de.KCD: German Competence Center Cloud-Technologies for Data Management and Processing

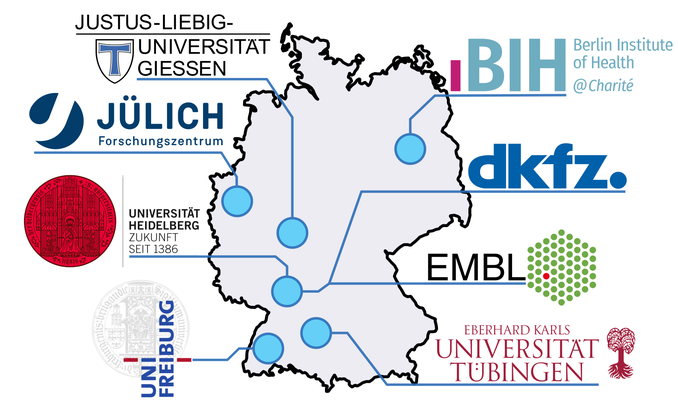
Sebastian Jünemann[1], The de.KCD Consotrium[2], Alexander Sczyrba[1]

[1] Institute of Bio- and Geosciences IBG-5, Research Center Jülich GmbH, c/o Centrum für Biotechnologie (CeBiTec), Bielefeld University, 33594 Bielefeld, Germany

[2] The de.NBI Cloud Consortium: A. Goesmann, Justus-Liebig-University Giessen; R. Eils, BIH-Zentrum Digitale Gesundheit, Charité - Universitätsmedizin Berlin; P. Bork, European Molecular Biology Laboratory Heidelberg; O. Kohlbacher, Eberhard Karls University Tübingen; U. Kummer, Heidelberg University; R. Backofen, Albert-Ludwigs University Freiburg; I. Buchhalter, Deutsches Krebsforschungszentrum Heidelberg; A. Sczyrba, Forschungszentrum Jülich GmbH

The de.KCD Consortium

The de.KCD addresses the challenges of the current phase of digital transformation by bundling expertise in cloud computing and data management, building upon the network, hardware capacities and services of the de.NBI Cloud. The project implements suitable measures for cloud-based data management, standardized data analysis and generic training content for knowledge transfer across different specialist areas. Beyond these, de.KCD also aims to promote collaboration and knowledge exchange between research locations by creating a networked, collaborative data space for national and international research projects.



Cloud Based Data Management

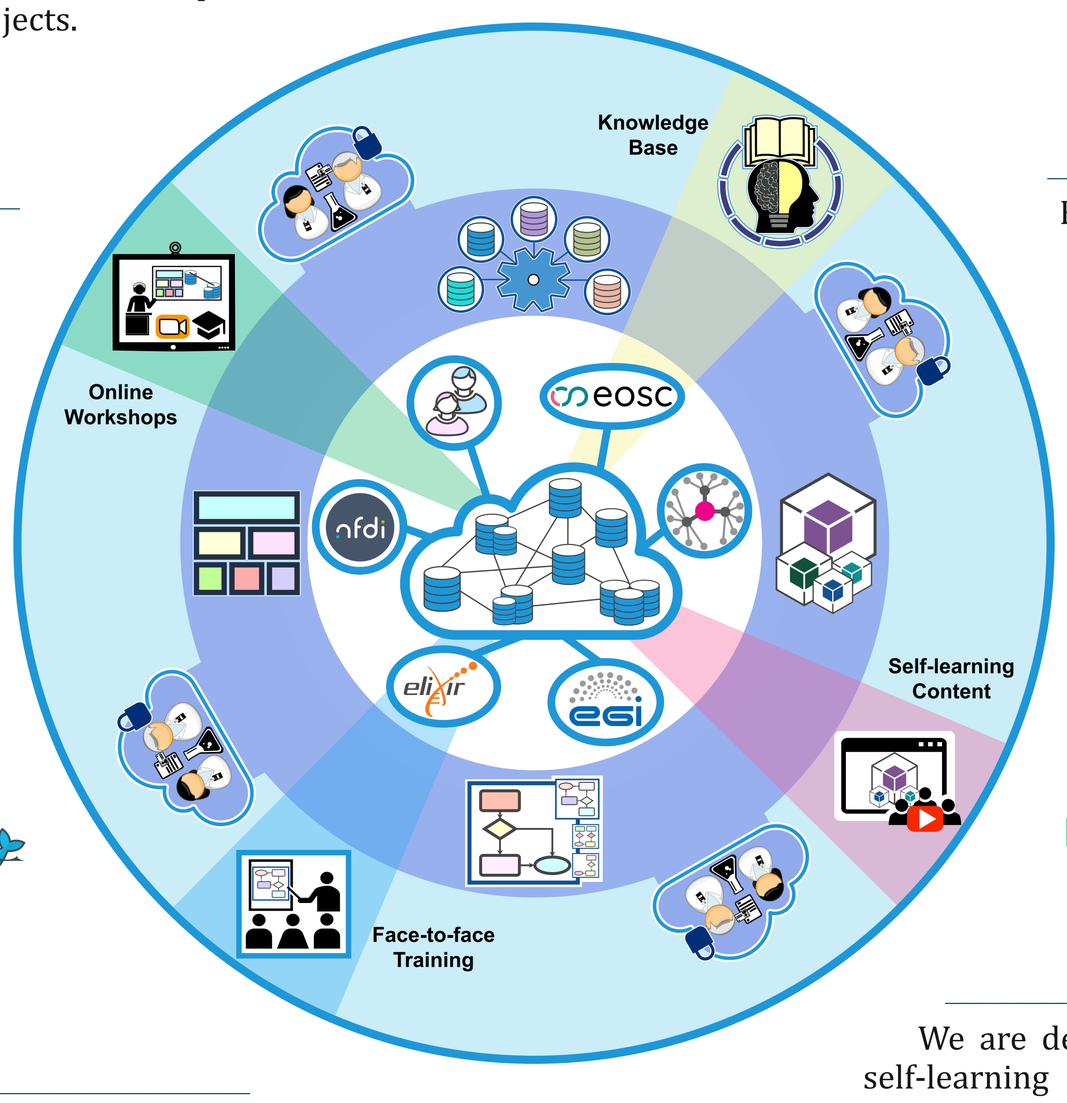
For data science methods, we focus on the use of cloud-based infrastructures for distributed and scalable data management, as well as the necessary skills for standardized and automated data processing. In accordance with the FAIR principles, this includes specific expertise for the reproducible handling of data and the use of appropriate software tools, for example, through the use of software container solutions (e.g. BioContainers) in conjunction with cloud-based data management systems (e.g. distributed databases). This will complement the efforts by resource providers and infrastructure initiatives within Europe (ELIXIR, EOSC) and Germany, specifically the different NFDI consortia.

nfdi 4 health · Data PLANT · NFDI4 MICROBIOTA · NFDI4 BIODIVERSITY · FAIRagro · NFDI Neuroscience

Software Stacks

Flexible virtualization and the use of different and specialized software solutions (e.g. GPU-based algorithms) for cloud-based data analysis often requires the use of dedicated cloud-based software stacks. To this end, scientists and data analysts from all specialist areas are actively supported (e.g. via a dedicated ticket system, chats, knowledge base) in the development and establishment of tailored software stacks, which are periodically renewed and automatically tested in order to automate and scale the underlying data management and the management of the virtualized compute

openstack. · SimpleVM · ANSIBLE



Workflows

For scalable and automated processing of research data, we teach skills in workflows (e.g. Nextflow, Snakemake, or Galaxy) and grid computing (e.g. slurm and BiBiGrid) that ensure easier traceability (data provenance) of the collection, generation, processing and reproducibility of research data (e.g. as machine-readable results in the form of Research Data Objects). Knowledge of the evaluation and use of version control systems for the audit-proof storage of e.g. differently parameterized workflows or individual evaluation scripts completes the data competence transfer in this area.

nextflow · useGalaxy.eu · slurm workload manager

Trusted Research Environments

Pre-configured virtual research environments (VRE) based on de.NBI custom SimpleVM environment provide easily accessible integrated development environments, data science notebooks and virtual desktops. The VREs can be accessed by browser or secure shell (ssh). VRE's and custom virtual machine's (VM) images can also be used to provide uniform VMs for trainings and workshops. Requirements for trusted research environments (TREs) will be described and material for the automated setup of these in cloud environments will be designed and tested with the aim of making them available at certified cloud locations. Internal audits and external ISMS certification (ISO27001 or BSI C5) ensure secure operations of the data centers.

Integrated Development Environments	Data Science Notebooks	Virtual Desktop Environments
THEIA · Visual Studio Code	jupyter · RStudio	XFCE

Training

We are developing a structured training program and self-learning units consisting of learning paths and modules, complemented by a scalable, cloud-based training infrastructure with pre-configured learning environments. As a whole, this forms the knowledge and competence base of de.KCD and represents a community-driven framework for the collection of FAIR training materials for software developers, system administrators and scientists. We place particular emphasis on train-the-trainer lessons so that the generic materials can be transferred to other subject-specific domains and used there, in an adapted form, for corresponding training courses. Material will be offered in the form of online, hybrid and face-to-face courses throughout the year, complemented by annual summer schools, for researchers at all career levels, which are announced in our portal with an integrated training calendar.

Workshop

Contact:

info@datenkompetenz.cloud

https://www.datenkompetenz.cloud

SCAN ME · SCAN ME

Funded by the European Union NextGenerationEU

SPONSORED BY THE Federal Ministry of Education and Research

de.NBI GERMAN NETWORK FOR BIOINFORMATICS INFRASTRUCTURE · JÜLICH Forschungszentrum · elixir