a focus on the programming language R. It shows how to trigger commands to generate targeted data-quality reports for selected data-quality aspects, along with very broad reports across multiple dimensions. Another aspect is how to interpret such reports.
4. Materials? [23]
5. Lessons Learned? „The course was very well attended and overall well received. However, it seems favorable to bring less theory in favor of more applied examples."

[23] https://zenodo.org/record/8042797

4.3 Training course on the German Central Health Study Hub

Duration: 3 hours | Group size: 5-12 participants

The National Research Data Infrastructure for Personal Health Data (NFDI4Health) is dedicated to managing data generated in clinical trials as well as epidemiological and public health studies. NFDI4Health provides services that aim to make research data more Findable, Accessible, Interoperable, and Reusable (FAIR), including permanent storage, semantic enrichment, and data merging from different sources while respecting privacy requirements.
This two-hour hands-on training workshop will repeat the principles of good research data management in the biomedical domain, introduce the NFDI4Health project, and allow participants to use NFDI4Health services to search and publish study metadata and instruments (hands-on). Participants who have carried out or are involved in a clinical or epidemiological study or work in public health can work with their own data set.

4.3.1 Target audience

This course is suitable for bachelor's, master's and PhD students as well as senior researchers in biomedical research. It is of particular interest to scientists with a basic knowledge of the FAIR data principles, and scientists who strive to publish their study protocols in order to make them accessible and more visible to their peers.

4.3.2 Learning objectives

- Understand the principles of research data management in the biomedical domain
- Learn about the NFDI4Health project and its services for FAIR research data
- Practise using NFDI4Health services to search and publish study metadata and instruments
- Apply NFDI4Health concepts on the German Central Health Study Hub (with given or own data)

4.3.3 Course outline and agenda

The course teaches different aspects of searching and submitting health studies in the German Central Health Study Hub. It supplements the specific services of the NFDI4Health project described in Section 4.1.2.

Participants will receive a login to a test system where they enter their study. They will also receive a handout for the exercises along with access to the presentation slides showcasing the use of the German Central Health Study Hub (for later reference).

DAY 1

Part 1: Welcome and programme overview (10 minutes)

1. Overview of the workshop and learning objectives
2. Housekeeping
3. Introductions and logistics
4. Setting the stage

Part 2: Introduction to RDM practices and Health Data/NFDI4Health (20 minutes)

5. Research Data + Management
6. Data Life Cycle
7. Publishing Data/Metadata

Part 3: Introduction to the NFDI4Health German Central Health Study Hub

8. Short walkthrough of the system: search and data entry (10 minutes)
9. Hands-on session (70 minutes)
    - Round of introductions: What data did I bring? (10 minutes)
    - Search and explore available data (20 minutes)
    - Add data to the German Central Health Study Hub (30 minutes)
    - Collect (and submit) bug reports and feature requests (10 minutes)

Part 4: Conclusion and feedback (10 minutes)

10. Wrap-up
11. Five-finger feedback

Note: Participants should bring their own data, such as study descriptions or auxiliary files containing metadata.

5 Summary

The multi-level approach presented in this handbook provides examples of how to accommodate data management training needs. Levels range from doctoral students and early career researchers to medical doctors, epidemiologists, and research support professionals. The modular design allows for reuse and adaptation to local contexts and needs.

One level is oriented towards a cross-disciplinary RDM and data science training programme for doctoral students. The original Data Train is running at Bremen University and other institutions can join the initiative. However, a natural constraint are the local resources available for the practical part of the programme. Therefore, the concept itself can be reused and adapted locally. The programme concept is also open to the addition of certain domain-specific elements. Focussing on research professionals in the health domain, a second set of training concepts addresses biomedical RDM as well as digital services developed by NFDI4Health. Examples of this are distributed data analysis and making personal health data FAIR (with special attention paid to available metadata schemes, **A**ccess, **I**nteroperability, and **R**euse).

This approach also includes two levels of competencies, i.e. introductory/awareness (broad, shallow) and expert level (narrow, deep/specific). At the time of writing, the focus is on the introductory level. Expert-level training is currently being developed, while related topics are also part of intensive consultation cases and are maturing in parallel with the development of NFDI4Health services.

Also, the health-domain-specific training of the emergent group of data stewards will be addressed in the future, adapting to the results of the Data Train programme, as well as those of FAIRsFAIR (Engelhardt et al. 2022).

For an ongoing development and modification of the learning offerings/training courses to meet the changing needs of the target group, it is essential to evaluate them constantly and get feedback from the participants. This fosters improvement of the content, the way the content is delivered, and other factors crucial to the learning experience.

Feedback can be collected both during and after the course. We suggest using a standardised evaluation form to be completed at the end of the course (or module) before participants leave. The survey should cover issues such as clarity of the learning objectives, pace, balance of theory and practice, etc. This guide includes an example of what such a survey might look like, but it can be modified to suit the specific needs of the reader (Chap. 6.5).

The NFDI4Health training activities are constantly evolving. As a result, this guide is a snapshot taken at its first major milestone.

Bibliography

Berman, Francine, Rob Rutenbar, Brent Hailpern, Henrik Christensen, Susan Davidson, Deborah Estrin, Michael Franklin, et al. 'Realizing the Potential of Data Science'. **Commun. ACM** 61, no. 4 (March 2018): 67–72. https://doi.org/10.1145/3188721.

Ecrin. 'Clinical Research Metadata Repository'. Accessed 23 March 2023. https://ecrin.org/clinical-research-metadata-repository.

Cuadrado-Gallego, Juan J., and Yuri Demchenko, eds. *The Data Science Framework: A View from the EDISON Project.* Cham: Springer International Publishing, 2020. https://doi.org/10.1007/978-3-030-51023-7.

'De.NBI – Services'. Accessed 23 March 2023. https://www.denbi.de/services.

Deutsche Initiative für Netzwerkinformation e.V. 'DINI/nestor-AG Forschungsdaten'. Accessed 23 March 2023. https://dini.de/ag/dininestor-ag-forschungsdaten/.

'ELIXIR | A Distributed Infrastructure for Life-Science Information'. Accessed 23 March 2023. https://elixir-europe.org/.

Engelhardt, Claudia, Katarzyna Biernacka, Aoife Coffey, Ronald Cornet, Alina Danciu, Yuri Demchenko, Stephen Downes, et al. *D7.4 How to Be FAIR with Your Data. A Teaching and Training Handbook for Higher Education Institutions* (version V1.2.1). Zenodo, 2022. https://doi.org/10.5281/zenodo.6674301.

'Ernährungsepidemiologie — Landwirtschaftliche Fakultät Universität Bonn'. Accessed 23 March 2023. https://www.ernaehrungsepidemiologie.uni-bonn.de/.

Garbuglia, Federica, Bregt Saenen, Vinciane Gaillard, and Claudia Engelhardt. 'D7.5 Good Practices in FAIR Competence Education', 16 December 2021. https://doi.org/10.5281/zenodo.6657165.

'German Central Health Study Hub'. Accessed 23 March 2023. https://csh.nfdi4health.de/.

Hoffmann, Wolfgang, Ute Latza, Sebastian E. Baumeister, Martin Brünger, Nina Buttmann-Schweiger, Juliane Hardt, Verena Hoffmann, et al. 'Guidelines and Recommendations for Ensuring Good Epidemiological Practice (GEP): A Guideline Developed by the German Society for Epidemiology'. *European Journal of Epidemiology* 34, no. 3 (1 March 2019): 301–17. https://doi.org/10.1007/s10654-019-00500-x.

Hörner, Tanja. 'About Dtaa Train / UBRA'. U Bremen Research Alliance. Accessed 23 March 2023. https://www.bremen-research.de/data-train/about-data-train.

Hörner, Tanja, Iris Pigeot, Frank Oliver Glöckner, and Rolf Drechsler. 'Disziplinübergreifendes Modell zur Ausbildung von Forschungsdatenmanagement und Data Science Kompetenzen: „Data Train – Training in Research Data Management and Data Science"'. *Bausteine Forschungsdatenmanagement*, no. 3 (26 October 2021): 56–69. https://doi.org/10.17192/bfdm.2021.3.8343.

Informatik (GI), Gesellschaft für. 'Data Literacy und Data Science Education: Digitale Kompetenzen in der Hochschulausbildung'. Web portal - LIVE, 1 June 2018. https://gi.de/themen/beitrag/data-literacy-und-data-science-education-digitale-kompetenzen-in-der-hochschulausbildung/.

Lindstädt, Birte, and Aliaksandra Shutsko. 'Publication Policy of the National Research Data Infrastructure for Personal Health Data (NFDI4Health) and the NFDI4Health Task Force COVID-19', 11 February 2022. https://doi.org/10.4126/FRL01-006431467.

March, Stefanie, Silke Andrich, Johannes Drepper, Dirk Horenkamp-Sonntag, Andrea Icks, Peter Ihle, Joachim Kieschke, et al. 'Good Practice Data Linkage (GPD): A Translation of the German Version'. *International Journal of Environmental Research and Public Health* 17, no. 21 (January 2020): 7852. https://doi.org/10.3390/ijerph17217852.

Michener, William K. 'Ten Simple Rules for Creating a Good Data Management Plan'. PLOS *Comput Biol* 11, no. 10 (22 October 2015): e1004525. https://doi.org/10.1371/journal.pcbi.1004525.

www.dfg.de. 'Nationale Forschungsdateninfrastruktur'. Accessed 23 March 2023. https://www.dfg.de/foerderung/programme/nfdi/index.html.

'NFDI | Nationale Forschungsdateninfrastruktur e. V.' Accessed 23 March 2023. https://www.nfdi.de/.

NFDI4Health Task Force COVID-19, Haitham Abaza, Sophie Anne Ines Klopfenstein, Martin Golebiewski, NFDI4Health, Carsten Oliver Schmidt, Aliaksandra Shutsko, Carina Nina Vorisek, and Johannes Darms. 'Metadata Schema of the NFDI4Health and the NFDI4Health Task Force COVID-19 (V3_0)'. Application/pdf, 2022. https://doi.org/10.4126/FRL01-006439110.

Panagiotopoulou, Maria, Sarhan Yaïche, Amélie Michon, Christian Ohmann, Jacques Demotes, Mihaela Matei, Steve Canham, et al. 'EOSC-Life Public Database Inventorying the National Health Databases and Registries and Describing Their Access Procedures for Reuse for Research Purposes', 5 October 2022. https://doi.org/10.5281/zenodo.7148861.

Petersen, Britta, Claudia Engelhardt, Tanja Hörner, Juliane Jacob, Tatiana Kvetnaya, Andreas Mühlichen, Hermann Schranzhofer, et al. 'Lernzielmatrix zum Themenbereich Forschungsdatenmanagement (FDM) für die Zielgruppen Studierende, PhDs und Data Stewards'. Zenodo, 14 June 2023. https://doi.org/10.5281/zenodo.8010617.

'RDMO4Life'. Accessed 23 March 2023. https://rdmo.publisso.de/

Ridsdale, Chantel, James Rothwell, Michael Smit, Hossam Ali-Hassan, Michael Bliemel, Dean Irvine, Daniel Kelley, Stan Matwin, and Bradley Wuetherick. 'Strategies and Best Practices for Data Literacy Education: Knowledge Synthesis Report'. Report, 2015. https://doi.org/info:doi/10.13140/RG.2.1.1922.5044.

Scholtens, Salome, Mijke Jetten, Jasmin Böhmer, Christine Staiger, Inge Slouwerhof, Marije Van Der Geest, and Celia W.G. Van Gelder. 'Final Report: Towards FAIR Data Steward as Profession for the Lifesciences. Report of a ZonMw Funded Collaborative Approach Built on Existing Expertise'. Zenodo, 3 October 2019. https://doi.org/10.5281/ZENODO.3474789.

Stodden, Victoria. 'The Data Science Life Cycle: A Disciplined Approach to Advancing Data Science as a Science'. *Communications of the ACM* 63, no. 7 (18 June 2020): 58–66. https://doi.org/10.1145/3360646.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 'The FAIR Guiding Principles for Scientific Data Management and Stewardship'. *Scientific Data* 3, no. 1 (December 2016): 160018. https://doi.org/10.1038/sdata.2016.18.

'WMA - The World Medical Association-WMA Declaration of Helsinki – Ethical Principles for Medical Research Involving Human Subjects'. Accessed 23 March 2023. https://www.wma.net/policies-post/wma-declaration-of-helsinki-ethical-principles-for-medical-research-involving-human-subjects/.

ZB MED (Ed.). 'Electronic Laboratory Notebooks in the Context of Research Data Management and Good Research Practice – a Guide for the Life Sciences'. Application/pdf, 2021. https://doi.org/10.4126/FRL01-006425772.

6 Appendix

6.1 Data Train – Training component profiles

6.1.1 Kick-off event

TABLE 6: ORGANISATIONAL SUMMARY OF THE KICK-OFF EVENT.

Format	TYPE:	Talks on the programme, organisational aspects and importance of DS and RDM followed by a get-together
	LOCATION:	hybrid or online
	TIME:	ideally at the end of January before the Starter Track begins
Workload	DURATION:	2-3 h
Target group	Doctoral researchers, open to all status groups and externals	
Max. number of participants	IN PERSON:	Limited by the room capacity
	ONLINE:	Limited by VC capacity
Prior knowledge	-	
Materials	Presentations, info-brochures, posters, etc.	
	DISTRIBUTION:	Password-protected distribution among all participants
Technical tools	- Laptop(s) - VC tool with chat and poll-function - Technical equipment for a hybrid event (camera, microphones, beamer etc.)	
Iteration	One iteration per year	
Lecturers	Programme speakers and coordinators, invited speakers	
Tutors	-	
Moderation	Programme staff	
Certification	Attendance will be listed in a certificate of participation which will be distributed once the Starter Track is completed	
Registration	WHEN:	Opens in November/December until 3 days before the event takes place
	WHAT:	Short description of the contents and organisational aspects
	WHO:	Target group can participate in person or online; others can register for online participation or, if there are spaces left, in person
Evaluation	-	

6.1.2 Starter Track

TABLE 7: ORGANISATIONAL SUMMARY OF THE STARTER TRACK.

Format	TYPE:	Interactive thematic overview lectures
	LOCATION:	Hybrid or online
	TIME:	Ideally a fixed day each week from the end of January until mid-May
Workload	DURATION:	2-3 h per lecture
Target group	Doctoral researchers, open to all status groups and externals	
Max. number of participants	IN PERSON:	Limited by the room capacity
	ONLINE:	Limited by VC capacity
Prior knowledge	-	
Materials	Presentations, books, articles, websites, tools, etc.	
	DISTRIBUTION:	Password-protected distribution among all participants
Technical tools	- Laptop(s) - VC tool with chat, poll function, breakout rooms, whiteboard - Technical equipment for a hybrid event (camera, microphones, video projector, etc.)	
Iteration	One iteration in the first half of each year	
Lecturers	Programme speakers and coordinators, invited speakers	
Tutors	-	
Moderation	WHO:	Programme staff
Certification	Attendance at the lectures will be recorded in a certificate of participation distributed once the Starter Track is completed	
Registration	WHEN:	Opens in November/December, closes approx. 3 days before the respective module takes place
	WHAT:	Short description of the learning content, outcomes, motivation and organisational aspects
	WHO:	Target group can register to participate in person or online; others can register to participate online or, if there are spaces left, in person
Evaluation	Online survey after each lecture	

6.1.3 Operator Tracks

TABLE 8: ORGANISATIONAL SUMMARY OF THE OPERATOR TRACKS.

Format	TYPE:	Hands-on workshops
	LOCATION:	In person or online
	TIME:	June to December; Data Steward Track and Data Scientist Track take place biennially
Workload	DURATION:	4 h to 24 h depending on the workshop content
Target group	Doctoral researchers, other status groups and externals, facilities permitting	
Max. number of participants	Decided by the respective lecturer; between 10 and 30 participants	
Prior knowledge	Stipulated for each workshop individually by the lecturers	
Materials	Presentations, installation instructions, exercises, code, data, tools, books, articles, websites, etc.	
	DISTRIBUTION:	Password-protected distribution among all participants
Technical tools	- Laptop(s)	
	ONLINE TOLLS:	R studio server, JupyterHub, Git, cloud services
	ONLINE SESSIONS:	VC tool, chat, polls, breakout rooms, whiteboard
	IN PERSON:	Video projector, whiteboard
Iteration	Biennial	
Lecturers	Internal or invited lecturers	
Tutors	Ideally one or two tutors for individual support	
Moderation	-	
Certification	Attendance at the workshops (full participation) will be recorded in a certificate of participation distributed once the Operator Track is completed	
Registration	WHEN:	Opens in April, closes approx. two weeks before the respective module takes place
	WHAT:	Short description of the learning content, outcomes, motivation and organisational aspects
	WHO:	Target group can register to participate in person or online; others can add themselves to a waiting list
Evaluation	Online survey after each lecture	

6.1.4 Further components

TABLE 9: ORGANISATIONAL SUMMARY OF THE ADDITIONAL COMPONENTS.

	Data Stories	Digital Toolkit	Hacky Hours	Highlight Modules	Data Factory
Format	**Type:** Inspiring talks **Location:** Hybrid or online **Time:** Approx. every 3 months; no fixed schedule	**Type:** Digital catalogue of materials and online courses for self-learning **Time:** On demand	**Type:** Informal meetings with topic-specific discussions **Location:** In person **Time:** Ideally a fixed day every schedule	**Type:** Hands-on workshops and lectures **Location:** In person, hybrid or online **Time:** Additional module within one of the three tracks schedule	**Type:** Summer school **Location:** In person **Time:** Annually, biennially or triennially; 2-4 sequential working days
Workload	**Duration:** 1 h per story and discussion plus get-together for networking	**Duration:** On demand	**Duration:** 1-2 h	**Duration:** 4 h to 24 h depending on the module content	**Duration:** 12 h to 32 h depending on the school's content
Target group	Doctoral researchers, open to all status groups and externals	Doctoral researchers, open to all status groups and externals	Doctoral researchers, open to all status groups and externals	Doctoral researchers, other status groups and externals if capacity available	Doctoral researchers and Post-doctoral researchers; externals if capacity available
Max. number of participants	**In person:** Limited by the room capacity **Online:** Limited by VC capacity	**N/A**	**In person:** Approx. 20	Decided by the lecturer; depending on the format (lecture or workshop), between 10 and 300 participants	Decided by the lecturers; between 10 and 30 participants
Prior knowledge	-	Depending on the material	-	Described for each module individually by the lecturers	Described by the lecturers
Materials	Presentation **Distribution:** Password-protected distribution among all participants	Links to resources	Introduction to the overall topic, data/examples to demonstrate the questions **Distribution:** Password-protected distribution among all participants	Presentations, installation instructions, exercises, code, data, tools, books, articles, websites, etc. **Distribution:** Password-protected distribution among all participants	Presentations, installation instructions, exercises, code, data, tools, books, articles, websites, etc. **Distribution:** Password-protected distribution among all participants
Tools	**Online sessions:** VC tool, chat, polls, breakout rooms, whiteboard **In person:** Video projector, whiteboard	-	• Laptop(s) • Whiteboard • topic-specific software for demonstrations	• Laptop(s) **Online tools:** R studio server, JupyterHub, Git, cloud services **Online sessions:** VC tool, chat, polls, breakout rooms, whiteboard **In person:** Video projector, whiteboard	• Laptop(s), beamer, whiteboard **Online tools:** R studio server, JupyterHub, Git, cloud services

	Data Stories	Digital Toolkit	Hacky Hours	Highlight Modules	Data Factory
Iteration	3-4 Data Stories per year or on demand	-	Approx. 6 Hacky Hours per year or on demand	Max. 1-2 Highlight Modules per track	Every 1-3 years
Lecturers	External and internal speakers	-	Internal lecturers	Internal or invited lecturers	Internal or invited lecturers
Tutors	-	-	Ideally one or two tutors for individual support	Ideally one or two tutors for individual support	Ideally one or two tutors for individual support
Moderation	Programme staff	-	-	None or Programme staff	-
Certification	-	- (no certificate is provided by the programme, but possibly by the individual course providers)	-	Attendance at the workshops (full participation) will be recorded in a certificate of participation distributed once the track is completed	Attendance at the summer school (full participation) will be recorded by a certificate of participation
Registration	**When:** Opens approx. a month before the talk takes place **What:** Short description of the content, introduction by the speaker and organisational aspects **Who:** Target group can register to participate in person or online; others can add themselves to a waiting list	-	**When:** Opens one month before a meeting takes place **What:** Short description of the topic, introduction of the expert(s) and organisational aspects **Who:** Open to everyone	According to the track's regulations (see above)	**When:** Opens approx. two months before the summer school takes place **What:** Short description of the learning content, outcomes and motivation, introduction by the expert(s) and organisational aspects **Who:** Doctoral researchers and post-doctoral researchers
Evaluation	-	-	-	Online survey after each module	Online survey after the summer school

6.2 Data Train – module descriptions

All modules are contributed and developed voluntarily by lecturers (mostly) from the U Bremen Research Alliance.

TABLE 10: OVERVIEW OF THE DATA TRAIN MODULES AND RESPECTIVE LECTURERS.

STARTER TRACK MODULES

	Title	Data Train lecturers	Duration (h)	Page
DATA SCIENCE, STATISTICAL THINKING, CRITICAL THINKING	Data science and big data	*Björn Tings, Karl Kortum, Dr. James Imber, Dmitrii Murashkin* Remote Sensing Technology Institute of German Aerospace Center (DLR)	2	70
	Statistical thinking	*Prof. Dr. Iris Pigeot* Leibniz Institute for Prevention Research and Epidemiology – BIPS; Faculty of Mathematics and Computer Science; NFDI4Health	2	76
	Asking (the right) research questions in data science	*Prof. Dr. Vanessa Didelez* Leibniz Institute for Prevention Research and Epidemiology – BIPS; University of Bremen, Faculty of Mathematics and Computer Science	2.5	64
	Philosophical reflections on data science	*Prof. Dr. Dr. Norman Sieroka* University of Bremen, Faculty of Cultural Studies	2	74
	Digital ethics	*Björn Haferkamp* University of Bremen, Faculty of Cultural Studies	2	69
	About the meaningfulness of data	*Prof. Dr. Hans-Christian Waldmann* University of Bremen, Faculty of Human and Health Sciences	2	63
RESEARCH DATA MANAGEMENT	Data and information management	*Prof. Dr. Frank Oliver Glöckner* Alfred Wegener Institute (AWI), Helmholtz-Center for Polar and Marine Research; German Federation for Biological Data (GFBio e.V.); University of Bremen, Faculty of Geosciences; NFDI4Biodiversity *Dr. Ivaylo Kostadinov* German Federation for Biological Data (GFBio e.V.); NFDI4Bidiversity	2	67
	Data protection and licences	*Prof. Dr. Dennis-Kenji Kipker* University of Bremen, Faculty of Law; University of Applied Science, Bremen; NFDI4Health *Prof. Dr. Benedikt Buchner* Augsburg University, Faclty of Law; NFDI4Health	2	68
	Managing confidential data	*Dr. Martin Dörenkämper, Dr. Julia Gottschall* Fraunhofer Institute for Wind Energy System (IWES	2	71
	Managing qualitative data	*Prof. Dr. Betina Hollstein* University of Bremen, Faculty of Social Sciences, KonsortSWD *Dr. Jan-Ocko Heuer* University of Bremen, Faculty of Social Sciences, KonsortSWD	2	72

STARTER TRACK MODULES				
	Title	**Data Train lecturers**	**Duration (h)**	**Page**
COMPUTER SCIENCE	Computer science basics	*Prof. Dr. Rolf Drechsler* German Research Center for Artificial Intelligence (DFKI); University of Bremen, Faculty of Mathematics and Computer Sciences	2	65
	Overview about programming languages	*Prof. Dr. Christoph Lüth* German Research Center for Artificial Intelligence (DFKI); University of Bremen, Faculty of Mathematics and Computer Sciences	2	73
	Cryptography basics	*Prof. Dr. Dieter Hutter* German Research Center for Artificial Intelligence (DFKI); University of Bremen, Faculty of Mathematics and Computer Sciences	2.5	66
	Security and privacy	*Prof. Dr. Dieter Hutter* German Research Center for Artificial Intelligence (DFKI); University of Bremen, Faculty of Mathematics and Computer Sciences	2	75
		Track total	30.5	

DATA STEWARD TRACK MODULES

	Title	Data Train lecturers	Duration (h)	Page
PROGRAMMING	Getting started with R	*Dr. Christian Fieberg* University of Bremen, Faculty of Business Studies and Economics	20	83
	Erste Schritte mit MATLAB (First steps with MATLAB)	*Dr. Christian Fieberg* University of Bremen, Faculty of Business Studies and Economics	20	81
	Getting started with Python	*Dr. Nikolay Koldunov* Alfred Wegener Institute (AWI), Helmholtz-Centre for Polar and Marine Research	15	82
DATA HANDLING, DATA ENGINEERING	How to write a DMP?	*Prof. Dr. Frank Oliver Glöckner* Alfred Wegener Institute (AWI), Helmholtz-Centre for Polar and Marine Research; German Federation for Biological Data (GFBio e.V.); University of Bremen, Faculty of Geosciences; NFDI4Biodiversity *Dr. Ivaylo Kostadinov* German Federation for Biological Data (GFBio e.V.); NFDI4Bidiversity *Jimena Linares* German Federation for Biological Data (GFBio e.V.); NFDI4Bidiversity *Tanja Weibuat* State Natural Science Collections of Bavaria (SNSB); NFDI4Biodiversity	4	85
	Versioning with Git/GitHub	*Nico Harms* Alfred Wegener Institute (AWI), Helmholtz-Centre for Polar and Marine Research	6	84
	Reproducibility in science: How and why?	*Dr. Arjun Chennu* Leibniz Centre for Tropical Marine Research (ZMT)	12	86
	Data base skills	*Prof. Dr. Sebastian Maneth* University of Bremen, Faculty of Mathematics and Computer Sciences	9	77
	Data preparation, climate model data	*Dr. Nikolay Koldunov* Alfred Wegener Institute (AWI), Helmholtz-Centre for Polar and Marine Research	4	79
	Data preparation, qualitative data	*Dr. Jan-Ocko Heuer* University of Bremen, Faculty of Social Sciences, KonsortSWD	4	80
	Data extraction from external online resources using R	*Prof. Dr. Kristina Klein* University of Bremen, Faculty of Business Studies and Economics	12	78
		Track total	114	

DATA SCIENTIST TRACK MODULES

	Title	Data Train lecturers	Duration (h)	Page
METHODS AND TECHNIQUES FOR DATA SCIENCE METHODS (INCLUDING ARTIFICIAL INTELLIGENCE METHODS)	Quantitative analyses for data science	*Prof. Dr. Thorsten Dickhaus* University of Bremen, Faculty of Mathematics and Computer Sciences	12	94
	Causal learning	*Prof. Dr. Vanessa Didelez* Leibniz Institute for Prevention Research and Epidemiology – BIPS; University of Bremen, Faculty of Mathematics and Computer Science	10	87
	Machine learning	*Prof. Dr. Marvin Wright* Leibniz Institute for Prevention Research and Epidemiology – BIPS; University of Bremen, Faculty of Mathematics and Computer *Science*	18	93
	Deep learning	*Prof. Dr. Peter Maas, Dr. Daniel Otero Baguer* University of Bremen, Faculty of Mathematics and Computer Science	10	90
	Evaluation machine learning/ artificial intelligence algorithms	*Prof. Dr. Werner Brannath* University of Bremen, Faculty of Mathematics and Computer Sciences *Dr. Max Westphal* Fraunhofer Institute for Digital Medicine (MEVIS)	16	91
	Hardware and system architectures for data science	*Christopher Metz* University of Bremen, Faculty of Mathematics and Computer Sciences	1.5	92
	Computational Social sciences	*Dr. Nikolitsa Grigoropoulou* University of Bremen, Faculty of Social Sciences	8	88
	Visual analytics with Geographical Information Systems (GIS)	*Dr. Antonie Haas* Alfred Wegener Institute (AWI), Helmholtz-Centre for Polar and Marine Research	16.5	96
		Track total	92	

6.3 Data Train – Individual modules

6.3.1 Starter Track

Starter Track Module: About the meaningfulness of data (2 hours)

Lecturer, affiliation: *Prof. Dr. Hans-Christian Waldmann, Univ. Bremen*

Background:

Data are not, as etymology suggests, 'the given', but are generated, constructed, or *made* (sometimes in the negative sense of the word). Therefore, we need to shed some light onto the hidden presuppositions in our scientific agenda. To get started, we assume that there is no meaning in the data per se; instead, we posit that meaning *happens* to data and is attached to it. In fact, YOU attach it, meaning you must assume liability for establishing a referential link between the data themselves and the phenomenon that your data are supposed to capture and whose epistemological status ("I got it, you see") and ontological status ("It exists, I mean, like 'really' there") might not be the same. You may find that technological sophistication and programming skills might not suffice here. In order to identify how your scientific attitudes and your decision making as a researcher along the rocky road of empirical research adds, detracts from, or alters the meaningfulness of data, we will cast a wide net, ranging from epistemological paradigms to specific choices of statistical models in handling missing data, bridged by measurement theory and its map of pathways from the (in-) tangible world to numbers. Welcome to the vast realm of philosophy of science.

Contents:

1. Notions of meaning, data, and information
2. Epistemology and ontology: how data refer to what is being measured
3. The ideal research process: are data decisive? A menu of paradigms
4. Data and theory. Realism – Anti-realism – Pragmatism. Models. Truth.
5. Introduction to measurement theory: do you abide by the rules?
6. Fuzziness, vagueness, uncertainty, incompleteness: 'bad' data?
7. Missing data: how your philosophical stance indeed impacts study results

Learning objectives:

Since, as a psychologist and statistician, I cannot claim expertise in your respective field of work, I will not, and cannot, tell you how to 'do it right'. But the patterns behind 'doing it wrong' are quite universal: unawareness and lack of transparency. My aim is to make some of the implicit explicit, and foster a critical mindset when it comes to relating data to meaning in your specific discipline.

6.3.1.1 Starter Track Module:

Asking the right research questions in data science (2.5 hours)

Lecturer, affiliation: *Prof. Dr. Vanessa Didelez – Leibniz Institute for Prevention Research and Epidemiology – BIPS; University of Bremen, Faculty of Mathematics and Computer Science*

Background:

"An approximate answer to the right question is worth a great deal more than a precise answer to the wrong question" said the renowned statistician John Tukey as early as 1969.

Based on my own experience in statistical consultations, a great deal of confusion arises due to a mismatch between research questions and data/methods. However, even more fundamentally, the research question is often not even clearly articulated at the outset – perhaps because researchers anticipate that the right question can only be answered approximately. But how can we discuss what data and methods are suitable if we are unclear or vague about the question to be answered? It seems that now, in the era of big data characterised by an abundance of data and a similar abundance of methods for analysing the data, the issue of asking the right question now has a new sense of urgency.

Contents:

In this course we will discuss the different types of research questions one might face in a variety of applied fields within data science, such as psychology, epidemiology, genetics, or political and social sciences. Key distinctions concern questions that are (i) descriptive, (ii) predictive, or (iii) causal (i.e. about counterfactual prediction). We will consider how these types of research questions are interrelated with the choices/requirements of data, methods of analysis, and the need for more or less specific background knowledge of the subject matter. We will see how starting with a clear and explicit research question helps with assessing, and perhaps avoiding, potential sources of (structural) bias in answering that research question.

Key topics covered:

- Types of research questions (descriptive, predictive, causal/counterfactual)
- Issues of validity and structural bias (e.g. selection, confounding, ascertainment)
- The target trial principle

Learning objectives:

Participants will be able to

- categorise research questions as descriptive, predictive or causal
- elicit a research question by formulating a target trial
- determine implications for the required data and choice of appropriate methods
- identify possible threats to the validity/sources of structural bias.

6.3.1.2 Starter Track Module:

Computer science basics for data science (2 hours)

Lecturer, affiliation: *Prof. Dr. Rolf Drechsler – German Research Center for Artificial Intelligence (DFKI); University of Bremen, Faculty of Mathematics and Computer Sciences*

Background:

Computer science is a key component for data science applications and research data management due to the fact that methods and procedures rely on it. For instance, to enable fast access to information, data sets must be stored efficiently in data structures. Clever modelling and algorithmic processing guarantee a fast search and selection of information, even in big data sets. This course will provide insights into the basics of computer science and an overview of topics relevant to data science.

Contents:

- Computer science and its subdisciplines: applied, technical, practical, theoretical
- Programming languages
- Data storage and processing
- Data structures
- Example: Sorting (Bubble Sort, Merge Sort, Quicksort)

Learning objectives:

Participants will acquire a basic overview of computer science and its subdisciplines, along with the basics of system engineering.

6.3.1.3 Starter Track Module:

Cryptography basics (2 hours)

Lecturer, affiliation: *Prof. Dr. Dieter Hutter – German Research Center for Artificial Intelligence (DFKI); University of Bremen, Faculty of Mathematics and Computer Sciences*

Background:

Cryptography is the key technology employed to ensure the security and privacy of IT systems. An understanding of the basic principles of cryptographic functions is an indispensable prerequisite for the development of modern IT systems.

Contents:

This course will provide basic theoretical and practical knowledge of cryptography, including asymmetric vs. symmetric encryption, cryptographic hash functions, digital signatures and public-key infrastructures, and post-quantum cryptography.

Learning objectives:

Participants will acquire the basics of cryptography, with a focus on being able to assess the strength of cryptographic methods in practice.

6.3.1.4 Starter Track Module:

Data & information management (3 hours)

Lecturers, affiliation: *Prof. Dr. Frank Oliver Glöckner – Alfred Wegener Institute (AWI), Helmholtz-Center for Polar and Marine Research; German Federation for Biological Data (GFBio e.V.); University of Bremen, Faculty of Geosciences; NFDI4Biodiversity*

Dr. Ivaylo Kostadinov – German Federation for Biological Data (GFBio e.V.); NFDI4Bidiversity

Background:

The comprehensive management of research data is integral to each research project and an essential aspect of sound scientific practice. It accompanies each phase of a research project, ranging from the proposal phase to data acquisition and data analyses and on to the publication phase. The overall goal of research data management is the production of findable (F), accessible (A), interoperable (I) and reusable (R), i.e. FAIR, data sets.

Good stewardship of data (following the FAIR principles; Wilkinson et al., 2016) and an open data culture (Nosek et al., 2015) foster reproducibility as well as sustainability in science, while forming the basis for data science applications.

Contents:

1. Research data: Data life cycle and accompanied challenges
2. Data management plans (DMP)
3. FAIR data principles
4. Metadata: standardisation and its significance
5. Archiving, publication and citation of research data sets

Learning objectives:

Participants will understand the significance of research data management and acquire an overview of concepts and approaches.

6.3.1.5 Starter Track Module:

Data protection and licences (2 hours)

Lecturer, affiliation: *Prof. Dr. Dennis-Kenji Kipker – University of Bremen, Faculty of Law; University of Applied Science, Bremen; NFDI4Health*

Prof. Dr. Benedikt Buchner – Augsburg University, Faclty of Law; NFDI4Health

Background:

Compliance with legal requirements in the handling of research data is an indispensable requirement for the long-term success of research data management.

Contents:

Legal framework of research data management with a special focus on questions of copyright law and data protection law.

Learning objectives:

Participants will acquire a basic legal knowledge of the possibilities and limitations of research data management.

6.3.1.6 Starter Track Module:

Digital ethics (2 hours)

Lecturer, affiliation: *Björn Haferkamp*

University of Bremen, Faculty of Cultural Studies

Background:

Digital ethics is of particular interest to data scientists, engineers, and managers because their results and products have a direct impact on individuals and society. Ethics involves respecting the dignity and vulnerability of human beings while contributing to a good life. Ethics for Data Science – like Medical Ethics or Engineering Ethics – is aspirational (providing values for goals to achieve) and preventive (providing tools for understanding and avoiding problematic outcomes). Given the numerous areas of application, the range of topics can be as varied and complex as it is fascinating.

Contents:

- Thematic areas of digital ethics
- Introduction to theories and tools for digital ethics
- Types of ethical concerns with algorithms
- Values for digital ethics

Learning objectives:

Participants will acquire an understanding of the many areas of ethical concern arising from digital technologies, and learn about the various principles and values at play when dealing with these concerns. A basic map of tools for ethical thinking will be presented to offer guidance for further research into questions of digital ethics

6.3.1.7 Starter Track Module:

Data science & big data (2 hours)

Lecturer, affiliation: *Björn Tings, Karl Kortum, Dr. James Imber, Dmitrii Murashkin*

Remote Sensing Technology Institute of German Aerospace Center (DLR)

Background:

The digital transformation has seen the development of a novel scientific discipline – data science. Data science allows new approaches for interdisciplinary (big) data analyses through complex algorithms and artificial intelligence machine learning, deep learning, etc.). Such approaches extract information from data sets that is beyond the scope of current scientific knowledge. Consequently, data science is of interest to practically every field of research, industry and the economy, which is why it is often referred to as a novel key discipline (e.g. Society of Informatics e.V., 2019).

This course provides a basic overview of the history, terminology and application of data science, and provides references to more detailed data train teaching. Ethical, legal, and social implications are provided and then addressed in subsequent courses and operator tracks.

Contents:

1. History
   - From the first computers and AIs to modern-day data science
   - Timeline comparison with CPU power and storage costs
2. Clarification of terms
   - Statistics <-> Machine Learning <-> Deep Learning
   - Data Mining <-> Big Data
   - Machine Learning <-> Artificial Intelligence
   - Data Collection, Analysis, Visualisation
3. Essentials of Data Science and Big Data
   - Machine Learning
     - Supervised Learning
     - Unsupervised Learning
     - Reinforcement Learning
   - Big Data, a combination of Data Science technologies
     - Five Vs Model
     - Data Privacy
4. Tools

Learning objectives:

Participants will acquire a basic overview of the history, terms, applications, methods, tools and implications of data science and big data.

6.3.1.8 Starter Track Module:

Managing confidential data (2 hours)

Lecturer, affiliation: *Dr. Martin Dörenkämper, Dr. Julia Gottschall*

Fraunhofer Institute for Wind Energy System (IWES)

Background:

Many data are classified as confidential because they contain sensitive personal or institutional information. Their confidentiality limits the use of such data, but they can still be of great benefit to research provided they are used and managed properly. In this course we will focus on data collected in applied and industry-related context-specific applications, using wind energy research as an example. We will discuss the challenges that arise in this context and how to deal with them to generate the best-possible research output (not least with the requirements of open science in mind).

Contents:

1. Classification of confidential data
2. Specific requirements for managing confidential data
3. Approaches for doing (open) science based on confidential data

Learning objectives:

Participants will gain an understanding of the specific requirements of confidential data and learning methods for working with/conducting research based on confidential data.

6.3.1.9 Starter Track Module:

Managing qualitative data (3 hours)

Lecturers, affiliation:

Prof. Dr. Betina Hollstein, SOCIUM Research Center on Inequality and Social Policy at the University of Bremen, Professor for Microsociology and Qualitative Methods at the University of Bremen, and Head of the Research Data Center (RDC) Qualiservice

Dr. Jan-Ocko Heuer, SOCIUM Research Center on Inequality and Social Policy at the University of Bremen, Postdoctoral Researcher at the Research Data Center (RDC) Qualiservice

Background:

Many disciplines within the social and cultural sciences and humanities deal with 'qualitative' data, which means that in general, the materials not only contain personal information, but also non-standardised forms of such information, thus posing particular challenges for research data management. Examples of qualitative materials include various types of text (e.g. interview transcripts, observation protocols, field notes), images, audiovisual data, or material artefacts. From a 'quantitative' research perspective, i.e. the application of statistical methods to standardised numerical data, qualitative materials merely appear to be data requiring additional structure. But qualitative material is a specific type of data that is usually richer, more context-dependent and more sensitive than quantitative data. On the other hand, qualitative data can be fruitfully analysed with common tools of quantitative inquiry (e.g. text mining). Thus, this lecture addresses both quantitative and qualitative researchers, and aims to introduce them to the ethical, legal and practical challenges of managing qualitative materials, e.g. in terms of data protection, informed consent, anonymisation, documentation and data sharing, so as to outline good practices of research data management.d).

Contents:

1. Introducing qualitative data and research
    - What is qualitative research? Aims, examples and characteristics
    - Quantitative versus qualitative data and research processes
    - Methodology, context and data in a qualitative inquiry
    - Mixing qualitative and quantitative data and research

2. Managing qualitative data in practice
    - Data collection and informed consent
    - Data transformations for analysis
    - Anonymisation/pseudonymisation
    - Documentation/contextualisation
    - Archiving and sharing data
    - Finding and reusing data

Learning objectives:

Participants will acquire a basic overview of qualitative research data and their management.

6.3.1.10 Starter Track Module:

Overview of programming languages (2 hours)

Lecturer, affiliation: *Prof. Christoph Lüth*

German Research Centre for Artificial Intelligence (DFKI) and the University of Bremen

Background:

Programming is essential for managing data sets and conducting data science work. Handling huge data sets manually is impossible, so we can only prepare, curate, analyse and evaluate them by way of programming. In addition, programming is crucial for documenting, creating graphical output, and presenting results (e.g. on the web). In order to write programs, we need to learn a *programming language* – but what is that?

Contents:

- What is a programming language, actually?
- What characterises a programming language and what purpose does it serve?
- Why is HTML not a programming language and what has Alan Turing got to do with it?

Approximately 700 programming languages exist – so how can we know about and keep pace with all of this? We learn to distinguish languages based on their degree of abstraction and programming paradigm (imperative, procedural, object-oriented, functional, logical, etc.), or their area of application. Furthermore, we determine how to choose the appropriate language for the task at hand, which programming languages you need to know, and see some of them briefly presented in this course.

Learning objectives:

Participants acquire an overview of programming languages, their features, significance and criteria for distinction.

6.3.1.11 Starter Track Module:

Philosophical reflections on data science (2 hours)

Lecturer, affiliation: *Prof. Dr. Dr. Norman Sieroka – University of Bremen, Faculty of Cultural Studies*

Background:

Critical awareness (see 'Critical Thinking' below) is crucial for an appropriate and reasonable assessment of data preparation, sharing and utilisation in the context of research data management, data protection and data science applications.

Critical thinking facilitates the establishment of common languages across disciplines while being aware of limits and difficulties, in turn making it essential for cooperative and future-oriented research.

Contents:

This session introduces the concept of critical thinking and provides a rough overview of broad societal concerns and philosophical debates surrounding data science and artificial intelligence. For instance, it seems inevitable that we are set to lose some of our own autonomy once our cars start driving autonomously and our houses become smarter and smarter. Computers have been outperforming us at number-crunching for decades now, but will they also outsmart us when it comes to creativity? Will they become the 'better scientists' or will there always remain a difference between 'pure prediction' and 'real understanding'? Is predictive success acceptable even if accompanied by a loss of transparency? After all, transparency is something we are very much worried about, not only in science but in all kinds of political and societal contexts. At the same time, privacy and data protection laws are a major issue in public discourse as well. Consider tracking apps, for instance— do we really want to become transparent citizens and consumers, X-rayed as it were by a machine learning algorithm that perhaps nobody actually understands?

Learning objectives:

Participants will engage in critical thinking to reflect critically on their own research/work and develop an appreciation of other disciplines, their mindsets and ways of thinking.

6.3.1.11 Starter Track Module:

Security & privacy (2 hours)

Lecturer, affiliation: *Prof. Dr. Dieter Hutter – German Research Center for Artificial Intelligence (DFKI); University of Bremen, Faculty of Mathematics and Computer Sciences*

Background:

Security and privacy are key aspects in developing and maintaining trustworthy systems. A lack of security results in vulnerable systems that are exposed to and unprotected against potential attackers, in turn presenting an incalculable economic and personal risk. As personal data has become the new currency in the digital era, its protection from unauthorised processing and distribution is a key issue when it comes to preserving the privacy and self-determination of individuals.

Contents:

Techniques to measure and enhance the security and privacy of IT systems

- Security: security protocols, security policies and their enforcement
  (e.g. access control, dataflow control)
- Privacy: GDPR, privacy-enhancing techniques
  (e.g. differential privacy, k-anonymity)

Learning objectives:

Participants will acquire basic knowledge of security and privacy techniques, and outline their underlying foundations.

6.3.1.12 Starter Track Module:

Statistical thinking (2 hours)

Lecturer, affiliation: *Prof. Dr. Iris Pigeot – Leibniz Institute for Prevention Research and Epidemiology – BIPS; Faculty of Mathematics and Computer Science; NFDI4Health*

Background:

Data science approaches are based on statistical/mathematical methods as well as computer science competencies. In this context, it is crucial to understand the basic principles of statistical methods as this will help to adequately apply statistical methods and to produce reliable statistical results.

Contents:

This course provides an introduction to statistical basics and concepts relevant for data science applications. After a brief presentation of the categories of statistics (descriptive, predictive, confirmatory) and their general ideas, selected basic methods will be explained and illustrated using practical examples: concept of probability, parameter estimation, confidence intervals and testing of hypotheses.

Learning objectives:

Participants will acquire a basic understanding of the major statistical principles.

6.3.2 Operator Track 'Data Steward'

6.3.2.1 Operator Track 'Data Steward' Module:

Database skills (9 hours)

Lecturer, affiliation: *Prof. Dr. Sebastian Maneth*

University of Bremen, Faculty of Mathematics and Computer Sciences

Background:

Relational data are ubiquitous, with the majority of data stored in relational database management systems (RDBMS). Anyone needing to analyse and query large data will come across RDBMS and will need to know how to use them and interface with them. Moreover, the value of knowing the ins and outs of the query language SQL in terms of scientific insight as well as market advantage cannot and should not be overlooked. According to a 2017 Stackoverflow Developer Survey, SQL is the second-most-popular programming language in use.

Contents:

This course will give a hands-on introduction to relational database management systems (RDBMS). You will learn how to structure your database and how to create tables within an RDBMS. You will learn how to import data from CSV files into your tables. We will use a large bibliography database as our example database. The main part of the course will about learning how to formulate interesting queries in SQL. We will also learn how to display the results of queries using other software (such as Python, R, of gnuplot). The ability to query and display results offers a powerful data analytics platform. We will discuss the limitations of RDBMs and scenarios in which modern NoSQL database systems are able to address and overcome these limitations.

Learning objectives:

Participants will learn how to structure data into relational tables, how to create such tables in a relational database management system (RDBMS), and how to query the data using the SQL query language. This will be accompanied by hands-on experience in formulating interesting queries on a large database.

Prior knowledge:

No prior knowledge is required, other than being able to run SQLite3 on your device, or how to run SSH on your device.

Technical requirements:

- Own PC, laptop
- Internet (access to eduroam), web browser (up to date)

Each participant should be able to run the SQLite3 database system on their own device (preferably a laptop or desktop machine). Alternatively, it is sufficient to know how to run SSH and we will provide a Linux-based machine for participants to log in via SSH to run SQLite3 remotely.

6.3.2.2 Operator Track 'Data Steward' Module:

Data extraction from external online platforms using R (12 hours)

Lecturer, affiliation: *Prof. Dr. Kristina Klein*

University of Bremen, Faculty of Business Studies and Economics

Background:

This course is open to anyone interested in analysing (and learning from) data from external (online) sources such as review platforms, Twitter, etc.

Contents:

The course will cover some basic insights of extracting data from external online platforms. It will discuss the general approaches as well as the legal requirements. We will work with R to program simple scrapers that systematically extract data from websites, and access application programming interfaces (APIs) that facilitate the extraction of data.

Learning objectives:

In this course, students will learn how to acquire data from external sources (such as online platforms) for follow-up statistical analysis for their own research.

Prior knowledge:

Some prior knowledge of R would be beneficial.

Technical requirements:

- Own PC, laptop
- Internet, web browser (up to date)
- For online format, a second screen might be beneficial
- Installation of latest version of R / R studio; participants will receive installation instructions prior to the workshop

6.3.2.3 Operator Track 'Data Steward' Module:

Data preparation: Climate model data (4 hours)

Lecturer, affiliation: *Nikolay Koldunov, Alfred Wegener Institute for Polar and Marine Research*

Background:

Data provide answers to all manner of questions, and analysis methods can extract such answers using 'data preparation' to link the two. Anyone interested in working with data will likely need to know at least some of the principles outlined.

Furthermore, cross-discipline data analysis is part of scientific progress. This workshop offers the opportunity to look into data preparation procedures for different data types, and to learn about their individual characteristics.

Contents:

This course will cover the following topics:

- Where to find climate model information
- netCDF data format
- Basic types of atmospheric, ocean, land and sea ice data
- Validation of model data
- Ways to extract weather and climate information for particular regions and times in the past and future

Course on GitHub: https://github.com/koldunovn/DT_model_data

Learning objectives:

Participants will gain a basic understanding of the strengths and weaknesses of data from Earth System models. This includes information on what kinds of data on weather and climate are available, how to get the data, and how to extract and post-process it.

Prior knowledge:

Some exercises will require a degree of familiarity with the Python programming language. Some specific extension packages will be used, but prior knowledge of these will not be assumed. Basic experience with programming in any language would be an advantage.

6.3.2.4 Operator Track 'Data Steward' Module:

Data preparation: Qualitative data (4 hours)

Lecturer, affiliation: *Dr. Jan-Ocko Heuer, SOCIUM Research Center on Inequality and Social Policy at the University of Bremen, Postdoctoral Researcher at the Research Data Center (RDC) Qualiservice*

Background:

Data holds the answers to all manner of questions and analysis methods can extract those answers - linking the two is "data preparation". Anyone interested in working with data will likely need to know at least some of the principles outlined.

Furthermore, cross-discipline data analysis is part of scientific progress. This workshop offers the opportunity to look into data preparation procedures for different data types and to learn about their individual characteristics.

Contents:

Social scientists usually deal with information about human beings, which poses particular ethical, legal and practical challenges for collecting, transforming, and analyzing such data and materials. These challenges are even more pronounced in 'qualitative' research, which involves open and flexible research processes and usually produces heterogeneous, complex, very information-rich and highly context-dependent materials. This hands-on workshop aims to familiarise participants with the major challenges of preparing qualitative social science materials for analysis, and to introduce them to good practices and RDM tools to deal with those challenges. Topics include data protection and informed consent, anonymisation, documentation, and sharing beyond the research project of origin. Particular attention is given to appropriate research documentation using a 'study report' and various context materials, and to the anonymisation/pseudonymisation of qualitative text data, including an introduction to the 'QualiAnon' tool developed by the Research Data Center Qualiservice (Nicolai et al. 2021).

Tom Nicolai, Kati Mozygemba, Susanne Kretzer, Betina Hollstein (2021): QualiAnon - Qualiservice tool for anonymising text data. Qualiservice. University of Bremen. For information and access, visit: https://www.qualiservice.org/de/help-desk/webinar/tools.html

This course will cover the following topics:

- What are qualitative data? Examples and characteristics
- Why, for whom and how are qualitative data important?
- Producing and preparing qualitative data for research/analysis
- Focus: documenting qualitative data and contexts
- Focus: anonymising qualitative (text) data

Prior knowledge:

-

Requirements:

Own computer with modern web browser; Wi-Fi: Access to eduroam

6.3.2.5 Operator Track ‘Data Steward’ Module:

First steps with MATLAB (approx. 20 hrs.)

Lecturer, affiliation: *Dr. Christian Fieberg*

University of Bremen, Faculty of Business Studies and Economics

Background:

Programming language skills such as Python, R or MATLAB are a necessary prerequisite for working with big data and applying advanced data science methods.

Contents:

- MATLAB basics
- Importing and exporting data
- Programming basics
- Data types
- Scripts
- Functions
- Basic plotting tools

Learning objectives:

Participants will learn how to use MATLAB in connection with data science methods and turn data into findings.

Prior knowledge:

-

Requirements:

- Own PC/ laptop
- Internet, web browser (up to date)
- For online format, a second screen might be beneficial
- MATLAB licence
- Installation will be performed during the workshop

6.3.2.6 Operator Track 'Data Steward' Module:

Getting started with Python (15 hours)

Lecturer, affiliation: *Nikolay Koldunov, Alfred Wegener Institute for Polar and Marine Research*

Background:

Python is currently one of the most popular general purpose programming languages, and continues to gain popularity due to its expressive syntax, vibrant community and large number of high-quality open-source libraries. Such libraries include those for data science, all branches of natural and computer sciences as well as tools for system administration and development of web applications and backend systems. Most of the popular machine learning libraries (not covered in this course) are written with Python as an interface language (TensorFlow, Keras, PyTorch). Parallel processing libraries (e.g. dask) allow people to work effectively with huge amounts of data, in turn making knowledge of Python desirable among data scientists as well as anyone working with data in various digital forms.

Contents:

In this course you will learn the basics of Python so you can perform data handling-related tasks with this programming language. The main emphasis will be placed on building up an overview and hands-on experience with the most popular Python libraries used for data processing and visualisation. In particular, we will cover numpy (array operations), pandas (table data and statistics), xarray (labeled arrays), dask (parallel data processing), scipy (scientific computing), matplotlib and bokeh (visualisation). The course will be held in an interactive Jupyter environment.

Course on GitHub: https://github.com/koldunovn/python_data_train

Learning objectives:

Participants will gain a basic understanding of Python syntax, data types and operations, along with basic information about several main Python libraries used for data processing and visualisation. Participants will also gather experience in solving simple data handling problems using those libraries, as well as experience in working within an interactive Jupyter environment. Finally, participants will acquire knowledge of the Python data science libraries ecosystem and learn where to find information about such libraries.

Prior knowledge:

Basic experience with programming in any language would be of benefit.

6.3.2.7 Operator Track ‘Data Steward’ Module:

Getting started with R (20 hours)

Lecturer, affiliation: *Dr. Christian Fieberg*

University of Bremen, Faculty of Business Studies and Economics

Background:

Programming language skills such as Python, R or MATLAB are a necessary prerequisite for working with big data and applying advanced data science methods.

Contents:

- R basics
- Importing and exporting data
- Programming basics
- Data types
- Scripts
- Functions
- Basic plotting tools

Learning objectives:

Participants will learn how to use R in connection with data science methods and turn data into findings.

Prior knowledge:

-

Requirements:

- Own PC, laptop
- Internet, web browser (up to date)
- For online format, a second screen might be beneficial
- Installation will be performed during the workshop

6.3.2.8 Operator Track 'Data Steward' Module:

Git / GitHub (6 hours)

Lecturer, affiliation: *Nico Harms – Alfred Wegener Institute (AWI), Helmholtz-Centre for Polar and MarineResearch*

Background:

Git is a powerful tool used for managing and tracking changes to code and documents. Paired with platforms like Github, it allows users to collaborate, automate tasks, and simplify their workflows. This course is designed to introduce Data Stewards to the basics of Git and Github, providing them with the skills and knowledge they need to use these tools effectively. Git's decentralised functionality also makes it ideal for use in environments with limited or no internet access.

It is worth noting that there are other tools similar to Github, such as GitLab and Bitbucket, and the principles learned in this course can be applied to these platforms as well.

Contents:

In this course, you will learn the basics of using Git and Github for managing and collaborating on projects. We will focus on using the command-line interface (CLI) tools of Git, and take a quick look at using graphical user interfaces (GUIs). Through a series of lessons and hands-on exercises, you will learn how to create projects, track changes, collaborate with others, automate tasks, and more. By the end of this course, you will have a solid foundation in using Git and Github for your projects.

Learning objectives:

By the end of this course, you will have a solid foundation in Git and be able to use it to simplify your workflow and collaborate with others on complex projects.

Prior knowledge:

This course is designed for beginners with no prior experience in Git or Github. If you are unfamiliar with the command-line interface (CLI), don't worry! We will provide a step-by-step guide to help you get up and running.

Requirements:

Participants are required to have their own PC or laptop, as well as an up-to-date web browser and a stable internet connection. It is also recommended that participants have Git installed on their device (https://git-scm.com/downloads) and have a Github account (https://github.com/) to fully participate in the course. Installation instructions will be provided prior to the workshop.

6.3.2.9 Operator Track 'Data Steward' Module:

How to write a data management plan (4 hours)

Lecturers, affiliation: *Prof. Dr. Frank Oliver Glöckner – Alfred Wegener Institute (AWI), Helmholtz-Centre for Polar and Marine Research; German Federation for Biological Data (GFBio e.V.); University of Bremen, Faculty of Geosciences; NFDI4Biodiversity*

Dr. Ivaylo Kostadinov – German Federation for Biological Data (GFBio e.V.); NFDI4Bidiversity

Jimena Linares – German Federation for Biological Data (GFBio e.V.); NFDI4Bidiversity

Tanja Weibuat – State Natural Science Collections of Bavaria (SNSB); NFDI4Biodiversity

Background:

Research data are constantly increasing in size and complexity. Besides new discovery opportunities, this also presents new challenges for managing the data. Key aspects here include the possibility to integrate data across scientific domains and to describe and publish data in a way that makes them reusable for future generations of scientists. Interest in this is also shared by the funding agencies and the general public. Like any successful venture, good data management starts with a plan, namely a Data Management Plan, which outlines in detail what data will be produced and how the data will be handled, preserved and shared.

Contents:

Participants will receive:

- A general introduction to Research Data Management (RDM), covering trans-disciplinary concepts
- A subject-specific focus on RDM for biodiversity and ecological research
- A detailed introduction to the concept of Data Management Plans (DMPs)
- An introduction to the GFBio service for DMP support
- Hands-on experience in creating a DMP according to the DFG Guidelines on the handling of Research Data in Biodiversity Research

Learning objectives:

Participants will acquire:

- A detailed overview of the most important aspects of RDM
- The ability to create their own DMP

The concepts and skills taught in this course are transferable to other disciplines.

Requirements:

- Own PC, laptop
- Internet, web browser (up to date)
- For online format, a second screen might be beneficial

6.3.2.10 Operator Track 'Data Steward' Module:

Reproducibility in science: How and why (12 hours)

Lecturers, affiliation: *Dr. Arjun Chennu*

Leibniz Centre for Tropical Marine Research (ZMT)

Background:

The reproducibility crisis in science stems not only from historically poor data availability, but also from a lack of the context used to glean knowledge from the data. Reproducible science seeks to package the analytical context of data – software environment, data organisation, analytical interdependencies, expert comments – into an operational product. This provides excellent benefits in multiplying the impact and usefulness of your scientific work for scientists, journalists – and yourself.

Contents:

- Why is reproducibility important in science?
- Why should I make my work reproducible?
- What does reproducible analysis mean?
- How can I rethink my workflow to be reproducible?
- Which tools help me to perform reproducible analysis?

Learning objectives:

Participants will learn the following in this course:

- Conceptual and operational understanding of reproducibility
- Structuring workflows for individual or collaborative work
- Tools for reproducible workflow management and data collaboration
- Guidance towards structuring projects

Prior knowledge:

- Useful for participants who (plan to) use programming in their analytical work: Python, R, julia, etc.
- Some basic knowledge of version control (Git)

Requirements:

- Own PC, laptop
- Internet, web browser (up to date)

6.3.3 Operator Track 'Data Scientist'

6.3.2.1 6.3.3.1 Operator Track 'Data Scientist' Module:

Causal learning (10 hours)

Lecturer, affiliation: *Prof. Dr. Vanessa Didelez – Leibniz Institute for Prevention Research and Epidemiology – BIPS; University of Bremen, Faculty of Mathematics and Computer Science*

Background:

In many instances, data analyses ultimately strive to answer causal research questions. We may want to assess and quantify the potential effects of certain decisions, interventions or policies, e.g. will a sugar tax or more playgrounds reduce childhood obesity? Will participation in a special training programme for the unemployed increase their chances of finding employment? Is a national mammography screening programme actually helpful in preventing deaths from breast cancer? Such questions involve causal relations and go beyond mere prediction; in fact, methods that are optimised for prediction will give biased results for causal targets. Above all, we use non-experimental, i.e. observational, data to try and answer questions about causal relations. This calls for tailored methods relying on specific assumptions, and data analysts should have a basic awareness of them. Ultimately, causal learning from data is used, for instance, in decision-making by public health authorities or policy makers both in general and on an individual level.

Contents:

This course offers a basic introduction to the main concepts, fundamental assumptions and basic principles of key methods for causal learning.

We will cover the following:

- Terminology and central concepts such as potential outcomes, counterfactual prediction, causal diagrams, the target trial principle
- Basic methods for estimation, such as inverse-probability weighting, standardisation/g-formula, as well as checks, e.g. for balance and overlap
- Introduction to causal random forests and double machine learning of causal effects
- Basic algorithms for causal discovery, such as the PC and IDA algorithms

Learning objectives:

Participants will be able to

- employ potential outcomes notation and causal diagrams to represent prior causal knowledge, identify key structural assumptions as well as potential sources of bias (such as confounding or selection)
- apply methods for learning (or estimating) the causal effects of hypothetical interventions using the appropriate data analytic tools, and assess the plausibility of required assumptions
- apply basic algorithms for causal discovery (structure learning) to real-world data, while appreciating their strengths and weaknesses

Prior knowledge:

Basic knowledge of statistical data analysis and familiarity with R would be helpful.

Requirements:

- Brain, coffee
- Own PC, laptop
- For online format, a second screen might be beneficial

6.3.3.2 Operator Track 'Data Scientist' Module:

Computational social sciences (8 hours)

Lecturer, affiliation: *Nikolitsa Grigoropoulou, Ph.D., Postdoctoral Researcher, SOCIUM Research Center on Inequality and Social Policy, University of Bremen*

Background:

Computational social science is a cross-disciplinary field that applies computational statistics to large-scale data, typically from digital (or digitised) sources, in order to understand human behaviour. It involves a variety of observational, experimental, and simulation research formats and is popular among industries, government agencies, and academia alike. Quantitative text analytics, in particular, leverage the abundance of historical and real-time textual data produced by humans to generate insights, observe trends, and uncover dynamic patterns in ways that were previously unattainable with small-scale, qualitative analysis of texts or other forms of data. As such, text analytics are deployed far and wide to examine phenomena such as cultural change, population-level health indicators, disease outbreaks, discrimination in the legal system, and political behaviour, among others, and support domains such as customer service, product development, disaster response, decision-making, and policy-making. Thus, researchers across disciplines can benefit from engaging with these methods and becoming familiar with their intricacies and potential applications in academic research or other industries.

Contents:

- What is computational social science?
- An overview of text analytics and its workflow, including:
  - Theoretical and methodological foundations
  - Types of data sources, potentials and pitfalls
  - Data cleaning and data transformation
  - Methods of quantitative text analysis, with an emphasis on supervised and unsupervised text classification, topic modelling, and sentiment analysis
- Brief overview of text analysis software and statistical packages
- Application of quantitative text analysis in R and SAS Enterprise Miner

Learning objectives:

Participants will be able to

- define what computational social science (CSS) is and how it is related to other fields
- define text analytics and have a good grasp of their theoretical and methodological foundations
- understand the core workflow for text analysis
- perform basic data cleaning and data transformation
- identify different methods of quantitative text analysis, their fundamental characteristics, advantages, and limitations
- conduct basic text classification (and/or sentiment analysis)

Prior knowledge:

Fundamentals of statistics and quantitative research

Requirements:

Hardware:

- Own PC, laptop
- For online format, a second screen might be beneficial

Software:

- R (or RStudio)
- SAS Enterprise Miner (accessed for free through SAS OnDemand for Academics https://www.sas.com/en_us/software/on-demand-for-academics.html)
- Microsoft Office Excel

Other:

- Wi-Fi

6.3.3.3 Operator Track 'Data Steward' Module:

Reproducibility in science: Deep learning (10 hours)

Lecturers, affiliation: *Prof. Dr. Peter Maas, Dr. Daniel Otero Baguer*

University of Bremen, Faculty of Mathematics and Computer Science

Background:

This course is for anyone interested in deep learning and industry applications. It is also a good introduction to the field and, in particular, to the PyTorch deep learning library. In this course, participants will get hands-on and build their own neural networks, not only for typical computer vision tasks, but also for solving more complex problems such as obtaining computer tomography reconstructions.

Contents:

- Introduction to training neural networks with PyTorch
- Model-based classic approaches, e.g. ISTA
- Introduction to data-based methods, e.g. LISTA
- Neural networks for trivial ill-posed inverse problems and fully data-based methods
- Combining model and data-based methods: learned post-processing and learned gradient descent
- Deep Image Prior and mathematical aspects
- Applications involving Computed Tomography (CT)

Learning objectives:

Participants will have a good understanding of how neural networks work, and also the mathematical theory behind it. They will also be able to program deep learning approaches themselves using Python and the PyTorch library.

Prior knowledge:

Some experience in Python programming is needed.

Requirements:

- Own PC, laptop
- Google Colab (https://colab.research.google.com/)
- For online format, a second screen might be beneficial

6.3.3.4 Operator Track 'Data Steward' Module:

Evaluating machine learning and artificial intelligence algorithms (16 hours)

Lecturers, affiliation: *Prof. Dr. Werner Brannath – University of Bremen, Faculty of Mathematics and Computer Sciences*

Dr. Max Westphal – Fraunhofer Institute for Digital Medicine (MEVIS)

Background:

Artificial Intelligence (AI) and Machine Learning (ML) methods have been successfully applied in many areas, with some very prominent examples such as the AlphaGo algorithm. However, there are also examples where the real-world performance of ML/AI systems fell short of expectations (e.g. IBM Watson). When building an ML/AI algorithm, a large number of candidate models (e.g. for prediction) are explored with the goal of selecting what appears to be the best one. This can lead to severe overestimation of the prediction or classification capability (also called 'performance'). The algorithm and its apparent performance also depend on the data used for its development and validation. Weak performance of an ML/AI algorithm is often difficult to identify, but can result in severe harm when the algorithm is applied.

In view of this, we need to evaluate ML/AI algorithms and apply specific methods and techniques (e.g. based on cross-validation or bootstrapping) to avoid overestimation. Care must also be taken to distinguish between the performance of the model building process itself (unconditional performance) and the prediction or classification capability of the finally selected model (conditional performance). In this course we will illustrate the difficulties and challenges involved in the judgment of ML/AI performance, and introduce a number of techniques to arrive at a reliable estimation of unconditional and conditional performance. The content of this course is a requirement for anyone looking to build a reliable ML/AI algorithm, and helpful to those wanting to apply an existing one.

Contents:

- Motivation: why do we need (quantitative) method evaluation in ML/AI?
- Definition of performance measures of ML/AI solutions, primarily for supervised methods (classification, regression)
- Statistical inference for selected performance measures (estimation, statistical testing, confidence intervals)
- Important terminology and concepts (in-sample vs. out-of-sample performance, conditional vs. unconditional performance)
- Practical aspects (experimental design, study planning)
- Application of evaluation methods to case studies in R

Learning objectives:

Participants will gain an understanding of the challenges and difficulties involved in evaluating ML/AI algorithms, as well as an understanding of how to apply basic and more advanced methods for ML/AI algorithm evaluation.

Prior knowledge:

- Basic knowledge of statistics (e.g. estimation, confidence intervals)
- Basic machine learning skills (e.g. applying a ML algorithm to data)

Requirements:

- Own PC, laptop
- For online format, a second screen might be beneficial

6.3.3.5 Operator Track 'Data Steward' Module:

Hardware and system architectures for data science (1.5 hours)

Lecturers, affiliation: *Christopher Metz*

University of Bremen, Faculty of Mathematics and Computer Sciences

Background:

The increasing application of Machine Learning (ML) techniques in various areas has led designers to leverage ML accelerators like GPUs, TPUs and many more. However, choosing the most appropriate accelerator for such algorithms is very challenging as they typically need to adhere to tight constraints, e.g. low power consumption, performance, and low cost. As a consequence, selecting the right device for a given ML algorithm becomes a tough challenge.

Contents:

1. Which device (Machine Learning Accelerator) is best for a given ML algorithm?
2. Difference between GPUs, TPUs, CPUs, FPGAs, etc.

Learning objectives:

Participants will be able to select the right hardware for machine learning algorithms and understand the advantages and disadvantages of different hardware and system architectures.

Prior knowledge:

-

Requirements:

-

6.3.3.6 Operator Track 'Data Steward' Module:

Machine learning (12 hours)

Lecturers, affiliation: *Marvin N. Wright, Leibniz Institute for Prevention Research and Epidemiology – BIPS, Bremen*

Background:

Nowadays, machine learning is everywhere: old and new questions, problems and challenges are tackled with machine learning, sometimes with great success, sometimes not. To successfully use machine learning and understand its limitations, we have to go beyond buzzword bingo and learn the basic and general concepts of machine learning.

Contents:

This workshop teaches the major concepts of machine learning. We focus on general principles rather than doing deep-dives into individual methods. We will learn the difference between supervised and unsupervised learning, important notation such as models and learners, why training errors are different to test errors, and how to optimise and evaluate prediction performance. Nevertheless, this course covers the most important machine learning methods such as k-nearest neighbors, decision trees, random forests, boosting, support vector machines and artificial neural networks. The methods will be introduced in a non-technical and intuitive way. Theory sessions will be complemented by hands-on sessions in R, where the methods are applied in practice.

Learning objectives:

Participants will understand the basic concepts of machine learning:

- Supervised and unsupervised learning
- Difference between models and learners, training and test errors, etc.
- Over- and underfitting
- Hyperparameter tuning
- Performance evaluation

In addition, participants will learn about the major machine learning methods:

- K-nearest neighbors
- Decision trees
- Random forests
- Boosting
- Support vector machines
- Artificial neural networks

Participants will also be able to perform machine learning analyses in R:

- Model fitting
- Hyperparameter tuning
- Performance evaluation
- Benchmarking
- Visualising results

Prior knowledge:

Advanced maths is generally NOT required (only for a short section on support vector machines); basic programming skills are required.

Requirements:

- Own PC, laptop
- Internet, browser (up to date)
- For online format, a second screen might be beneficial

6.3.3.7 Operator Track 'Data Steward' Module:

Quantitative analyses for data science (12 hours)

Lecturers, affiliation: *Prof. Dr. Thorsten Dickhaus*

University of Bremen, Faculty of Mathematics and Computer Sciences

Background:

Proficiency in (mathematically grounded) quantitative data analysis is key to many modern applications in data science. Understanding the basic underlying principles helps to interpret data analysis results, even if data analysis is performed elsewhere.

Contents:

This course covers the basic notions of mathematical and applied statistics. It also presents some prototypical statistical models, in particular regression models and time series models, in more detail. Major topics include point estimation, confidence estimation, testing, and prediction. At times, connections to statistical (machine) learning are also presented. This course is a mixture of lectures and practical hands-on sessions.

Learning objectives:

Participants will learn the following:

- Principles of decision-making amid uncertainty
- Statistical data modelling
- Statistical data analysis
- Interpretation of statistical data analysis results

Prior knowledge:

- Basic mathematical education (maps, matrices, taking derivatives, solving integrals, matrix-vector multiplication, etc.)
- Knowledge of basic probability theory (probability spaces, random variables, random vectors, probability distributions, central limit theorem, etc.)

Requirements:

- Own PC, laptop with R software installed (R + R-Studio)
- For online format, a second screen might be beneficial
- Paper and pens

6.3.3.8 Operator Track 'Data Steward' Module:

Visual analytics with Geographical Information Systems (GIS) (16.5 hours)

Lecturers, affiliation: *Dr. Antonie Haas, Alfred Wegener Institute for Polar and Marine Research*

Background:

Geographical Information Systems (GIS) are powerful when it comes to managing, analysing and visualising georeferenced data. GIS software has the capability to be applied to data in a broad variety of scientific disciplines due to its location-based data processing. Every data table or raster data file that has coordinate information can be imported and processed. The multi-layer concept of GIS software allows data to be embedded in base maps or any other environmental data that support spatial analysis and visualisation of the data.

Contents:

This GIS introduction course includes the basics of GIS, import of data (GIS and non-GIS formats), basics of coordinate systems, handling and editing of data tables, querying of data tables and creation of selections, establishment of GIS projects, and the creation and design of highly informative figures and maps.

Learning objectives:

Participants will learn how about GIS software and how to use it for data analysis and visualisation so they can apply it to their own data.

Prior knowledge:

-

Requirements:

- Own PC, laptop

6.4 Evaluation of tutorials/workshops, etc.

Please state the extent to which you agree with the following statements

[Strongly disagree (1) / Disagree (2) / Neither disagree nor agree (3) / Agree (4) / Strongly agree (5)].

___ The learning objectives of the workshop were clearly stated.

___ I understood what the workshop was about.

___ The workshop content was easy to follow.

___ Each topic was given enough time to be covered properly.

___ The workshop was a good mix of theory and practice.

___ I got a good insight into the topic of research data management.

___ I learned a lot of new and interesting things.

___ The contents of the workshop are helpful for my scientific work.

___ All in all, the workshop met my expectations.

Did your workshop have any practical exercises? ◯ Yes ◯ No

If so, Please state the extent to which you agree with the following statements

[Strongly disagree (1) / Disagree (2) / Neither disagree nor agree (3) / Agree (4) / Strongly agree (5)].

___ The instruction of the practical exercises was clear.

___ The content of the practical exercises was interesting.

___ The practical exercises were of practical relevance.

On a scale of 1 (very good) to 5 (very bad), how would you rate the workshop overall? ____

On a scale of 1 (very likely) to 5 (very unlikely), how likely are you to recommend this workshop to a colleague? ____

What did you like the most? _____

Please name one thing we could improve. _____

The workshop was held online. Would you participate in an online workshop format like this again?

◯ definitely participate ◯ rather participate ◯ rather not participate ◯ definitely not participate ◯ not specified

Optional questions on NFDI4Health

Have you heard of NFDI4Health? ◯ Yes ◯ No

If so, who/where from? _____