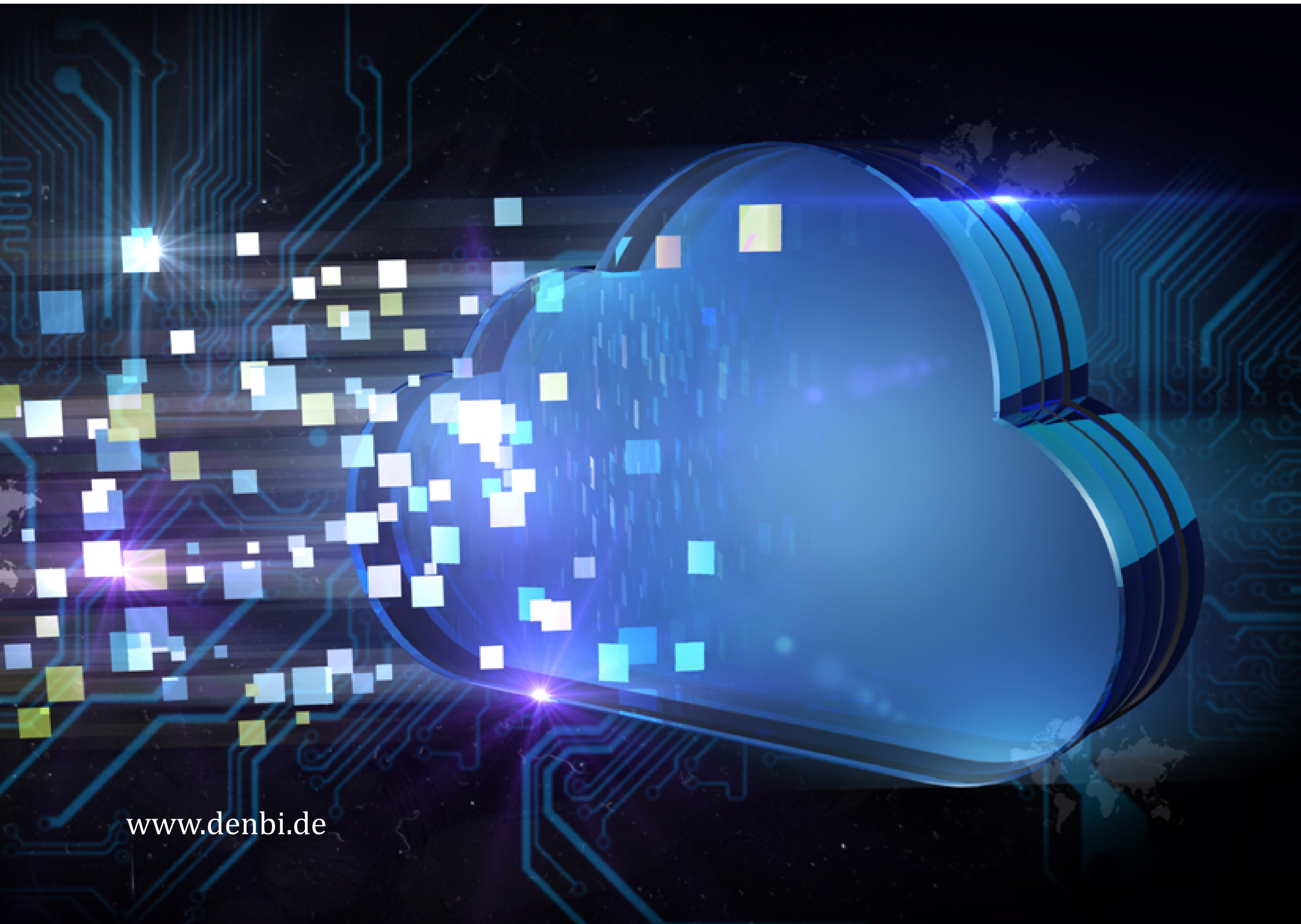
GERMAN NETWORK FOR BIOINFORMATICS INFRASTRUCTURE

DEVELOPMENT AND OPERATION OF THE FEDERATED de.NBI CLOUD

Contributions of the German Network for Bioinformatics Infrastructure



www.denbi.de

FOREWORD

Dear Reader,

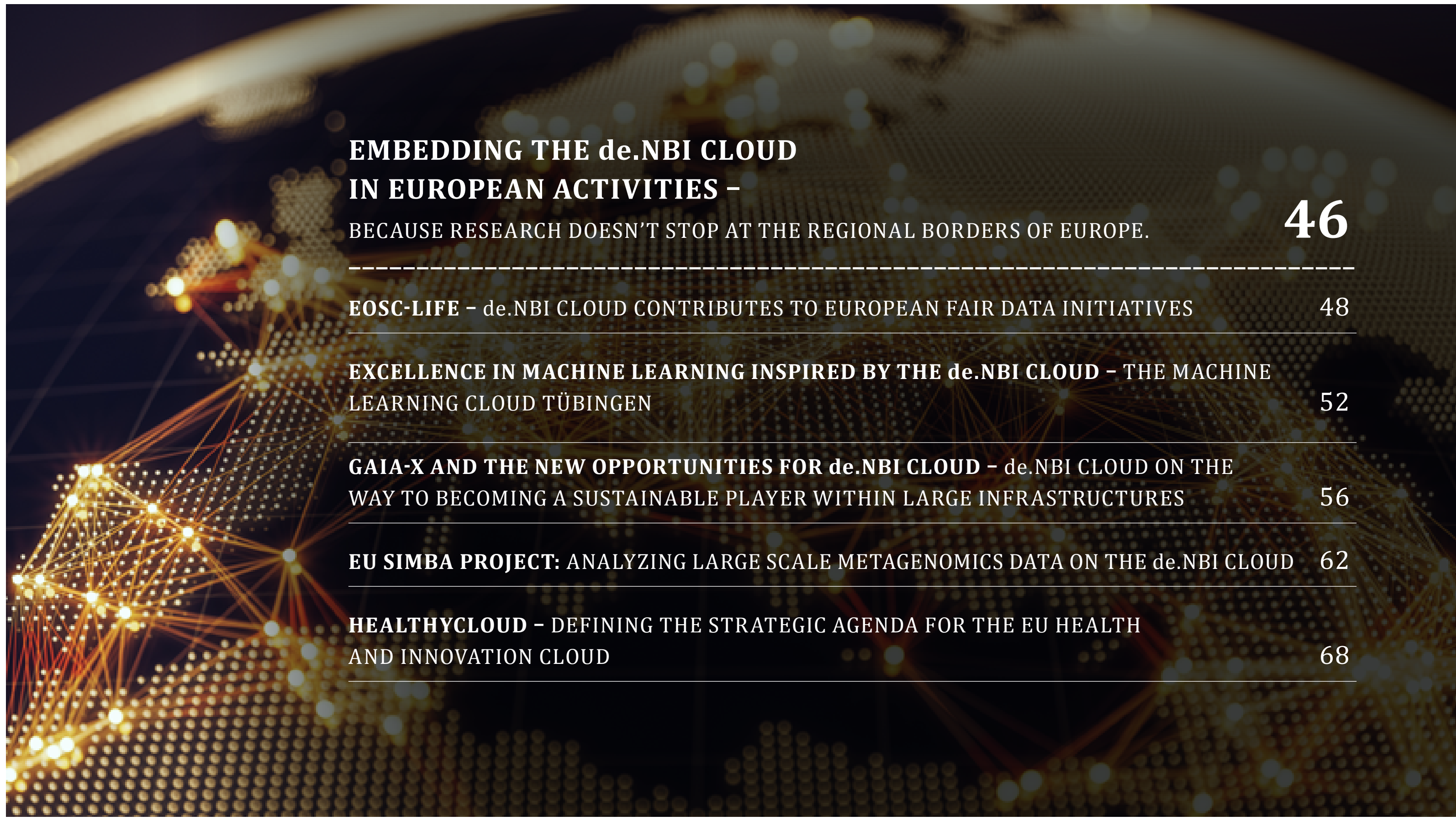
Access to research data has never been easier than with cloud solutions. Soon, the cloud will take a key focus in the life sciences, given its benefits of improved accessibility, scalability and flexibility. Following this trend, the de.NBI Cloud was initiated at an early stage after the German Network for Bioinformatics Infrastructure (de.NBI) was launched in 2015, with the aim to provide life scientists in Germany with access to compute and storage capacities to be able to analyze their own research data adequately. In the meantime, seven years later, the de.NBI Cloud is the largest German academic cloud for life science purposes with more than 1800 registered users. It is also the scientific and collaborative backbone for new major German initiatives like NFDI, DeCOI, and GHGA or EOSC-Life in the European sector of computational biosciences.

This booklet showcases the power, the strength and the range of the de.NBI Cloud for the life science community by providing the ability to unlock data and better collaborate across the ecosystem, which in turn enables us to innovate and scale in this way. Beginning with the information on establishment and operation of the distributed de.NBI Cloud this brochure also sheds light on the role of the de.NBI Cloud in national and international projects. Furthermore, it gives a wide range of application examples of how the scalability of cloud projects on existing hardware infrastructures can tailor specific requirements for individual projects.

I hope you find stimulating ideas and useful methods here. I wish you lots of pleasure reading our de.NBI Cloud brochure and trust that you find it useful.

Alexander Sczyrba
de.NBI Cloud spokesperson,
Bielefeld University

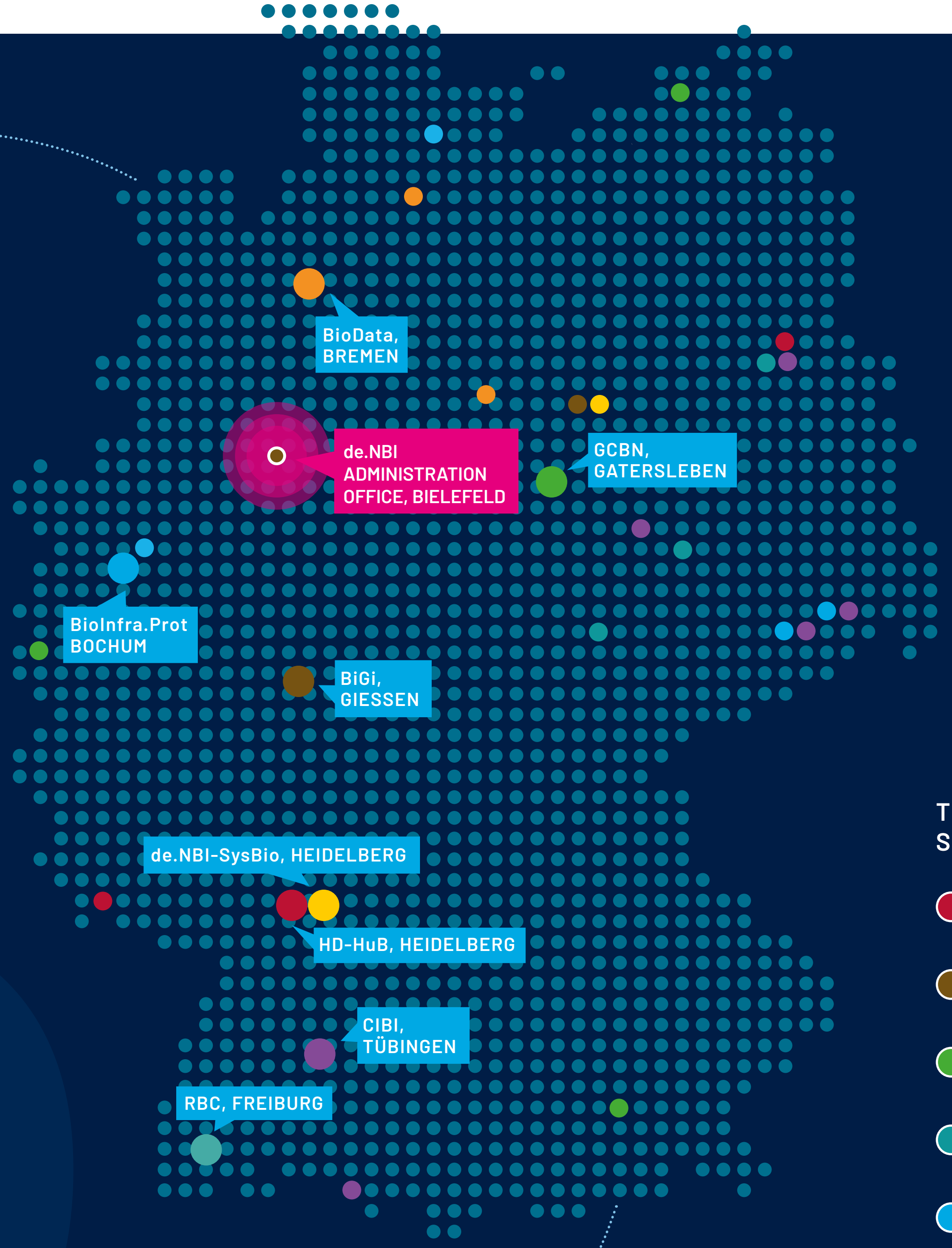
CONTENT



de.NBI
GERMAN NETWORK FOR BIOINFORMATICS INFRASTRUCTURE

THE GERMAN NETWORK FOR BIOINFORMATICS INFRASTRUCTURE – de.NBI

de.NBI is a distributed bioinformatics infrastructure which started in March 2015 as an academic and non-profit initiative of the German Ministry of Research and Education (BMBF). The de.NBI network aims to provide high standards of bioinformatics services, comprehensive training, powerful computing capacities (de.NBI Cloud) as well as connections to industrial companies that assist researchers to more effectively exploit their own data. In addition, a de.NBI Cloud resource has been established at six locations in Germany that can be used free of charge for the analysis of life science datasets. On international level, the network is integrated into the ELIXIR consortium via the German node. Since 2022 the German government provided the Jülich Research Centre, member of the Helmholtz Association, with funding to continue the de.NBI network with the primary goal to contribute to the advancement of life science research in Germany and Europe.



THEMATIC FOCUSES & SERVICE CENTERS:

● **HUMAN BIOINFORMATICS**
HEIDELBERG CENTER FOR HUMAN BIOINFORMATICS (HD-HuB)

● **MICROBIAL BIOINFORMATICS**
BIELEFELD-GIESSEN RESOURCE CENTER FOR MICROBIAL BIOINFORMATICS (BiGi)

● **PLANT BIOINFORMATICS**
GERMAN CROP BIOGREENFORMATICS NETWORK (GCBN)

● **RNA BIOINFORMATICS**
RNA BIOINFORMATICS CENTER (RBC)

● **PROTEOME BIOINFORMATICS**
BIOINFORMATICS FOR PROTEOMICS (BioInfra.Prot)

● **INTEGRATIVE BIOINFORMATICS**
CENTER FOR INTEGRATIVE BIOINFORMATICS (CIBI)

● **BIODATABASES**
CENTER FOR BIOLOGICAL DATA (BioData)

● **DATA MANAGEMENT/SYSTEMS BIOLOGY**
de.NBI SYSTEMS BIOLOGY SERVICE CENTER (de.NBI-SysBio)

● **LOCATIONS OF SERVICE CENTERS**

● **LOCATIONS OF PARTNERS**

SERVICE

- TOOLS
- WORKFLOWS
- DATABASES
- CONSULTING

TRAINING

- TRAINING COURSES
- SUMMER SCHOOLS
- HACKATHONS
- WEBINARS

de.NBI CLOUD

- INFRASTRUCTURE AS A SERVICE
- PLATFORM AS A SERVICE
- SOFTWARE AS A SERVICE

INDUSTRIAL FORUM

- INDUSTRY SERVICES
- CONSULTING
- NETWORKING

USER MEETINGS

- CONTINUOUS DEVELOPMENT OF TOOLS
- CONSULTING
- EXCHANGE OF OPINIONS AND EXPECTATIONS
- BOTTOM-UP FEEDBACK

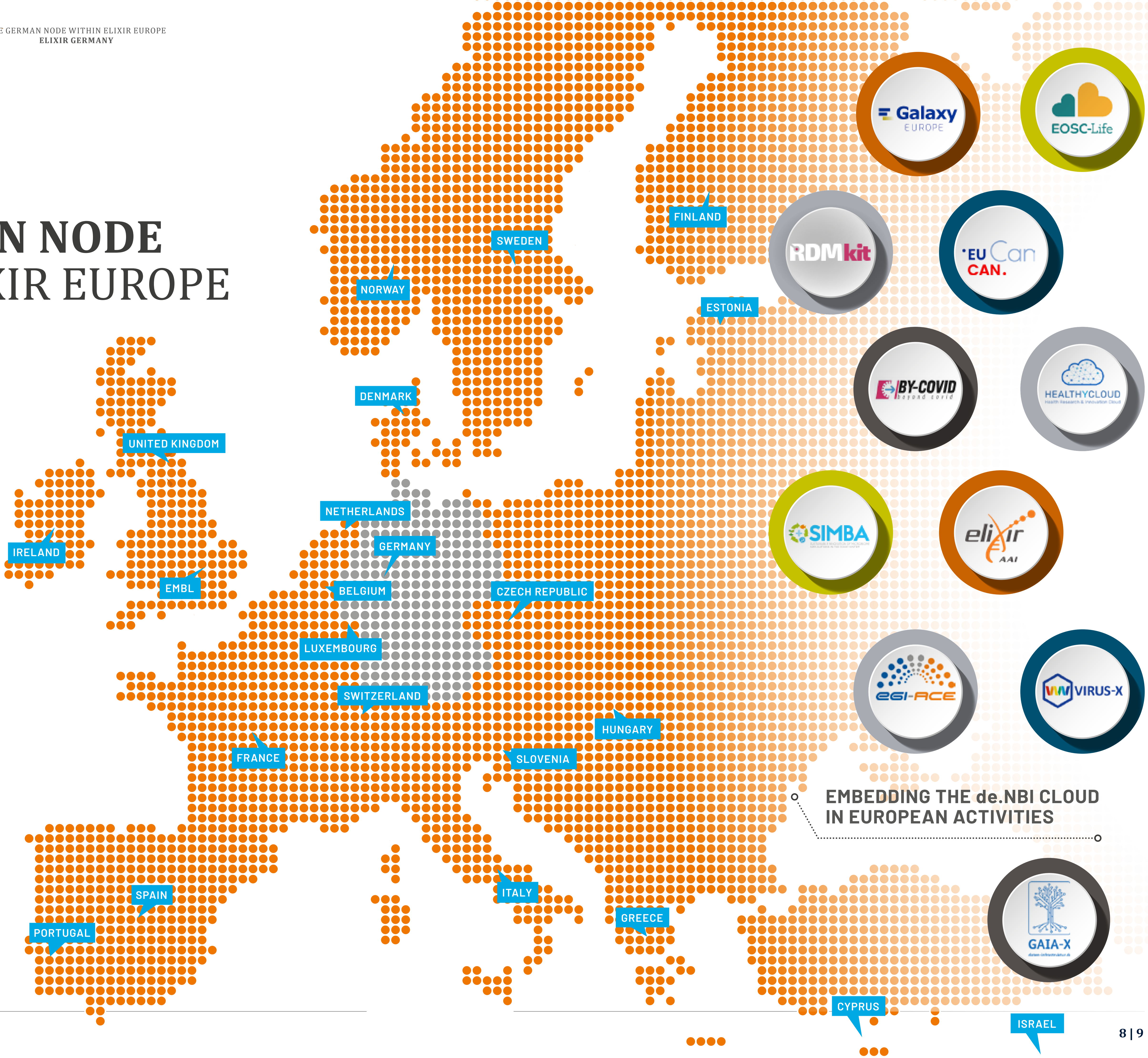
CONTACT

www. denbi.de
@denbiOffice
linkedin.com/company/de-nbi

elixir GERMANY

THE GERMAN NODE WITHIN ELIXIR EUROPE

ELIXIR activities are structured around Platforms, Communities, and Focus Groups. de.NBI Cloud sites contribute to the central tasks of the ELIXIR Compute Platform e.g., via Life Science Login, ELIXIR Hybrid Cloud and Data Integration for Compute Resources, a project which aims to coordinate the transfer of large amounts of data across ELIXIR clouds. Furthermore, when setting up the de.NBI Cloud, various efforts were made to create synergies and establish connections to the ELIXIR cloud infrastructure. To meet the needs of big data consortia, the de.NBI Cloud will be the scientific and collaborative backbone for new major initiatives in computational biosciences in Germany and Europe.



EMBEDDING THE de.NBI CLOUD IN EUROPEAN ACTIVITIES

CONTACT

Andreas Tauch
Head of Node
tauch@cebitec.uni-bielefeld.de
www.denbi.de/elixir-de

GOING BEYOND THE GRID TO ENABLE LIFE SCIENCE DATA ANALYSIS – **DEVELOPMENT AND OPERATION OF THE DISTRIBUTED de.NBI CLOUD**

With the life sciences becoming increasingly data-driven, cloud technologies will become more important to the sector. An appropriate IT infrastructure is crucial to perform analyses with such large datasets and to ensure secure data access and storage. The de.NBI Cloud is an excellent solution to enable integrative analyses and the efficient use of data in research and application.

de.NBI CLOUD FEDERATION

Right from the start in 2017 the de.NBI Cloud searched for a way to simplify the access and handling of computational resources in the cloud federation. With the mission to bring compute power to the life sciences in Germany, users with different background knowledge not only in cloud computing but also in command line usage were expected. In order to satisfy these varying requirements, the de.NBI Cloud created early in its history appropriate access mechanisms and innovative software solutions towards a fully federated cloud with project types tailored to the life science community.

A BRIEF HISTORY OF THE de.NBI CLOUD

The need to provide substantial compute resources to the life sciences community in the framework of de.NBI, preferably in form of a cloud infrastructure, was substantiated by the de.NBI Special Interest Group SIG 4 - Interoperability and Data Management at an early stage after the start of the de.NBI network in March 2015, as the initial funding of the project comprised only few hardware resources to meet the demand of the network itself and of its user community.

A primary impulse for the establishment of a de.NBI Cloud infrastructure came from the de.NBI reviewer panel, which met in Berlin in March 2016 and pointed out the need for a cloud infrastructure for the German life sciences in a letter to the Federal Ministry of Education and Research (BMBF). Subsequently, at a special meeting of the de.NBI Central Coordination Unit (CCU) in Braunschweig in July 2016, the de.NBI consortium was informed about an extraordinary BMBF funding for the establishment of the de.NBI Cloud of 5 million euros.

The first cloud computing infrastructures were set up at five universities in Bielefeld, Giessen, Heidelberg, Freiburg and Tübingen during winter 2017. Technical staff was hired at these locations to guarantee the establishment and continued operation of the de.NBI Cloud. This set the course for the successful establishment of the distributed de.NBI Cloud federation, which adds an important third pillar to the de.NBI network alongside services and training.

As the next major step in the development of the de.NBI Cloud, an extension by a storage component, was suggested by the newly founded Special Interest Group SIG 6 - de.NBI Cloud in summer of 2017. This storage component is of great importance for the efficient operation of the de.NBI Cloud and additional funding was finally granted by the BMBF. From this point on, the de.NBI Cloud could advance from the set-up phase to the production phase, i.e. all computing services requested by users could also be served by the de.NBI Cloud. The de.NBI Cloud thus reached another milestone and has been available as a bioinformatics infrastructure in Germany to all users from the life sciences since then.

Continued discussions about the further development of the de.NBI Cloud soon revealed additional requirements in the areas of certification and security. Additional funding by the BMBF was acquired in 2018 to initiate a process towards information security certification and to generally strengthen security-related aspects. Subsequently, from 2019 onwards, the certification and security management track was pursued which led to the first successful certification of three de.NBI Cloud locations in 2021.

In the course of time the de.NBI Cloud was subject to a structural change. A sixth cloud site was established at the Berlin Institute of Health and integrated into the de.NBI Cloud federation. In addition, the EMBL in Heidelberg joined the de.NBI Cloud as an associated partner in March 2020.

The COVID-19 pandemic proved to be both a great challenge for society and the de.NBI network, as well as providing an opportunity to demonstrate the value of the de.NBI Cloud for the public. de.NBI partners have been and still are involved in more than 30 projects that contribute to solving current SARS-CoV-2 research issues. A large number of these projects use the de.NBI Cloud to tackle computational problems with direct relation to COVID-19. Here it is of particular importance that all researchers from the life sciences have access to the de.NBI Cloud free of charge. de.NBI Cloud resources for COVID-19 related projects were fast-tracked via a dedicated submission process to provide them without further delay to applicants .

Currently, the de.NBI Cloud is heavily involved in the realization of National Research Data Infrastructure Germany (NFDI) projects in the field of life sciences. The already approved NFDI projects GHGA, NFDI4Biodiversity, DataPlant and NFDI4Microbiota make use of de.NBI Cloud resources for their implementations.
The de.NBI Cloud continues to serve as a prime example of a highly successful dis-

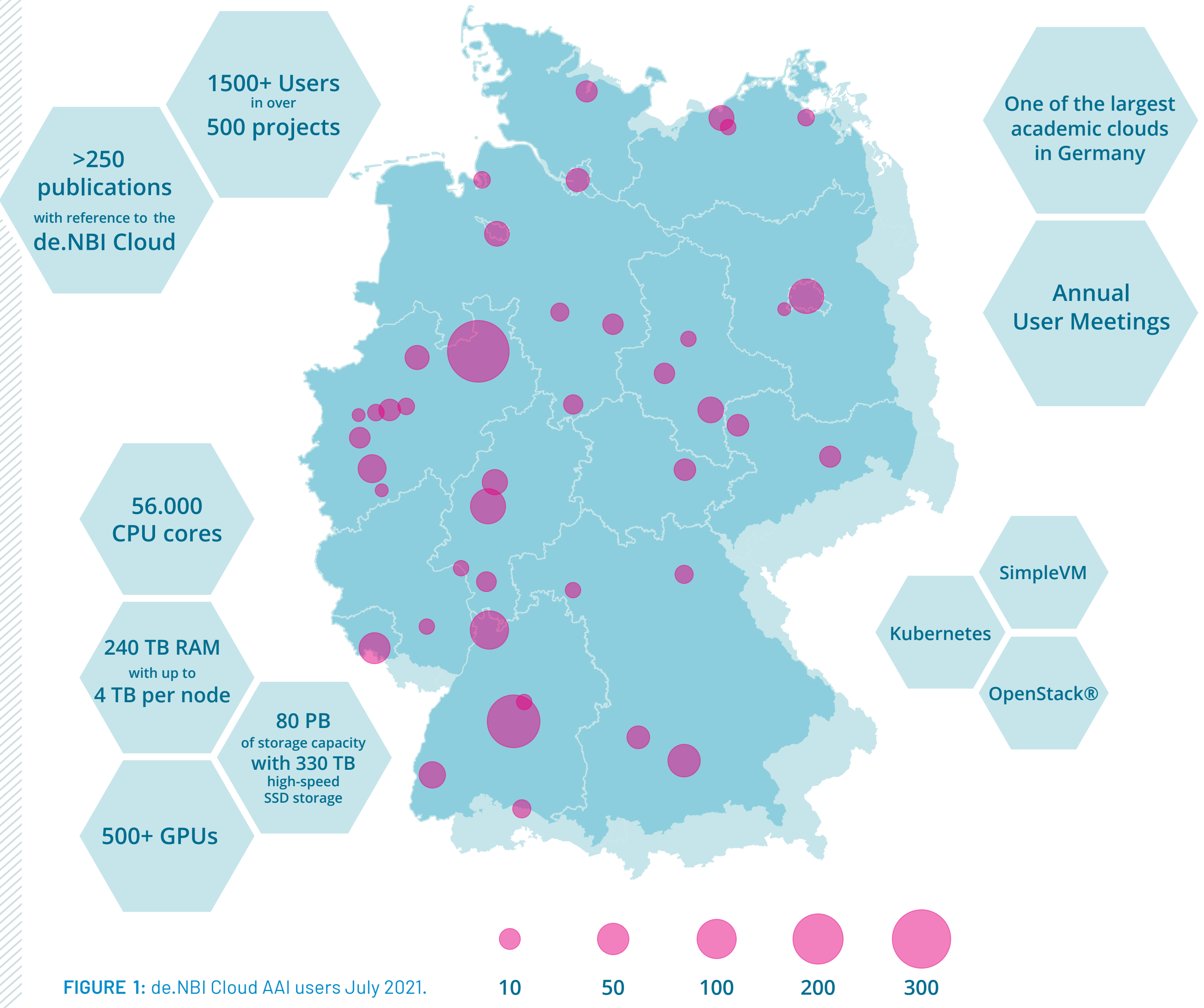


FIGURE 1: de.NBI Cloud AAI users July 2021.

tributed computational infrastructure for the life sciences in Germany and beyond.

de.NBI CLOUD FEDERATION POWERED BY LIFE SCIENCE LOGIN

The de.NBI Cloud offers computational resources for the life sciences community in Germany and applications were submitted from users in more than 40 different universities and research institutions distributed all over Germany (Figure 1).

This positive development also involves challenges, especially with respect to the federated setup of the de.NBI Cloud. An early decision in the de.NBI Cloud development was to integrate with the ELIXIR authentication and authorisation infrastructure, short Life Science Login [1].

All cloud sites and the de.NBI Cloud portal integrate with Life Science Login, which allows to handle user permissions on all cloud sites. de.NBI Cloud users benefit from this integration by a unified login procedure on all cloud sites which eliminates the need for additional accounts on every cloud site. Users are able to access all de.NBI Cloud services via their home institutional account which can be connected to their Life Science Login identity. User permissions at all cloud sites are centrally managed in Life Science Login only by the de.NBI Cloud portal which makes it the central access and management component of the cloud federation. Besides the simplified access for users, Life Science Login helps de.NBI Cloud employees to identify user affiliations and simplifies the review process.

de.NBI CLOUD PROJECT TYPES

Project applicants have the choice between the two project modes OpenStack and SimpleVM. OpenStack and SimpleVM target different use cases and users regarding their prior knowledge in cloud computing.

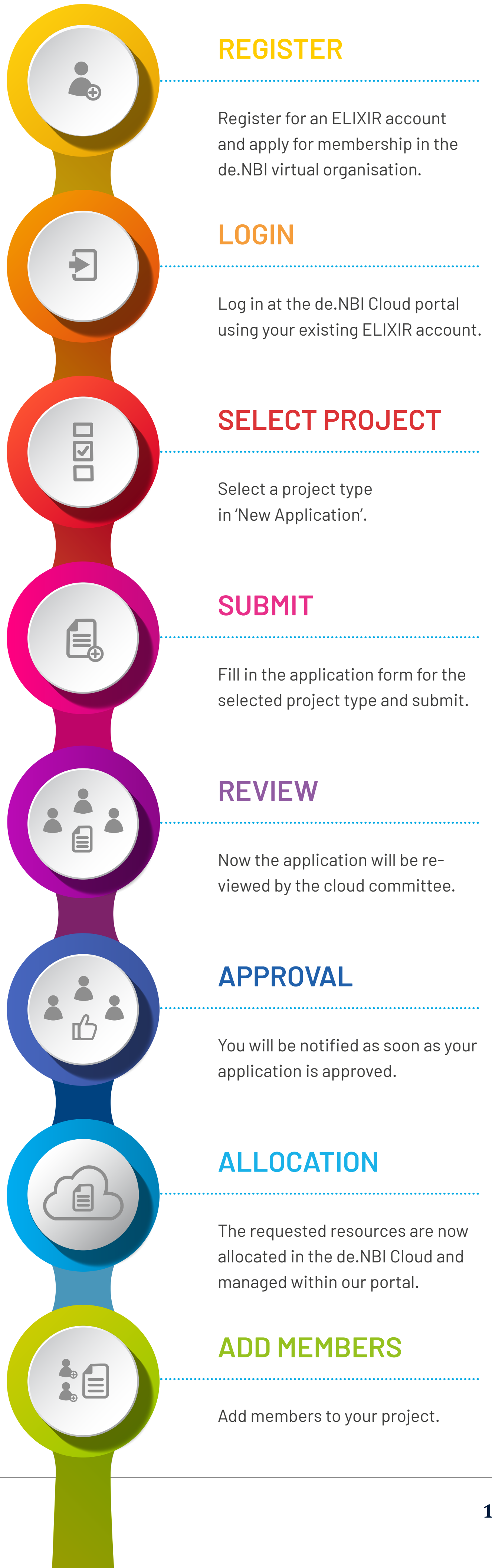
OpenStack as an 'Infrastructure as a Service' system allows high configurability of any resource type available, such as virtual machines, network, block and object storage.

OpenStack eases the scaling of virtual machines and distribution of data for efficient computations. Any interaction with OpenStack can be automated via its API, e.g. starting and stopping of virtual machines, which allows it to manage any data analysis or orchestration framework of the cloud computing ecosystem.

For users who just want to start a virtual machine or a cluster of machines without the need to configure network properties, the de.NBI Cloud developed its own project type, called SimpleVM. SimpleVM is an abstraction layer on top of OpenStack that allows the management of single machines or even clusters of machines with just a few clicks. In addition, SimpleVM allows the deployment of well-known research and development web-based environments such as Rstudio, Guacamole Remote Desktop and Theia IDE.

de.NBI CLOUD PORTAL: A CLOUD FEDERATION MANAGEMENT TOOL

de.NBI Cloud sites are independent OpenStack installations that are connected to the de.NBI Cloud portal. The de.NBI Cloud portal allows the management of applications and projects. As already mentioned, the de.NBI Cloud portal integrates with Life Science Login and is able to manage users project memberships that are automatically synchronized with all cloud sites that are connected to the cloud federation. Different user roles that have been defined from the start of the de.NBI Cloud are mapped to the portal functionality in order to allow the application procedure depicted in figure 2. Besides the user role that allows every Life Science Login user to apply for projects, the cloud access committee role is able to review incoming applications and upon approval to delegate the requested resources to one of the cloud sites. Cloud site administrators always have the possibility to view all projects that belong to their cloud site and get



REGISTER

Register for an ELIXIR account and apply for membership in the de.NBI virtual organisation.

LOGIN

Log in at the de.NBI Cloud portal using your existing ELIXIR account.

SELECT PROJECT

Select a project type in 'New Application'.

SUBMIT

Fill in the application form for the selected project type and submit.

REVIEW

Now the application will be reviewed by the cloud committee.

APPROVAL

You will be notified as soon as your application is approved.

ALLOCATION

The requested resources are now allocated in the de.NBI Cloud and managed within our portal.

ADD MEMBERS

Add members to your project.

FIGURE 2: Cloud Access.

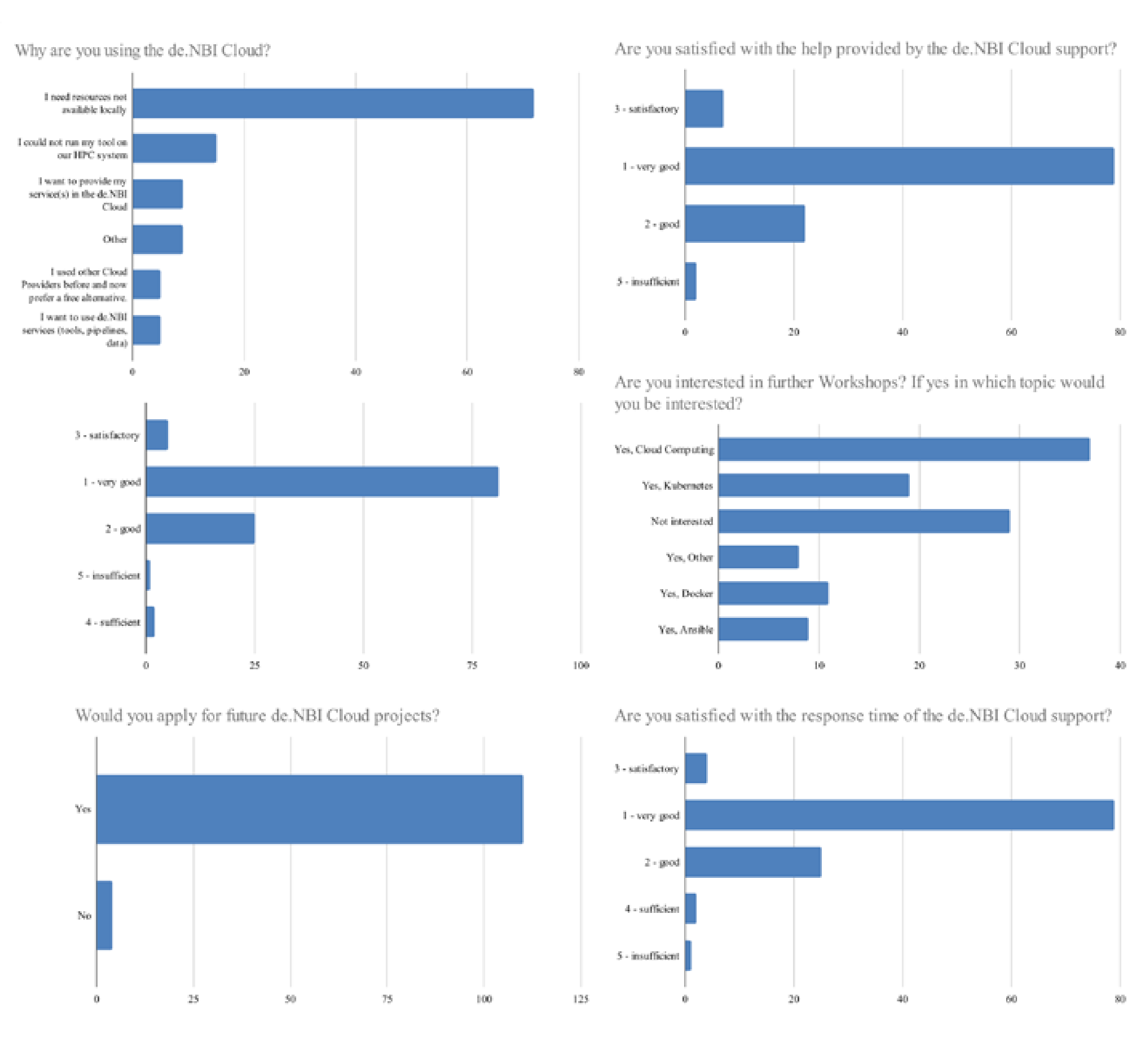


FIGURE 3: Cloud user survey results (2021).

notified if a new project comes in for a last parameter check before the actual resource allocation takes place. In addition, the de.NBI Cloud portal eases everyday tasks in administration like sending announcements to all users, specific project members or members of one cloud site.

Upon application approval the applicant gets notified about the cloud committee's decision. Afterwards the applicants have access to the project overview page that allows them to add new members or request modifications and lifetime extensions. The newest de.NBI Cloud portal feature is the de.NBI Cloud credits system that provides the ability to account for the consumption of computational resources at all cloud sites. The more cores and the higher the amount of RAM that is used as part of the machine, the more credits will be consumed per hour. More valuable hardware like GPUs and high memory nodes can have a higher weighting factor. This federation-wide accounting system increases awareness of resource consumption and reduces thereby the number of idling virtual machines.

de.NBI CLOUD USERS AND USER MEETINGS

de.NBI Cloud users have different use cases due to different research areas (Figure 3) and different background knowledge in cloud computing and command line usage. For this reason the de.NBI Cloud staff organize annual user meetings for training purposes and community building. Experienced de.NBI Cloud users present their projects and demonstrate the manifold use of the de.NBI Cloud. Workshops are organized to teach newcomers and advanced users in state-of-the-art cloud computing technologies. User meetings are also used to collect valuable feedback from the community and give users the chance to talk to experts in cloud computing for learning how to approach their research challenges in a de.NBI Cloud way.

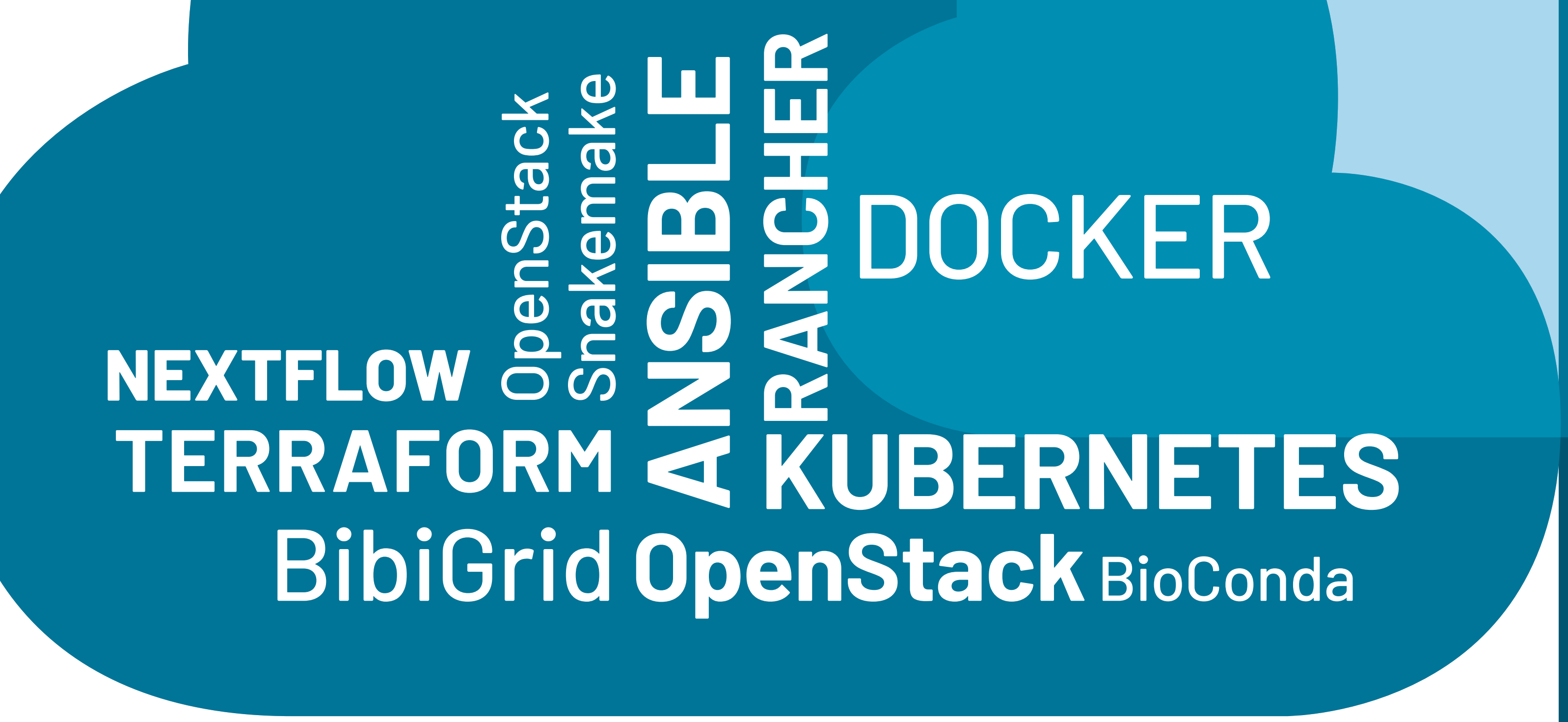
CONCLUSION & OUTLOOK

The de.NBI Cloud portal allows to manage multiple independent cloud sites without increasing the complexity of project and resource management and without compromising the user experience which seems to have worked out according the a recent survey (Figure 3). As a result of the current user requests additional project types such as Kubernetes will be added. Future user meetings will be held for teaching and community building. The feedback of the community will help to shape the future of the de.NBI Cloud.

REFERENCE: [1] F1000Research 2018; 7(ELIXIR):1199. DOI: https://doi.org/10.12688/f1000research.15161.1

AUTHOR: Peter Belmann[1]

[1] Institute of Bio- and Geosciences (IBG-5) - Computational Metagenomics, Forschungszentrum Jülich GmbH, 52425 Jülich, Germany



ANNUAL de.NBI CLOUD **USER MEETINGS**

The de.NBI Cloud provides computing and storage capacity for the life sciences and makes important de.NBI services accessible via the cloud. In addition to that, the de.NBI Cloud provides the opportunity for current users to learn more about best practices in cloud computing and introduces newcomers and potential future users to the capabilities provided by the de.NBI Cloud that could help accelerate their research in future projects. In our annual user meetings we would like to highlight topics of interest for our growing community. Since bioinformatics and the technologies in the cloud computing environment are constantly changing, it is essential to constantly realign the services offered by the de.NBI Cloud. In the user meetings, suggestions and tips for improved de.NBI Cloud operation are collected and then implemented if applicable.

CLOUD USER MEETING BIELEFELD
2018
47 PARTICIPANTS

CLOUD USER MEETING^{2} HEIDELBERG
2019
60 PARTICIPANTS

CLOUD USER MEETING^{3} ONLINE
2020
75 PARTICIPANTS

CLOUD USER MEETING^{4} ONLINE
2021
63 PARTICIPANTS

WHAT OUR **USERS SAY** ABOUT THE de.NBI CLOUD

The de.NBI Cloud user meeting supports the interaction with our growing community. We are always interested to hear from you the needs of the scientific life sciences community allowing us to shape the future of the de.NBI Cloud according to specific use cases.

Steffen Neumann
from Leibniz Institute of Plant Biochemistry

"Having the de.NBI Cloud, we can easily crunch through some large-scale metabolomics data sets using the containerized workloads we have developed in the PhenoMeNal-H2020.EU project."

"The de.NBI Cloud has been a great resource for GHGA software development team to test the concept in a sandbox project, and utilize state-of-the-art cloud and container tools like OpenStack and Kubernetes."

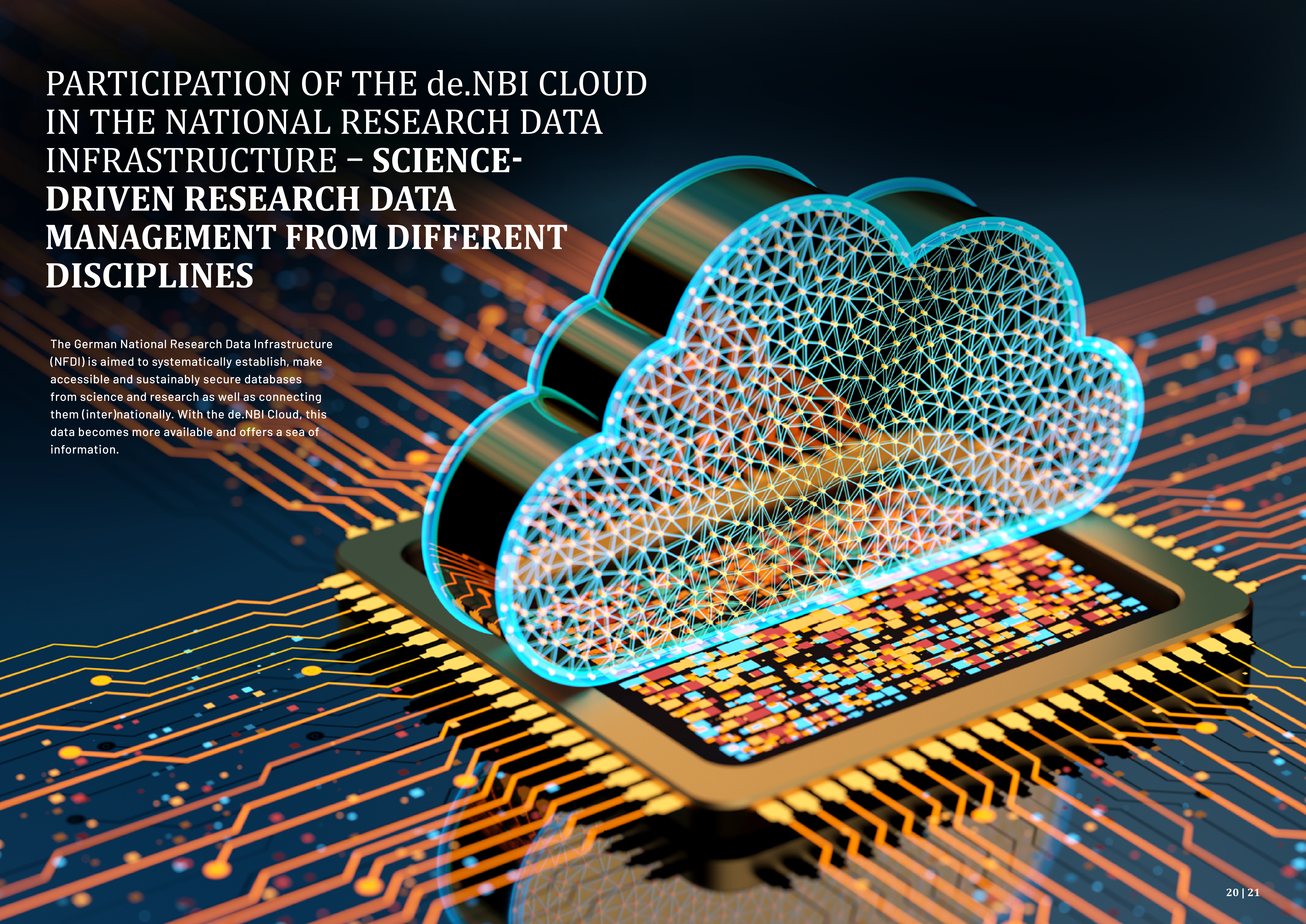
Koray Kirli
from The German Human Genome-Phenome Archive, DKFZ Heidelberg.

"A major advantage of the de.NBI Cloud is that we save time managing hardware and servers and instead focus on the actual science."

Dr. Matthias König
from Systems Medicine of the Liver, Humboldt-University Berlin
https://livermetabolism.com

PARTICIPATION OF THE de.NBI CLOUD IN THE NATIONAL RESEARCH DATA INFRASTRUCTURE – **SCIENCE-DRIVEN RESEARCH DATA MANAGEMENT FROM DIFFERENT DISCIPLINES**

The German National Research Data Infrastructure (NFDI) is aimed to systematically establish, make accessible and sustainably secure databases from science and research as well as connecting them (inter)nationally. With the de.NBI Cloud, this data becomes more available and offers a sea of information.



NFDI4Biodiversity

NFDI Consortium for Biodiversity, Ecology and Environmental Data

NFDI4Biodiversity is a consortium within the German National Research Data Infrastructure (NFDI) dedicated to the collaborative use of biodiversity and environmental data. Countless studies document the decline of biodiversity on our planet. To understand the complex coherences and to develop action items, policy, research and public authorities need reliable data. The mission of NFDI4Biodiversity is to mobilize existing data, develop efficient workflows, and provide services to support data archiving, findability and analyses. NFDI4Biodiversity's large partner network combines a broad range of scientific and technical competencies, including IT, software engineering, bioinformatics, and cloud technologies.

NFDI4Biodiversity – MISSION AND BACKGROUND

NFDI4Biodiversity is a consortium within the German National Research Data Infrastructure (NFDI) dedicated to the collaborative use of biodiversity and environmental data. Funding is obtained through the German Research Foundation (DFG), based on a 10-year agreement between the Federal Government and the federal states concerning the Establishment and Funding of a National Research Data Infrastructure (NFDI)[1].

Biological diversity is a highly important field of action in research as well as policy making. The less biodiversity there is, the poorer is our planet – and so are we. The approaches to recording biodiversity are as diverse as biodiversity itself, and equally heterogenous are the data sets generated. With the increasing use of genetic and -omics technologies, data-intensive imaging methods and sensor technology in ecology, the amount of the data obtained grows well beyond the scope of data gathered with traditional observation methods in the field. Consequently, in addition to trustworthy data archives, there is a need for computing and storage capacity.

NFDI4Biodiversity aims to mobilize, structure, and standardize relevant portions of such data sources [1]. Use cases include the German Barcode of Life project, long-term observations within the framework of the European eLTER network[2] as well as data streams from the AMMOD prototype data stations[3]. The consortium can build on extensive preliminary work of the German Federation for Biological Data (GFBio)[4]. As part of this collaboration, the data centers of natural history museums as well as established data archives such as PANGAEA[5] were connected to a common portal, workflows for data submission were standardized, and common (meta) data standards as well as tools and services related to research data management were established.

NFDI4Biodiversity will expand these efforts with additional partners and technologies. One major goal is to move data and services into the cloud in order to facilitate data flows and applications. To this end, several de.NBI service centers and cloud providers are represented in the NFDI4Biodiversity consortium, as well as partners representing governmental nature conservation and citizen science. As of June 2022, the consortium consists of 50 organizations from

[1] corresponding DFG project number: 442032008. For NFDI as a whole, see also: https://www.dfg.de/foerderung/programme/nfdi/
[2] https://www.lter-europe.net/lter-europe, last accessed on 2021-08-08.
[3] https://ammod.de/
[4] www.gfbio.org, last accessed on 2022-04-01.
[5] www.pangaea.de, last accessed on 2022-04-01.

all over Germany. By joining forces, the consortium pools professional, scientific and technical competences in order to provide users from research and practice with a broad service portfolio for biodiversity and environmental data.

WHICH DATA ARE RELEVANT FOR NFDI4Biodiversity?

NFDI4Biodiversity focuses on biodiversity data concerning animals, plants, and microorganisms all of which are collected in a variety of ways, for example as observation data recorded in the form of tables, photos, videos or audio files. These observations may occur once (e.g., during a walk or field excursion) or be part of targeted long-term studies. Specimens are also frequently collected and preserved for further analysis or description. In addition, there is a growing amount of genetic information on the observed species, i.e., sequence data – from so-called barcoding (a method for species identification via genetic markers) to genome sequences and metagenome analyses. Of equal interest are data from disciplines as metabolomics, glycomics or transcriptomics, which provide information on functional relationships. Furthermore, data on environmental conditions in the species' habitat, as well as on land use and colonization, are relevant. Due to the increasing use of sensors that measure certain environmental parameters over time, the amount of collected data is constantly growing.

These data are not only of interest to the NFDI4Biodiversity professional community, but also provide valuable information to other fields of research. At the same time, essential data for biodiversity research are being collected within other disciplines. There is, for example, a broad overlap with earth and environmental sciences, agricultural science, and cultural geography. In terms of society as a whole, there is traditionally a close link between biodiversity research and environmental and nature conservation (biodiversity monitoring). Therefore, within the consortium and beyond there is a high interest in a standardized structuring and description of the data to optimize their reusability. National and international networking is particularly relevant here, for example with the European Research Infrastructure for the Life Sciences ELIXIR[6], via the German Network for Bioinformatics Infrastructure de.NBI[7] and the Global Biodiversity Information Facility GBIF[8].

MOBILIZATION, ARCHIVING, AND PUBLICATION OF DATA

As mentioned above, GFBio already connected the data centers of the natural science collections as well as the three research data repositories. Within NFDI4Biodiversity, they play a central role as data and service providers. In the context of structured long-term archiving and data publication, they provide intensive curation and standardization of datasets by experts. This provides tangible added value with regard to later scientific use and enables, for example, the combination of datasets of different provenance.

In the course of the development of NFDI4Biodiversity, the circle of data centers will be gradually expanded. Also, within the use cases, data sets from the other partner organizations will be rendered ('mobilized') to prepare them to become part of the planned Research Data Commons (RDC) – a cloud infrastructure in whose environment the mobilized data will be available for broader applications.

[6] https://elixir-europe.org, last accessed on 2021-08-08
[7] https://www.denbi.de, last accessed on 2021-08-08
[8] https://www.gbif.org, last accessed on 2021-08-08

The technical connection of these resources will be one of the tasks during the project. The connection of additional data sources will be pursued, especially in collaboration with other NFDI consortia, in order to bring the standardization and simplification of workflows even further into the professional communities and to mobilize the largest possible amount of relevant data for the NFDI Research Data Commons.

→ OUR VISION: THE NFDI RESEARCH DATA COMMONS

Most of the collected biodiversity data NFDI4Biodiversity deals with is of cross-disciplinary and societal relevance. Depending on the research question and methodology, target-group-oriented data compilations and products are needed. In order to be able to draw as much added value as possible from the provided data, a basic infrastructure is needed that allows a wide variety of actors to a) access a large amount of heterogeneous data from different disciplines (mobilization), b) bring it into a comparable structure (integration and transformation), and c) process it in an event-related and target-group-specific manner (harmonization and provenance). This vision drives the development of the Research Data Commons (RDC).

The NFDI RDC is the technical framework for a variety of intermediary services that mediate between the needs of the data demand side and the producer side. These are organized into different layers that are intended to provide a) the producer side with a coupling of holdings and b) end users with flexible access to relevant data from different sources and associated services.

In the so-called mediation layer, the technical connection of the technically differentiated data sources and the data available there takes place. The transformation of this heterogeneous data makes it integrable and searchable. On top of this, a semantic layer is planned that enables semantic enrichment and description of the data, so that data from different disciplines can be translated into a common language and thus prepared for different target groups. In the application layer, users will find user-friendly products such as data portals, dashboards and other applications so that the harmonised data can be accessed, analysed with RDC-compatible, standardised tools or integrated with their own data. The aim is to create a kind of ecosystem of user-generated applications. At this level, data submissions are also possible in order to archive and publish research datasets from individual and collaborative research at one of the connected data centres.

Technically, the Research Data Commons are set up as a multi-cloud to ensure high flexibility and scalability. Hybrid constructions and edge components are also planned. It is planned to use already established cloud computing resources of the German Network for Bioinformatics (de.NBI). The individual components will be modular and communicate with each other via standardized interfaces (e.g. REST, gRPC). Deployment will be done using (Docker) containers managed via Kubernetes.

RDC development is defined as an NFDI cross-cutting topic. Several NFDI consortia have expressed interest in becoming involved in the development of a common interoperable infrastructure [2] (Figure 1).

→ INTEGRATING THE STATUS QUO

NFDI4Biodiversity builds on the existing services of GFBio [3]. For data producers, these include individual advice on RDM issues and the preparation of a data management plan (DMP) as well as a data submission system through which se-

quence data and other outputs from research projects are submitted centrally and transferred to the appropriate data centers. For data requests, there is a data portal that allows users to search across all connected data centers. Furthermore, they can use a software which allows them to process the searched data spatially and temporally. Additionally, there is a terminology service as well as a close cooperation with two software tools for ongoing data management in the research process (Diversity Workbench and BEXIS). Part of the IT development capacities in NFDI4Biodiversity will go into making these services 'fit for the cloud'.

ENLARGING THE PORTFOLIO: THE USE CASE PROJECTS

The consortium is working with a set of more than 20 representative use cases in order to ensure that the new service infrastructure meets the needs of the heterogeneous community. These include, for example, the European eLTER research infrastructure with its long-term ecological data, the collections of several professional societies specializing in specific animal or plant families, as well as actors in the fields of environmental protection, nature conservation, and citizen science.

While the vision of Open Science in politics and society often consists of free-flowing, FAIR data, the problems at the working level are others. In some use cases, support is needed for semantic enrichment of (meta)data, for example by providing species reference lists or a structured repository for observational data. Other needs concern basic IT services, such as the development of database solutions or web portals, or the provision of already existing data via standardized interfaces. Within the AMMOD use case, integration of sensor-generated data streams is being prototyped. Several natural science collections work on 3D digitization of specimens. To bring the resulting, rather large image data into software applications, powerful working storage is needed, which can be provided in the cloud.

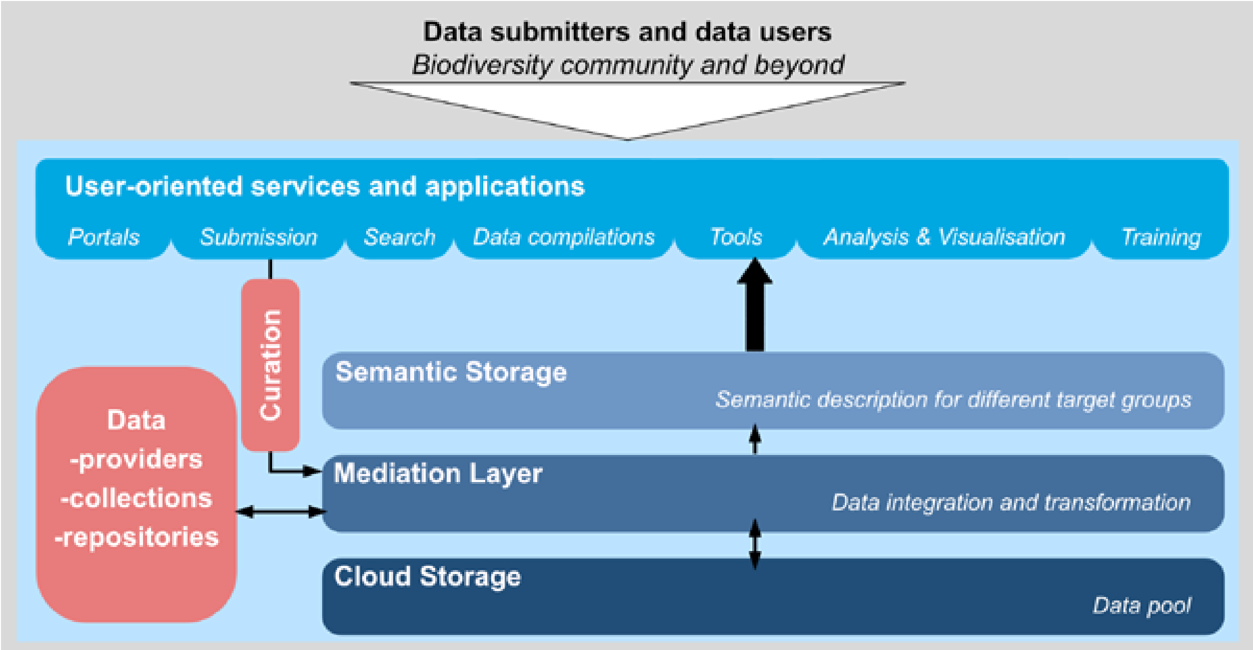
In any of these cases, the NFDI4Biodiversity network offers access to new technologies and added value in terms of best practices and workflows, which can be adapted by the partners in order to integrate with the common infrastructure.

→ CHALLENGES

The establishment of an NFDI as a network of networks is a challenging socio-technical experiment for everyone involved. In particular for NFDI4Biodiversity with its broad-based partner organizations, the challenge is to build bridges between scientific and civil society stakeholders whose interests in data use must be reconciled. Mobilising data is hard enough within academia, where many researchers and research groups understandably place a high value on control over their data and where the effort required to make it accessible and publish it is still rarely acknowledged.

For non-scientific partners, it is important to combine the mobilization of existing data for research with concrete added value for their own day-to-day business. This can also mean accelerating the digitization of work processes, which is already in the pipeline. Efforts to consolidate national biodiversity monitoring, accompanied by the establishment of new data centers and parallel networking initiatives, advance our goal of effective data mobilization across sectors. Such powerful efforts at the federal and state levels mean that NFDI4Biodiversity will need to position itself well in both science and environmental policy to gain recognition as a network of stakeholders. The provision of concrete, useful and appealing services and tools that support a variety of applications in and together with the community will be a key factor of success. Access to the tried-and-tested cloud services of de.NBI is part of the equation.

Figure 1: Schematic representation of the cloud infrastructure Research Data Commons (RDC). The RDC are organized in different layers that are intended to provide the end users with flexible access to relevant data from different sources and associated services.


Data submitters and data users
Biodiversity community and beyond
User-oriented services and applications
Portals
Submission
Search
Data compilations
Tools
Analysis & Visualisation
Training
Curation
Data
-providers
-collections
-repositories
Semantic Storage
Semantic description for different target groups
Mediation Layer
Data integration and transformation
Cloud Storage
Data pool

CONCLUSION & OUTLOOK

The NFDI4Biodiversity consortium assembles a broad variety of stakeholders in the biodiversity community. Services for the Long Tail of Science have been established out of the GFBio network to help open up data from individual and collaborative research and to provide services for working groups within biodiversity research. This approach will be significantly expanded in NFDI4Biodiversity to include data from governmental agencies and citizen science. Following technical advances in the field, the demand for big data applications will rise significantly. Bringing services into trusted and tailored cloud environments is therefore an important step to enable a variety of data-related applications. In turn, the aim is to generate tangible added value for science, politics and society.

ACKNOWLEDGEMENT

The authors of this article have used various preparatory works from the GFBio and NFDI4Biodiversity networks to create this portrait, and references have been made where possible. We are grateful to all the people who have helped to shape and support the structures and results presented here. With regard to this article special thanks go to Katharina Frohne for translation and editing of the text.

Links:
www.nfdi4biodiversity.org
www.gfbio.org

REFERENCES: **[1]** NFDI4BioDiversity - A Consortium for the National Research Data Infrastructure (NFDI). Zenodo 2020. DOI: http://doi.org/10.5281/zenodo.3943645 **[2]** Zenodo 2021. NFDI Cross-cutting Topics Workshop Report. Zenodo. DOI: http://doi.org/10.5281/zenodo.4593770 **[3]** Potential für ein starkes Netzwerk zwischen GFBio und FDM-Beratenden an Universitäten und Forschungsinstituten. Bausteine Forschungsdatenmanagement, Nr. 1 (März) 2021; German:22-31. DOI: https://doi.org/10.17192/bfdm.2021.1.8311.

AUTHORS: Barbara Ebert[1], Michael Diepenbroek[1], Judith Sophie Weber[1], Ivaylo Kostadinov[1], Frank Oliver Glöckner[2,3,4]

[1] *GFBio - Gesellschaft für Biologische Daten e.V., c/o Research II Campus Ring, 128759 Bremen*

[2] *MARUM - Center for Marine Environmental Sciences, University of Bremen, Leobener Str. 8, D-28359 Bremen*

[3] *Alfred Wegener Institute, Helmholtz Center for Polar and Marine Research, Am Handelshafen 12, 27570 Bremerhaven,*

[4] *Jacobs University Bremen gGmbH, Campus Ring 1, 28759 Bremen*

GHGA AND de.NBI CLOUD

German Human Genome-Phenome Archive for FAIR, safe and secure omics data sharing on a federated infrastructure

A growing body of research connects genomics data to improvement of clinical diagnosis. However, the decreasing cost of technologies that ease access to sequencing on both research and clinical studies pose a challenge for storing, searching, and accessing these datasets. The German Human Genome-Phenome Archive (www.ghga.de) is setting up a federated data network in Germany to provide harmonized services to enable data sharing in collaboration with international partners like the European Genome Archive. de.NBI Cloud sites at the two largest German nodes - Heidelberg and Tübingen - will speed up the start-up phase by providing the required efficient and secure infrastructure.

GHGA: SETTING UP A NATIONAL ARCHIVE FOR GENOMICS WITHIN THE NFDI

Collecting genomic data from patients is an essential part of biomedical research, with major applications in basic biology, translational research and medicine. Increasingly, this type of high-volume data is also generated in the context of personalized therapies as a tool for precision diagnostics. The rapid growth of available data is a major challenge, but also an unprecedented opportunity for research. Extensively characterized data sets are currently generated at a growing number of research institutions and hospitals across Germany. Integration of these local data sets is essential for their optimal use in diagnostics and research since modern computational services such as machine learning or artificial intelligence only unfold their full potential with access to combined big data.

The need to make data accessible for the research community according to the FAIR principles [1] must always be weighed against the protection of the patient's privacy. While other countries have already established large national programs to generate, analyse and share

The German Human Genome Archive

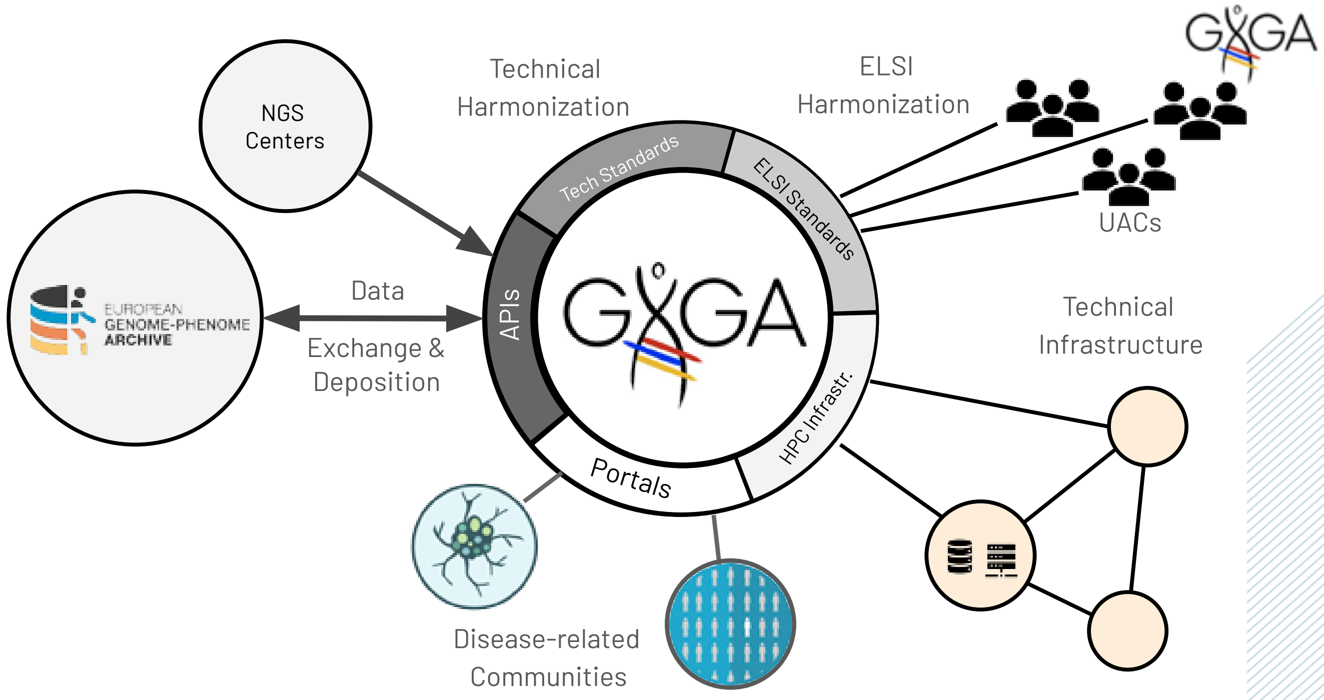


FIGURE 1: GHGA and surrounding communities

genomics data on a broad scale, e.g. the Genomics England program[1], the Estonian Biobank[2] or the deCODE Genetics program in Iceland[3], Germany is lacking behind with respect to an overarching initiative to archive and share human genomic data in an appropriate way. To fill this gap, recently the **German Human Genome-Phenome Archive (GHGA)** was initiated within the National Research Data Infrastructure (NFDI e.V.).

GHGA's core mission is to establish a national infrastructure for the submission, controlled access, management and long-term secure archival of sequence-based human omics data (e.g., genomics, transcriptomics, proteomics). This will allow highly sensitive genome data to be merged, saved and analyzed within a uniform and data protection-compliant framework.

Being deeply rooted in other international initiatives, most importantly the European Genome Archive (EGA, see below), GHGA will build on existing German omics data providers such as the Next Generation Sequencing Competence Network (NGS-CN) and their IT infrastructures to establish a harmonized distributed infrastructure for omics data.

Going significantly beyond mere archival functionality, in the mid-term GHGA will enable and democratize FAIR access to even the largest population-scale datasets. Responding to the need of the research community to analyze protected human large-scale data in a user-friendly manner, GHGA will build an analytics platform for distributed data processing and research use. Secondary use of clinical omics data in research will enable biological discovery and allow replication across distinct cohorts. Access portals tailored to community-needs combined with curated reference data collections will ensure the utility of GHGA's datasets to researcher and clinician communities.

Collaborations with other NFDI consortia such as NFDI4Health and NFDI4Microbiota, can combine healthcare, medical research, and public health data with omics-centric raw data archival, processing and analytics. Linking these data sets creates an invaluable bridge between biomedical research and healthcare, opening the door for scientists in Germany to participate in key international research networks and boost the field of genome sciences in Germany.

INTERNATIONAL NETWORKING AND STANDARDS

The current home of personally identifiable genetic and phenotypic data collected within biomedical research projects across Europe is the European Genome Phenome Archive (EGA)[4], hosted as a joint venture between EMBL-EBI in the UK and CRG in Spain. As more and more countries start personalized medicine initiatives, the need for national infrastructure has become apparent and has led to introduction of the concept of a federated European Genome Archive (fEGA), consisting of a network of national nodes connected via joint standards and infrastructures[5]. With increased stringency of information access for genome-phenome data that is produced in the healthcare context, both European, national and local legislation comes into play.

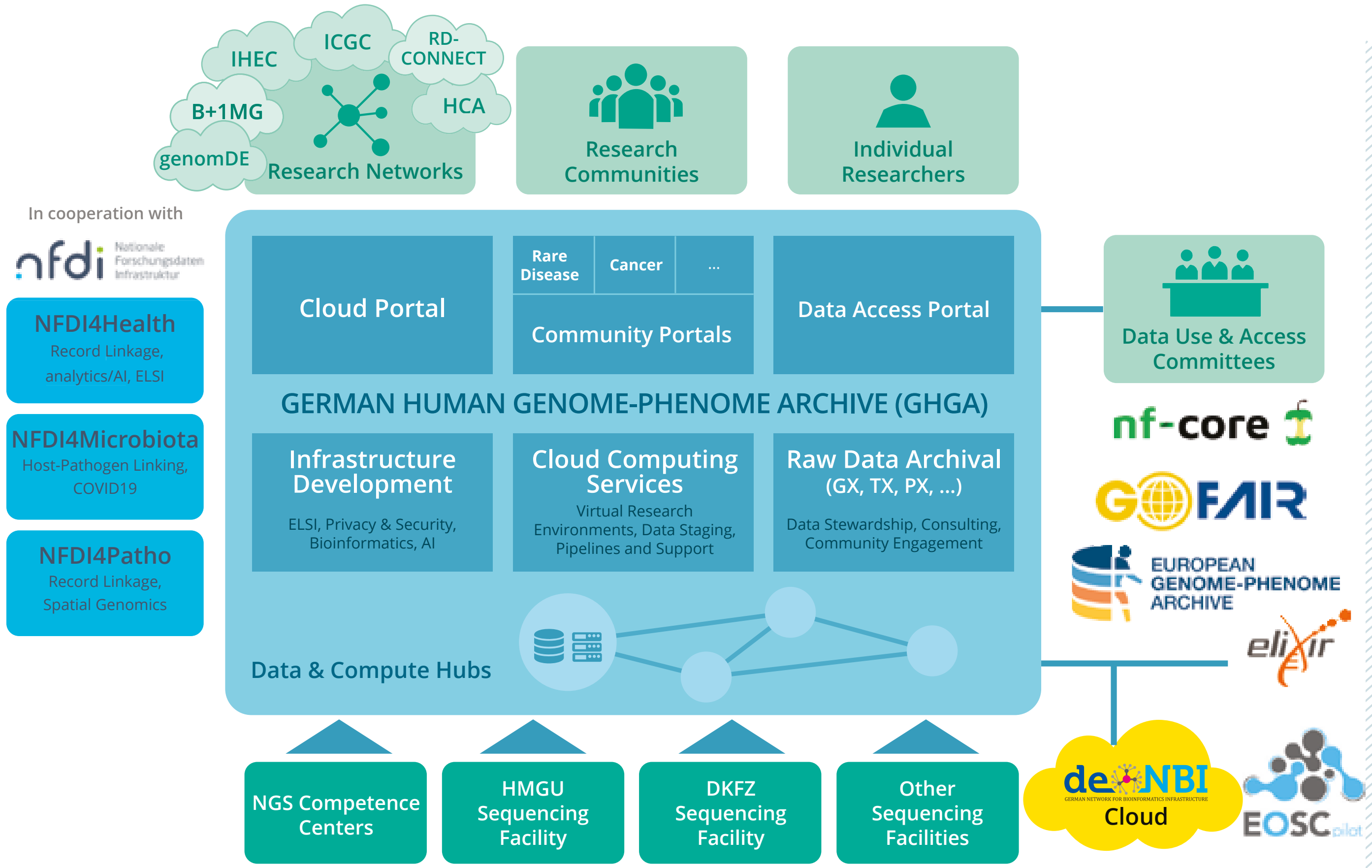

IHEC
ICGC
RD-CONNECT
B+1MG
HCA
genomDE
Research Networks
Research Communities
Individual Researchers
In cooperation with
nfdi Nationale Forschungsdaten Infrastruktur
NFDI4Health
Record Linkage, analytics/AI, ELSI
NFDI4Microbiota
Host-Pathogen Linking, COVID19
NFDI4Patho
Record Linkage, Spatial Genomics
Cloud Portal
Rare Disease
Cancer
...
Community Portals
Data Access Portal
GERMAN HUMAN GENOME-PHENOME ARCHIVE (GHGA)
Infrastructure Development
ELSI, Privacy & Security, Bioinformatics, AI
Cloud Computing Services
Virtual Research Environments, Data Staging, Pipelines and Support
Raw Data Archival
(GX, TX, PX, ...)
Data Stewardship, Consulting, Community Engagement
Data & Compute Hubs
Data Use & Access Committees
nf-core
GO FAIR
EUROPEAN GENOME-PHENOME ARCHIVE
elixir
NGS Competence Centers
HMGU Sequencing Facility
DKFZ Sequencing Facility
Other Sequencing Facilities
de.NBI Cloud
EOSC pilot

FIGURE 2: Overall structure of the GHGA consortium and connection to international activities.

As a national node of the fEGA, GHGA can follow national regulations on data protection and at the same time be closely linked to international data infrastructures. This makes the data sets easy to find, accessible and optimally usable for national and international research – enabling German researchers to shape future international standards for data exchange and take on leading roles in international research consortia (e.g., the 1+ Million European Genomes Initiative [2] and The Global Alliance for Genomics and Health (GA4GH)).

GA4GH focuses on policies and technical standards for responsible genomic data sharing within a human rights framework[6]. GA4GH has already established a range of standards, of which GHGA would like to utilize the following: (1) Data Use Ontology (DUO) for standardized consent information, (2) Passports for data access policy, (3) Crypt4GH for data encryption, (4) Data Repository Service (DRS) for data storage, and (5) Workflow Execution Service (WES) for data processing. Besides aligning to set international standards, GHGA aims to actively engage and help define new standards.

DATA FEDERATION WITHIN GHGA - THE NEED FOR A FEDERATED ARCHITECTURE

Big data comes with many promises in the healthcare setting. Diagnosis of diseases that result from mutations in the genome and understanding the underlying mechanism of such phenomena well

[1] https://www.genomicsengland.co.uk/
[2] https://genomics.ut.ee/en/content/estonian-biobank
[3] https://www.decode.com/
[4] https://ega-archive.org/
[5] https://ega-archive.org/federated
[6] https://pubmed.ncbi.nlm.nih.gov/35072136/

enough to explore treatment options is a highly challenging problem. Artificial intelligence (AI) is an emerging solution for these complex tasks and combines computational systems, data management and AI algorithms to provide researchers with tools to get most out of their research. Access to large, well-curated datasets is crucial for researchers since AI applications depend on very large training datasets for homogeneous coverage and good performance.

GHGA will facilitate sharing of existing and future datasets in Germany for many areas of research including cancer and rare diseases. Diagnosis and treatment of cancer depends on precise detection of the disease subtype, and mapping the spectrum of classifications requires many datasets to come together. On the rare disease side, patients often suffer from a yet undiagnosed disease with rare mutations causing that, and the need for data discovery is massive.

While the omics data production rate is steadily increasing, the resulting data is often stored only locally at the original institutions. Although there are various technical options for bringing these data together, privacy and security concerns and the surrounding legislation renders centralized infrastructure solutions unsuitable. Many countries and organizations use federated architectures to overcome these issues. In a federated data architecture, the data is stored and managed in different locations while these nodes are harmonized at the service level.

GHGA will set up a federated architecture made up of a central node located at DKFZ and an expandable list of local data hubs. Currently planned local data hubs are located in Heidelberg, Tübingen, Munich, Cologne, Kiel and Dresden, spanning multiple institutions. Choosing this federated approach, GHGA will be able to continue onboarding new data hubs later on.

The GHGA central node in Heidelberg will be focused on handling data operations, search queries, download requests and metadata exchange with EGA. Local data hubs will be responsible for secure storage, encryption, processing, and archiving of the omics datasets. The GHGA software development team will provide the framework required for these services with guidance on their implementation. In Heidelberg and Tübingen, these services reside on the respective de.NBI Cloud.

TECHNICAL CONCEPT OF GHGA AND POSSIBLE IMPLEMENTATIONS WITHIN THE de.NBI CLOUD

During the first year, GHGA focused on planning, recruitment, and start-up activities. Various teams were formed spanning ELSI, outreach, data stewardship, software development, bioinformatics and metadata task areas. Being an infrastructure project, software development and implementation takes a central position and needs to quickly build the tools required by other task areas.

The GHGA software development team is building a suite of microservices that will be distributed between the central and local hubs. This distributed complexity which enables a healthier growth cycle for GHGA comes at the expense of an increased initial complexity. de.NBI Cloud has been a great resource for the GHGA software development team to test the concept in a sandbox project, and utilize state-of-the-art cloud and container tools like OpenStack and Kubernetes.

As the two major GHGA sites - Heidelberg and Tübingen - are running local de.NBI Cloud instances, they will become the early data hubs of the project. In addition, the direct connection to these clouds will enable GHGA to democratize big data by allowing researchers throughout Germany to work on huge data sets using state-of-the-art compute resources (CPU and GPU) without local infrastructure requirements.

CONCLUSION & OUTLOOK

As genomics and other omics research is shifting from bench to bedside, the importance of protecting, sharing and analysing this data is increasing. GHGA will speed up this transition by providing an efficient data dissemination platform with utmost respect to data privacy and consent. Availability of the de.NBI Cloud enables GHGA to develop our architecture in a harmonized environment across our federation and to run all GHGA services on the existing efficient cloud infrastructure.

REFERENCES: [1] Scientific Data 2016;3:160018. DOI: https://doi.org/10.1038/sdata.2016.18. **[2]** Nature Reviews. 2019;20:693-701. DOI: 10.1038/s41576-019-0156-9. DOI: https://doi.org/10.1038/s41576-019-0156-9 **[3]** Nature. 2020;578(7793):82-93. DOI: https://doi.org/10.1038/s41586-020-1969-6.

AUTHORS: Koray Kirli[1] , Ulrike Träger[1], Jan Eufinger[1], Ivo Buchhalter[1,2], Jens Krüger[3], for the GHGA Consortium

[1] German Human Genome-Phenome Archive (W620), DKFZ, Im Neuenheimer Feld 280, 69120 Heidelberg
[2] Omics IT and Data Management Core Facility (W610), DKFZ, Im Neuenheimer Feld 280, 69120 Heidelberg
[3] University of Tübingen, High Performance and Cloud Computing Group, IT Center, Geschwister-Scholl-Platz, 72074 Tübingen

DataPLANT

Platform for Research Data Management

Modern science relies on effective research data management services and infrastructures facilitating the acquisition, processing, exchange and archiving of research data. In fundamental plant research, modern approaches need to integrate analyses across different system levels ranging from *omics to imaging data. This is necessary to understand system-wide molecular physiological responses as a complex dynamic interplay between genes, proteins and metabolites. The overall goal of DataPLANT (https://nfdi4plants.de) is to provide research data management practices and tools for the plant community, relying on the de.NBI Cloud as infrastructure basis.

FUNDAMENTAL PLANT RESEARCH AND RESEARCH DATA MANAGEMENT

The main goal of DataPLANT is to support fundamental plant researchers in research data management [1]. Fundamental plant research employs modern approaches including genomics, transcriptomics, proteomics, metabolomics, phenomics, and imaging data. As such system-wide molecular physiological responses can be related to individual genes, proteins and metabolites, deciphering the complex dynamic interplay in plants.

DataPLANT as a compact consortium addresses the important core issues of a subject-oriented data management in three central task areas on (1) standardization, on (2) software, services and infrastructure, and on (3) personal support on-site. Especially the latter is provided through Data Stewards, a core element of the DataPLANT strategy.

DataPLANT follows a gradual and iterative approach, ensuring the commitment and alignment of expectations of all stakeholders. This is mirrored in the modular and federated technical infrastructure building on the de.NBI Cloud infrastructure. The evolution of standards, tools and services get accompanied by comprehensive training to ensure data literacy through lectures, courses, workshops and hackathons and providing open training material. The set of tools and services developed in the first year of DataPLANT focused on linking the pre-existing digital landscape of the average plant scientist to various cloud services.

The above-mentioned areas are also addressed at the level of the NFDI, specifically the NFDI association (https://www.nfdi.de/verein/) and their individual sections [2].

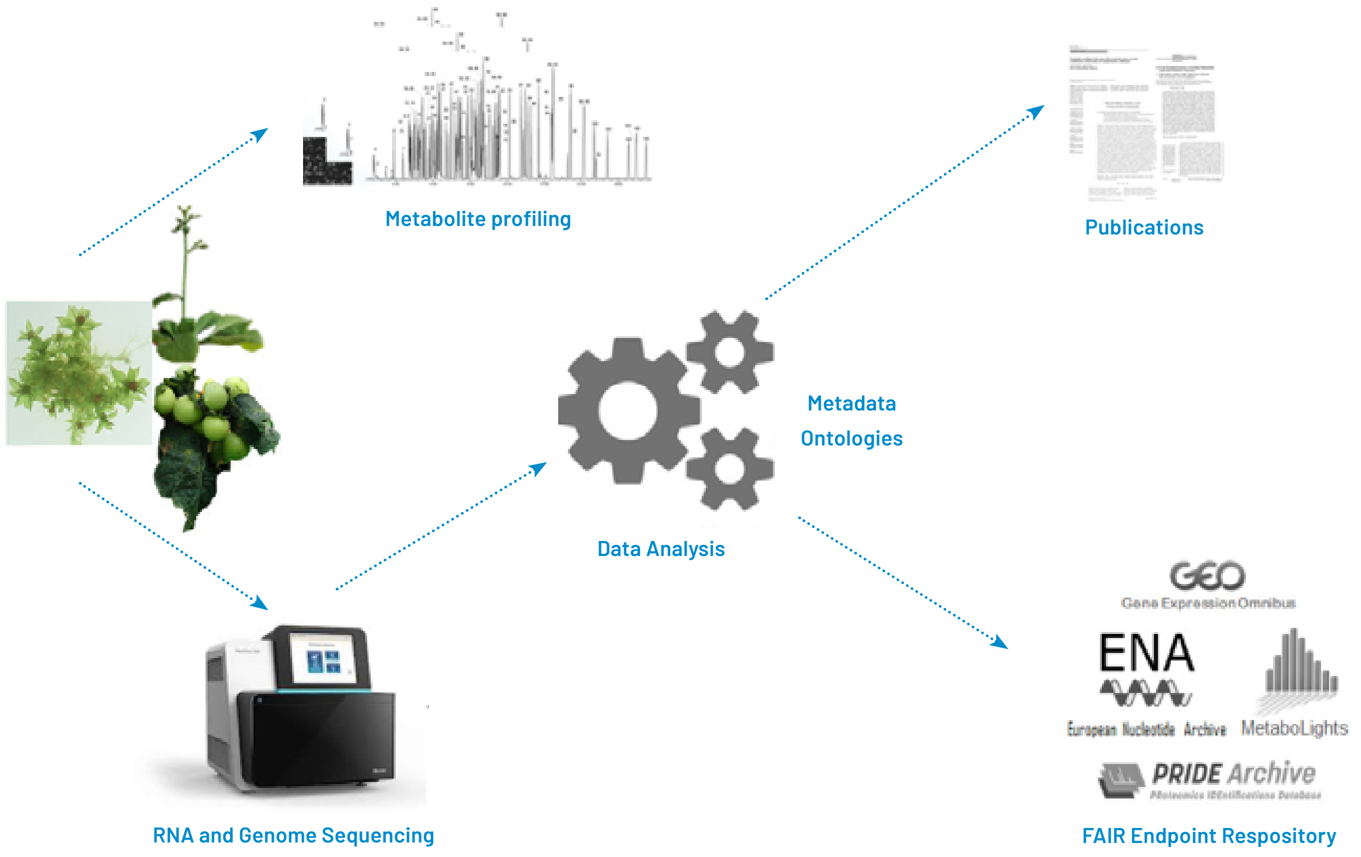


FIGURE 1: DataPLANT is supporting fundamental plant researchers on research data management.

DataPLANT SERVICES

The services of DataPLANT rely on the compute and storage cloud facets of the de.NBI Cloud. They cover a broad range of mechanisms of function ranging from stand-alone webservices, over integrated services connected through mutual trust, to versioning and publication services. All services are conceptualized to have a low entry barrier for novice users while providing extra value, facilitating scientific research.

Well annotated research data objects – ARC

DataPLANT defined a specification of the Annotated Research Context (ARC), a structured way to organise research data oriented along the ISA-TAB logic. The backend infrastructure for actual data versioning and sharing relies technology-wise on Git LFS in combination with GitLab. The de.NBI Cloud provides the technical basis for implementing the backend infrastructure for the data management of the plant research community.

Metadata – ISA-TAB – SWATE

In DataPLANT ARCs form a single-entry point to structured research data and allow practical research data management from a user perspective. They are Git repositories to track changes and build on existing standards like ISA-TAB for administrative and experimental metadata plus CWL for analysis and workflow metadata. ARCs are digital objects that fulfil all FAIR principles and are therefore referred to as FAIR Digital Objects (FDO) including a Research-Object-Crate (ROC) manifest for international interoperability. The fundamental axiom of DataPLANT is to enable users and support their data management strategies. Therefore, in addition to personal support, we build tools and services focusing on the ARC for digital assistance. While the tool ARC-commander facilitates creation and management of ARCs, help for users with the metadata annotation process is provided through an Excel Add-in named SWATE (Swate Workflow Annotation Tool for Excel).

Ontology management – Swobup

To support assisted research data management SWATE helps with the metadata annotation process based on the SWATE database, which integrates the required external ontologies. The database is implemented as cloud service populated from a public GitHub repository. To enable the users to use new, non-integrated terms, Swobup (Swate OBO Updater) bridges the gap to the NFDI4plants ontology, which stores these terms temporarily until incorporation into existing ontologies.

Authentication and authorization – LIFE SCIENCE LOGIN – Keycloak

Similar to the de.NBI Cloud, DataPLANT relies on the LIFE SCIENCE LOGIN for user authentication and authorization. As such, users authenticate themself to DataPLANT services using their Life Science Login and credentials from their home institutions which act as identity providers. Further identity providers such as ORCID may be used. Building on Keycloak as access management system, a continuous chain of trust is established between the user and DataPLANT services, as well as between the services, offering a secure and resilient environment for their research data.

Workflows – Galaxy and Nextflow

Galaxy and Nextflow are workflow systems to allow simple to complex analyses chains of bioinformatics tools. Both approaches can readily be used on ARCs, tight integration into the DataPLANT service environment is anticipated in the near future following the PaaS paradigm on top of the de.NBI Cloud. As such scientists are able to use state-of-the-art data analysis pipelines facilitating their research. Workflow-based results of analysis pipelines become part of their corresponding ARC. Researchers are enabled to run data analysis workflows on their data, share their analyses with others, and enable others to reuse the analysis pipelines.

Publication Service – InvenioRDM

ARCs comprising well-annotated research data objects, contain various *omics or imaging data, along with workflows and the analysis results. It's good scientific practice and the main motivation for the NFDI to publish these datasets and share them with peers and community. Usually, this happens through the deposition in e.g. suitable NCBI or EBI repositories but normally only processed secondary data is deposited. To close this gap in the DataPLANT Publication Service based on InvenioRDM has been established, enabling the deposition and publishing of whole ARCs, including raw data, scripts, pipelines, result data and the whole of metadata. Published ARCs receive a DOI as persistent identifier.

de.NBI CLOUD INFRASTRUCTURE AS BASIS

The DataPLANT science gateway as assembly of individual services is made accessible through a stable and highly available entry point residing above the cloud layer. Services are addressed through URL or services ports. All aforementioned DataPLANT services are hosted on the de.NBI Cloud infrastructures at the sites in Freiburg and in Tübingen.

The services of DataPLANT follow the cloud paradigm: they are meant to be automatically deployed on-demand in appropriate cloud infrastructures like the OpenStack-based de.NBI Cloud. This generates resilient operation for each service using deployment technologies like Ansible or Packer.

The layered DataPLANT infrastructure is complemented through a persistent storage layer forming the basis for the GitLab version service.

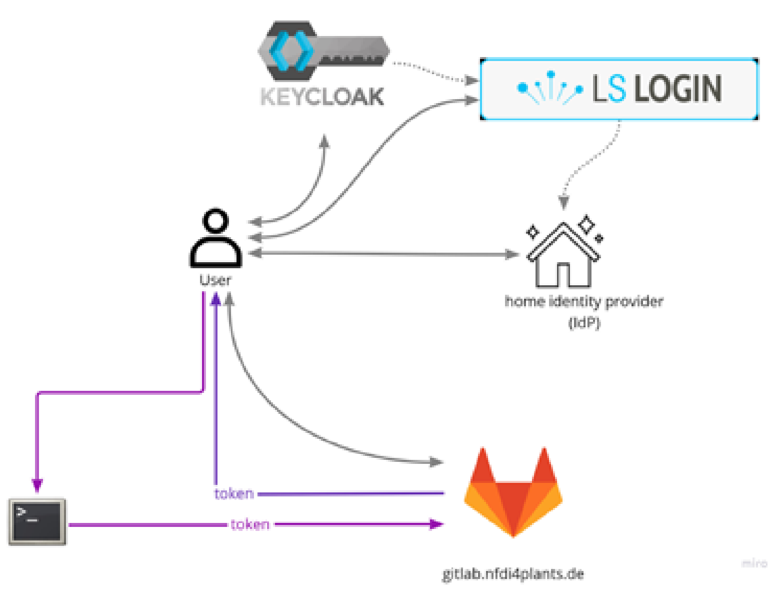
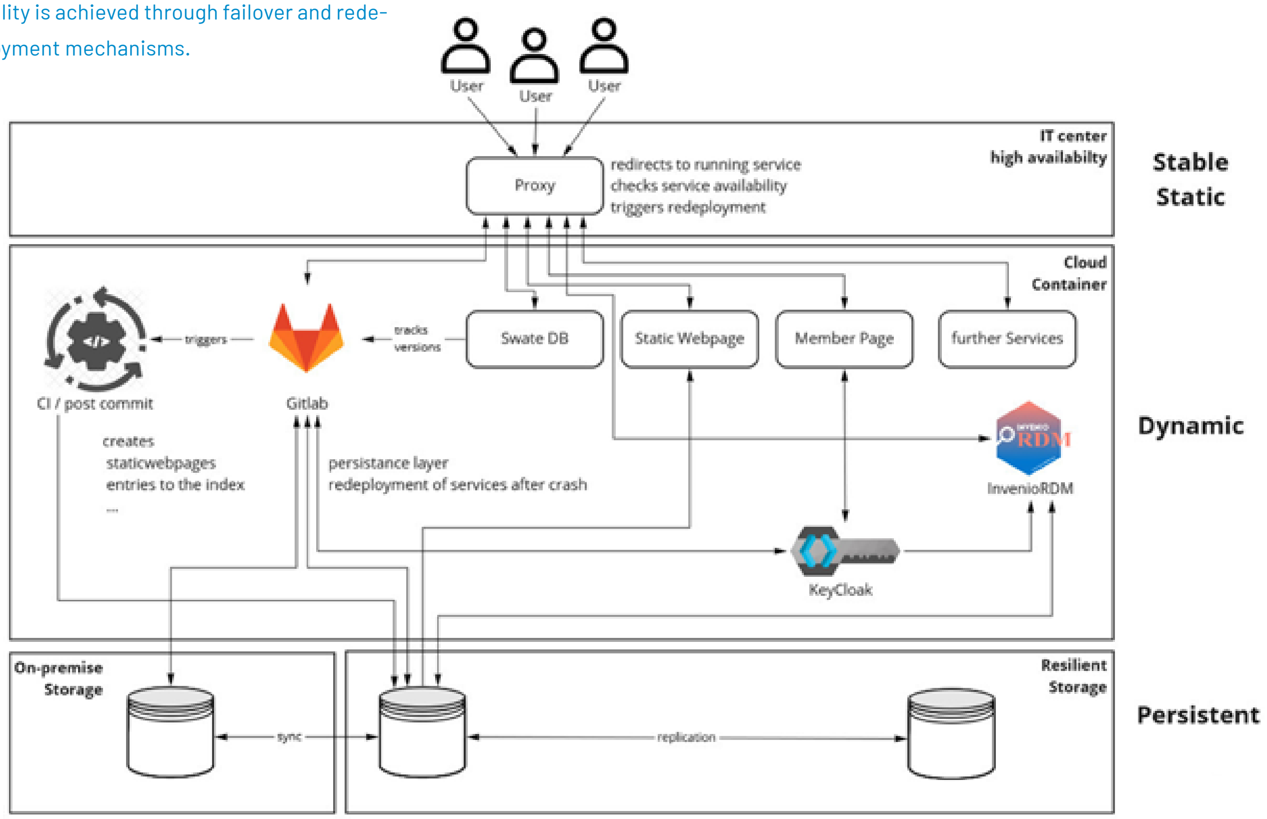


FIGURE 2: DataPLANT relies on the Life Science Login for identity management, part of the information flow among the DataPLANT services is depicted.

CONCLUSION

Building on de.NBI Cloud resources and the expertise from the de.NBI consortium DataPLANT is providing modern research data management to the fundamental plant research community. With the help of data stewards, researchers are enabled to store, annotate and process their research data. Well-annotated research data, the ARCs, facilitate data sharing among peers, improve reproducibility and promote the gain of scientific insight.

FIGURE 3: The DataPLANT environment is built on three layers, decoupling services and infrastructure. An optimal service availability is achieved through failover and redeployment mechanisms.



REFERENCES: [1] BFDM 2021, 2:46-56. DOI: https://doi.org/10.17192/bfdm.2021.2.8335. **[2]** Zenodo 2020; DOI: https://doi.org/10.5281/zenodo.3895209.

AUTHORS: Jens Krüger[1], Timo Mühlhaus[2], Björn Usadel[3], Dirk von Suchodoletz[4], Cristina Martins Rodrigues[4]

[1] *University of Tübingen, High Performance and Cloud Computing Group, IT Center, Wächterstraße 76, 72074 Tübingen*

[2] *Technische Universität Kaiserslautern, Computational Systems Biology, Erwin-Schrödinger-Straße 52, 67663 Kaiserslautern*

[3] *Forschungszentrum Jülich GmbH, IBG-4 Bioinformatics, Wilhelm-Johnen-Straße, 52428 Jülich*

[4] *University of Freiburg, eScience, Fahnenbergplatz, 79085 Freiburg im Breisgau*

NFDI4MICROBIOTA

Enabling data-heavy research on microorganisms and their communities

As one of currently 19 NFDI (National Research Data Infrastructure) consortia, NFDI4Microbiota - launched in October 2021 - aims to support the German microbiology research community in many ways: definition and establishment of standards for (meta-)data, computational resources, mass storage, long-term preservation, training and community building. To achieve this efficiently NFDI4Microbiota builds on infrastructure provided by de.NBI. This article reflects on the future work of NFDI4Microbiota and the role of de.NBI within.

INTRODUCTION

Microbes including bacteria, archaea, unicellular eukaryotes, parasitic microorganisms and viruses have a strong impact on human, animal and environmental health. Understanding these species and their interaction has relevance for agriculture, water treatment, ecosystems remediation, and the diagnosis, treatment as well as prevention of various diseases and much more. Microbiological research today provides several potential solutions to address existing and upcoming needs. Thanks to technology advancements that allow high-throughput molecular characterisation of microbial species and their communities there is rapid growth in research output.

NFDI4Microbiota's mission is to serve and link the diverse microbiology subcommunities by improving the accessibility and quality of research data. It intends to make microbiological data generation, management, interpretation, sharing and reuse easier. The consortium will do this through constructing a German microbial research network as well as by offering a cloud-based infrastructure that will enable the storage, integration, and analysis of microbiological data, particularly omics data. In addition a training program will enable broad and effective use of these services. The adoption of standards of research data is another important objective. This should ensure the re-use and reproducibility of research in accordance with the FAIR [1] and Open Science principles. To reach this NFDI4Microbiota aims to bundle existing capacities, and make them generally available in a structured manner instead of building new infrastructure where possible. The consortium consists of ten well-established partner institutions and an extensive network of more than 50 participating institutions.

Cloud-based compute infrastructure

NFDI4Microbiota will offer computational resources and storage solutions to the microbiology research community in order to make data-heavy projects easier to conduct and to enable researchers to perform large scale analyzes independent of their local IT setup. The de.NBI infrastructure serves as the key platform and thereby will help to provide the underlying scalable, high performance, cloud-based computing and store infrastructure.

In collaboration with de.NBI, NFDI4Microbiota members will provide services and (meta-)data in scalable cloud computing environments. This approach enables efficient scaling by pooling computational resources. As part of this virtual environments will allow maximum, flexibility and software-container-based solutions enable scientists to generate reusable environments or workflows to ensure reproducible research.

DATA STORAGE AND FINDABILITY

Developing and defining data and metadata standards, as well as the effective and long-term storage and provision of microbial metadata and data are key objectives of NFDI4Microbiota. In order to achieve these goals, NFDI4Microbiota will provide a platform for scalable and accessible data storage and to this end will use the object storage provided by the de.NBI Cloud to store the actual data. NFDI4Microbiota will also provide a platform for data submission from our storage system to a number of reputable archives, such as the European Nucleotide Archive (ENA). This system will not only take care of all repository-specific parts of the submission, but it will also ensure that the data and metadata meet all submission standards. If the repository supports it, data can also be sent in for restricted access until it is published. Access to and findability of data in public repositories is another key feature of data management. NFDI4Microbiota will create tools for searching and retrieving data from public repositories, allowing users to query many repositories at the same time.

DIGITAL ARCHIVING

NFDI4Microbiota will support microbiologists with long-term archiving of their research data and associated metadata. Proper archiving starts with selecting suitable file formats, which NFDI4Microbiota will give advice on, as well as on integrity checks. Archiving of molecular data will be done in EMBL-EBI repositories, whereas other data types will be published in the PUBLISSO Repository for Life Sciences, and subsequently preserved in ZB MED's dark archive. Bitstream preservation of data and metadata will be done in the de.NBI Cloud.

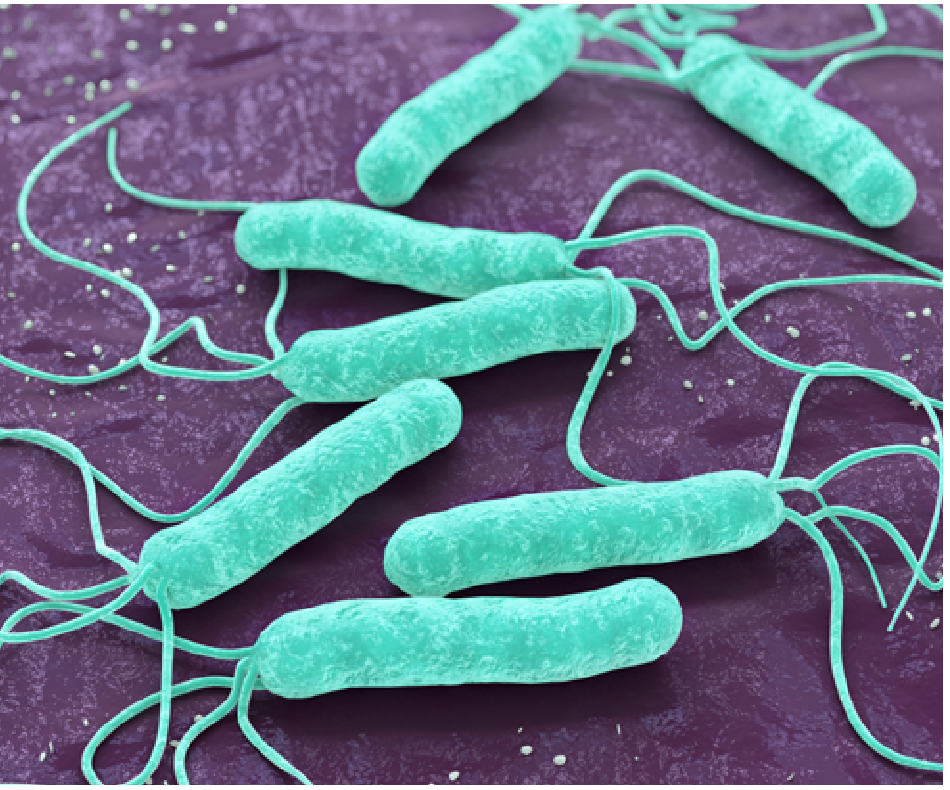
NFDI4Microbiota also intends to establish digital preservation solutions for data types that do not yet have dedicated repositories. The first step will be to raise awareness about the importance of archiving research datasets over the longterm. To assess the archival value of such datasets, criteria will then be defined together with the microbiology research community. To standardize and control the quality of archived data and metadata, NFDI4Microbiota will also select and implement relevant metadata standards. Last but not least, NFDI4Microbiota will provide training, consulting and individual support on digital archiving for the community.

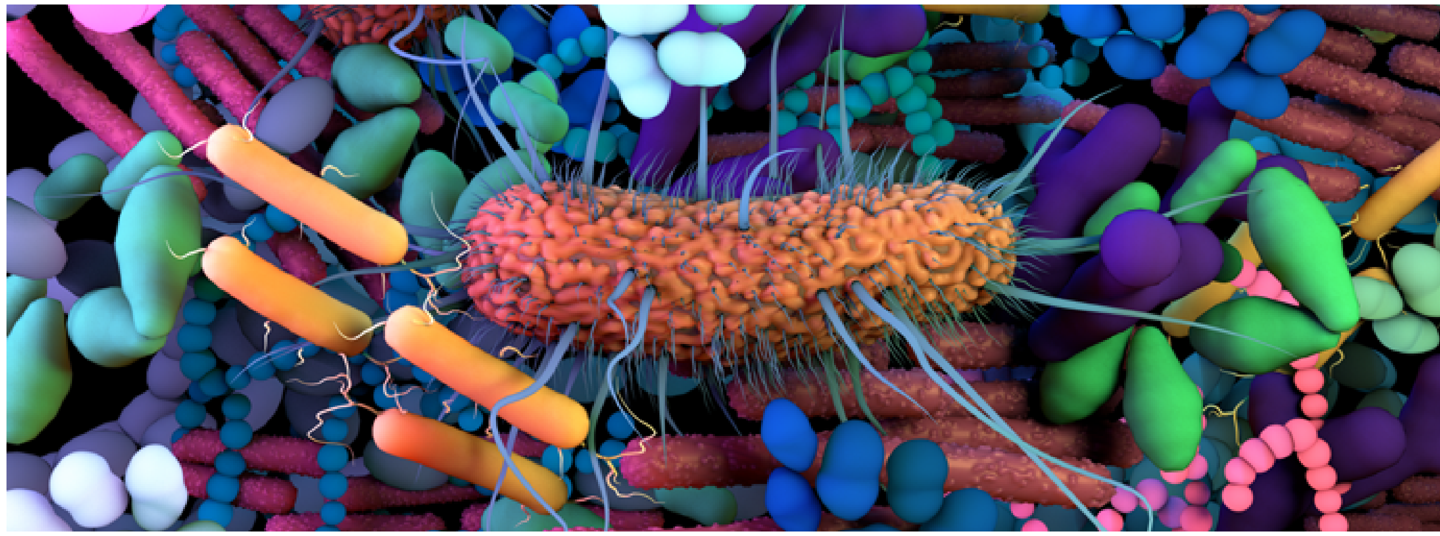
TRAINING

One major task of NFDI4Microbiota is the coordination of training events and to create a comprehensive digital literacy training program. Training topics include Infrastructure & Software (i.e. cloud and de.NBI, bioinformatics, workflow engines), Research Data Management (including data science skills and cloud compute platform services), as well as domain-specific training (e.g. -omics, bioinformatics). Online courses and online training material will be generated to foster the FAIR data and Open Science principles [1] as well as best practices in research data management.

STANDARDS AND DEFINITIONS

The re-use of research data as well as the reproduction of research results is good scientific practice. At the moment, however, it remains a challenge to use previously deposited data and scientific findings. The numerous deviating procedures for this, their often incomplete documentation and missing metadata are a regular challenge. Having uniform, easily accessible and documented standards of metadata and research processes is a prerequisite for effective and efficient reuse and integration of data. NFDI4Microbiota meets these challenges by facilitating the exchange and reuse of data. In addition, it supports simple and open access to microbial and related data in accordance with the FAIR principles. The developed guidelines ensure open access, reproducibility, consistency, transparency, and interoperability of NFDI4Microbiota services. In order to establish standards that are also accepted by the research community, these are discussed in close exchange with several stakeholders.





COMMUNITY BUILDING

NFDI4Microbiota will strengthen the community by supporting the exchange between potential research partners as well as by supplying support infrastructure. The community is not only central to the active research process but also to the development of infrastructure and standards as well as their acceptance and use. There are different channels NFDI4Microbiota establishes e.g. active participation in important conferences, mailing lists, social media presence and interaction with DFG-funded research consortia. Furthermore, the consortium will continue to investigate and actively promote synergies with existing and future NFDI consortia, develop common use cases and organize joint events, among other things by using the connections of the applicant institutions that are already part of other NFDI consortia. International networking is particularly important with regard to the development of standards.

CONCLUSION & OUTLOOK

The NFDI4Microbiome consortium is engaged in the support of microbiome research and for this builds on the high-performance cloud infrastructure provided by de.NBI. The de.NBI computational infrastructure supplies the network with computational resources on the one hand and massive storage capacities on the other. Furthermore, de.NBI and other partners work together to develop best practices and standards for metadata, training programs and community building. In conclusion, the de.NBI network represents a key component in the provision of NFDI4Microbiota's solution and the works of de.NBI in the recent years lay the foundation for the efficient implementation of several NFDI consortia including NFDI4Microbiota.

REFERENCES: [1] Sci Data 2016;3:160018. DOI: https://doi.org/10.1038/sdata.2016.18.

AUTHORS: Barbara Götz[1], Kristin Sauerland[2], Eva Seidlmayer[2], Justine Vandendorpe[1], Alice McHardy[2], Konrad U. Förstner[1]
[1] Deutsche Zentralbibliothek für Medizin (ZB MED) - Informationszentrum Lebenswissenschaften, Gleueler Straße 60, 50931 Köln
[2] Helmholtz-Zentrum für Infektionsforschung GmbH (Helmholtz Centre for Infection Research), Inhoffenstraße 7, 38124 Braunschweig

MANAGING RESEARCH AND RESEARCH DATA IN THE de.NBI CLOUD

How the de.NBI Cloud infrastructure is used to make research data FAIR. An overview over the BioDATEN project.

The state of Baden-Württemberg supports four Science Data Centers (SDC) with the intention to create and establish services and tools helping specific scientific communities to make their research data FAIR. BioDATEN is one of the sponsored SDCs and aims to support the bioinformatics community in Baden-Württemberg by providing the necessary infrastructure. FAIR data and research data management gets increasingly important and is expected by funding agencies but also reflects the good scientific practice and is also proof of citeable scientific activities. To develop a reliable and sustainable infrastructure, BioDATEN builds on the de.NBI Cloud, both for storage and for computation.

Striking advances in high-throughput technologies have enabled researchers around the globe to analyse nowadays whole organisms or even entire ecosystems. However, with the advance of these technologies the computational demand for computation and storage has vastly increased which impacts all steps from data generation, data analysis, data sharing to data publishing. Currently, the amount of data and demand for computational resources have exceeded by far the capacities of most biological labs which often requires expertise in data analysis and data management.

In the context of growing data, well-annotated research data becomes more and more important as it improves reproducibility and standardization in research. Curated metadata facilitates and simplifies the identification of comparable research projects. However, until today accurate metadata information is often lacking. Important efforts have been made in order to standardize the metadata collection. One example is the Minimum Information about any Sequence (MIxS) schema developed by the Genomics Standard Consortium (GSC) which is an integral component of sequence data submission at SRA or ENA. Unfortunately, this only addresses the metadata collection during the sequence submission which almost always is carried out immediately before manuscript submission. A consistent metadata collection workflow that includes – besides sequence data – also reproducible workflows and intermediate results is still lacking. One step toward a sustainable research data management along the research data life cycle is the implementation of the FAIR principles that aim to make research data findable, accessible, interoperable and reusable. While the FAIR principles formulate the target state for research data the actual reality still shows a considerable gap which needs to be bridged.

In order to fill this gap, the BioDATEN project (https://portal.biodaten.info/) is sponsored by the Ministry of Science, Research and Arts of the state of Baden-Württemberg to support the state's bioinformatic community and will provide a uniform research platform which includes computation, storage and data publishing resources and thus, supports the researchers within their whole research data management cycle. In this way, BioDATEN helps researchers to make their data FAIR.

WHAT IS FAIR DATA?

FAIR data implies that data is managed and curated according to the four FAIR key concepts. (i) To be findable, a globally unique persistent identifier needs to be assigned to (meta)data. Furthermore, data must be described with rich metadata, metadata must include an identifier that leads to the data and (meta) data must be indexed in a searchable resource. (ii) To be accessible, (meta)data have to be retrievable using their identifiers via standardised communication protocols. Even if the data are no longer available, the metadata has to remain accessible. (iii) To be interoperable, (meta)data must use a formal, accessible, shared, and broadly applicable language for representation. (Meta)data must use vocabularies including qualified references to other (meta)data. (iv) To be reusable, (meta)data must be richly described with relevant information such as clear and accessible data usage license, detailed provenance, and domain-relevant community standards. The advantages of FAIR data lie in the reuse of data as a possible starting point to answer further scientific questions, e.g. by comparing data sets and the display of good scientific practice. This implies, however, that truly FAIR data need an environment that supports and facilitates the necessary annotation with metadata.

WHAT IS THE RESEARCH DATA LIFE CYCLE?

The research data life cycle in life sciences commonly includes the following stages: data generation including sample collection and preparation, data processing and analysis, and data publication. Within each stage of the life cycle different metadata are produced and all of them need to be FAIR. The first stage requires metadata about what and how sample preparation and sequencing took place. As mentioned earlier, there are schemata like MIxS to support researchers. The second stage involves information about computational tools, pipelines, settings, etc. The third stage involves metadata about the creators of the data and granted licenses (see Figure 1).

The stages mentioned here do not represent the complete cycle but only represent major stages. As research is not a straight line from data to publication, not every required bit of metadata can be recorded beforehand. Researchers might change their affiliation, new tools and workflows can be released, and the target repository might change its metadata requirements. While metadata of living research data are dynamic and require flexibility, the publication at the end of the life cycle freezes them and renders them static. The idea of data publication is to get a citable persistent identifier – such as a DOI - to facilitate data exchange and to display scientific activities beside journal publication in the form of a data publication.

BioDATEN INFRASTRUCTURE

To implement the FAIR principles and to provide services for the bioinformatic community in Baden-Württemberg, BioDATEN sets up a distributed infrastructure consisting of several components operated on de.NBI Cloud resources. Each component targets a different stage at the research data life cycle (see Figure 2).

The central component is the BioDATEN science gateway which allows scientific project management and knowledge sharing. This science gateway is extensible with portlets that can introduce further functionalities. Hence, the science gateway serves as a central entry point that grants access to further functionality. One example for such a portlet provides access to the de.NBI Cloud S3 storage. Research data publication is managed by InvenioRDM. InvenioRDM registers DOIs as persistent identifiers for research datasets. Thus, data publication is taken care of while DOIs increase the accessibility of research data and serve as citable proof of research activity. The whole process of (meta)data handling is geared to get publication-ready (meta)data towards the end of the research data life cycle in the most convenient way for the researchers. They can search for research data of interest within the science gateway by using the open-source and community supported search frontend VuFind which is based on the search server Apache Solr. In order to facilitate single sign-on, Keycloak is used as a one-stop-shop for authentication and is therefore connected to the Life Science Login to enable easy access to services and resources.

As mentioned before, the metadata annotation is crucial for data to be FAIR. In order to achieve that goal, the science gateway will also be connected to the BinAC bwHPC infrastructure next to the de.NBI Cloud infrastructure. As soon as a computation job is completed on the BinAC cluster, a core set of workflow-related metadata is automatically created, attached to the research dataset and moved to a staging area on BinAC infrastructure. This core set of metadata focuses on finable and accessible FAIR principles by integrating the DataCite schema. This schema is required for DOI registration and includes also a set of common terms which are widely used by research data portals such as species ID, tissue type, and experiment type. Users will be able to annotate and complete the metadata core set via the BioDATEN science gateway by selecting a domain-specific respective target metadata schema such as MIxS and PRIDE. This directly aims at fulfilling the reusable principle while the provided metadata framework aims at the interoperable principle by using standard formats and vocabularies. By annotating and enriching their research data with metadata from the beginning in the research data management cycle, they can make their data not only publication-ready during the entire research process, but can also keep track of their data, workflows and analyses.

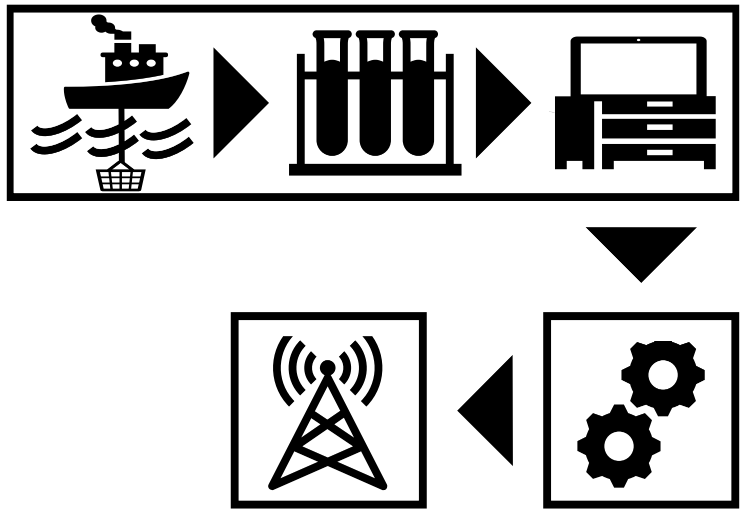
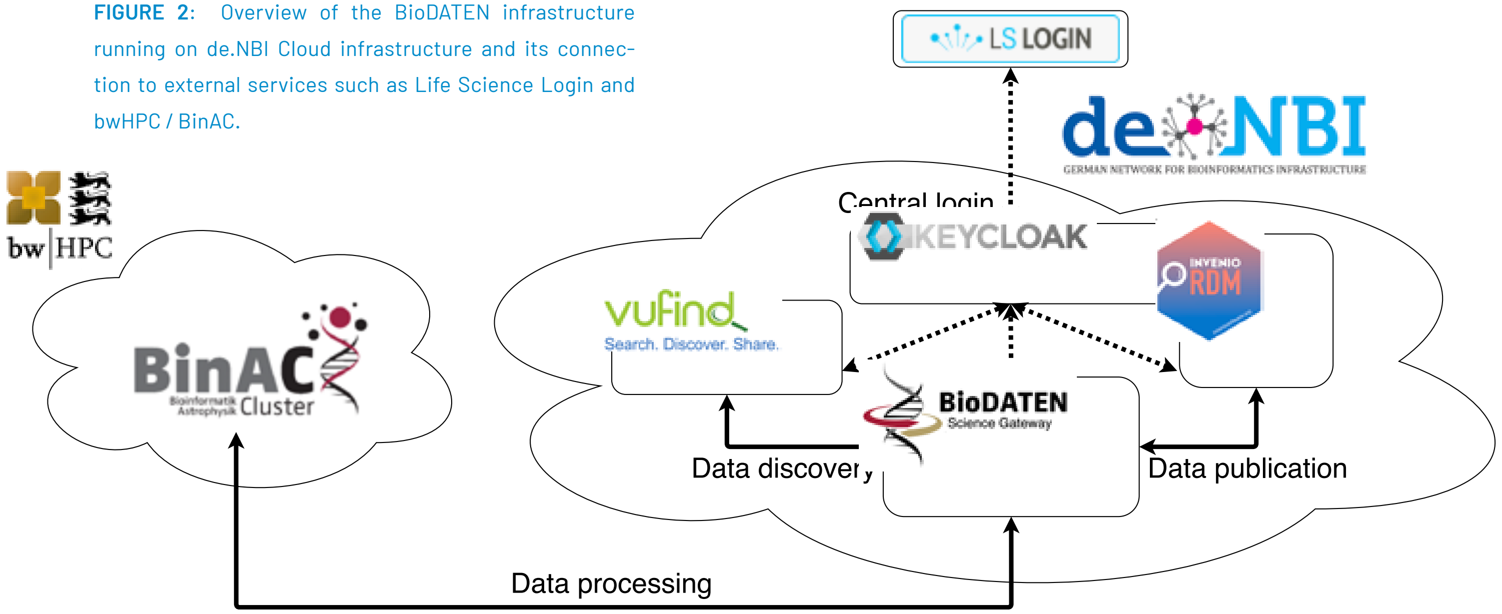


FIGURE 1: Three stages of the research data life cycle in life sciences. 1st stage: Sample gathering, preparation, and data generation. 2nd stage: Data analysis and processing. 3rd stage: Data publication.

FIGURE 2: Overview of the BioDATEN infrastructure running on de.NBI Cloud infrastructure and its connection to external services such as Life Science Login and bwHPC / BinAC.



CONCLUSION & OUTLOOK

FAIR research data requires not only engagement from users and service providers alike but also an integrated and combined view on infrastructure and its components. The de.NBI Cloud infrastructure enables BioDATEN to provide such an environment and to integrate all components under one roof which in turn facilitates further component integration. During the ongoing BioDATEN project, more portlets for the science gateway will be developed to further integrate de.NBI Cloud capabilities and to help researchers to make their data FAIR. In turn, the aim is to generate tangible added value for science, politics and society.

AUTHORS: Jens Krüger[1], Johannes Werner[1] and Holger Gauza[1]
[1] University of Tübingen, High Performance and Cloud Computing Group ZDV, Wächterstraße 76, 72074 Tübingen

EMBEDDING THE de.NBI CLOUD IN EUROPEAN ACTIVITIES – **BECAUSE RESEARCH DOESN’T STOP AT THE REGIONAL BORDERS OF EUROPE.**

In the new era of Big Data, toward which European life sciences are rapidly transitioning, cloud technology will become more important to the sector. However, cloud computing is not alone what emerges, but a networked system that links many cloud federations providers together.

EOSC-LIFE

de.NBI Cloud contributes to European FAIR data initiatives.

EOSC-Life[1] is a consortium of 13 Life Science Research institutes across Europe. The goal of EOSC-Life is to 'create an open, digital and collaborative space for biological and medical research', based on the principles of FAIR data, and a service catalog that allows researchers to manage, store, and use data within the European Open Science Cloud. The authors are involved in EOSC-Life with Harald Wagener directly as Lead of WP7 (Cloud Services); de.NBI Cloud involvement goes beyond that and fulfills the goals of cloud training and infrastructure provisioning on a European scale via EOSC-Life.

ABOUT EOSC AND EOSC-LIFE

The European Open Science Cloud is a European partnership to drive Open Science practices and skills as the common standard of practice, make research results findable, accessible and reusable, and build sustainable and federated infrastructures to enable open sharing of federal results [1].

In this context, EOSC-Life provides policies, guidelines, and processes for secure and ethical data re-use; populating an ecosystem of life-science tools in EOSC, and enabling data-driven research in Europe, connecting life scientists to EOSC via Open Calls [2].

These goals and ideals align closely with the approach of the de.NBI Cloud as a federated research infrastructure for life sciences, and in this article we describe how de.NBI Cloud support fulfills some of the goals above today.

EOSC-LIFE TRAINING AND BOOTSTRAPPING

de.NBI Cloud is supporting and driving the deliverables of various work packages. There is a natural connection between WP2 (Tools and Workflows) and WP7 (Cloud Services), but with the focus on life sciences, there also is a tight interaction between WP4 (Sensitive Data) and Cloud Services. One of the deliverables of WP7 was a questionnaire for infrastructure providers about their capabilities to host sensitive data; several de.NBI nodes graciously provided feedback on this questionnaire and filled it.

Based on input from WP2, de.NBI Cloud representatives in WP7 also provide training so that scientists can make effective use of state-of-the-art cloud technologies during their projects. This aligns with de.NBI Cloud's core mission and we provided tailor-made training such as basic and advanced kubernetes courses for container workload management. On top of that, de.NBI Cloud user meetings are open to EOSC-Life participants; this covers a large gamut of cloud technologies from basic OpenStack to life science analytics pipeline usage.

[1] https://www.eosc-life.eu/

EOSC-LIFE CALLS AND SUPPORT

WP3 (Use Cases) profits most directly from the de.NBI Cloud in terms of consultation for applicants to the various calls, as well as adjudicating submissions for technical feasibility and maturity to help with the selection of funded calls.

For all three calls, de.NBI representatives consulted on more than 30 submissions, reviewed more than 50 submissions for technical feasibility and provided consultancy to the applicants. de.NBI Cloud members were involved in a handful of submissions to these calls as well.

WP9 (Training) also had two open calls where the de.NBI Cloud provided resources and personnel for training.

EOSC-LIFE CALLS AND de.NBI CLOUD CONTRIBUTIONS

de.NBI Cloud directly or indirectly provides resources to projects; directly via the de.NBI Cloud portal and onboarding process (after going through the EOSC-Life Resource Allocation Process), or indirectly since a lot of use cases use Galaxy which runs on the infrastructure of multiple de.NBI Cloud sites.

While the third WP3 call is still in the evaluation and selection phase, de.NBI Cloud provides resources for three out of eight selected projects from the first open call (three via Galaxy, one via cloud infrastructure); and two from the sensitive data call (Reference Data Source, Expression Atlas).

For the first WP9 Open Training Call, de.NBI Cloud provides resources for one out of four selected projects, FATES [3], and for the second WP9 Open Training Call, we provided both Admin training and resources for Training Infrastructure as a Service, which allows smoother training sessions by giving trainees' workflows preference for processing [4].

CONCLUSION

While de.NBI Cloud's primary mission is to support Life Science research in Germany with the required training and infrastructure to appropriate and use cloud technologies successfully to accelerate and enable new methods of data-driven and analytics workflows, the many interactions with EOSC-Life via the ELIXIR umbrella and direct involvement in the EOSC-Life project show that the present and future of Life Science research is European at its core. In this vein, it's not surprising that EOSC is aligning with other European efforts such as GAIA-X which is mainly industry-driven, but has strong ties to research in some domains, especially in health. The de.NBI Cloud is a valuable and valued partner in European efforts aside from other research infrastructures in Europe.

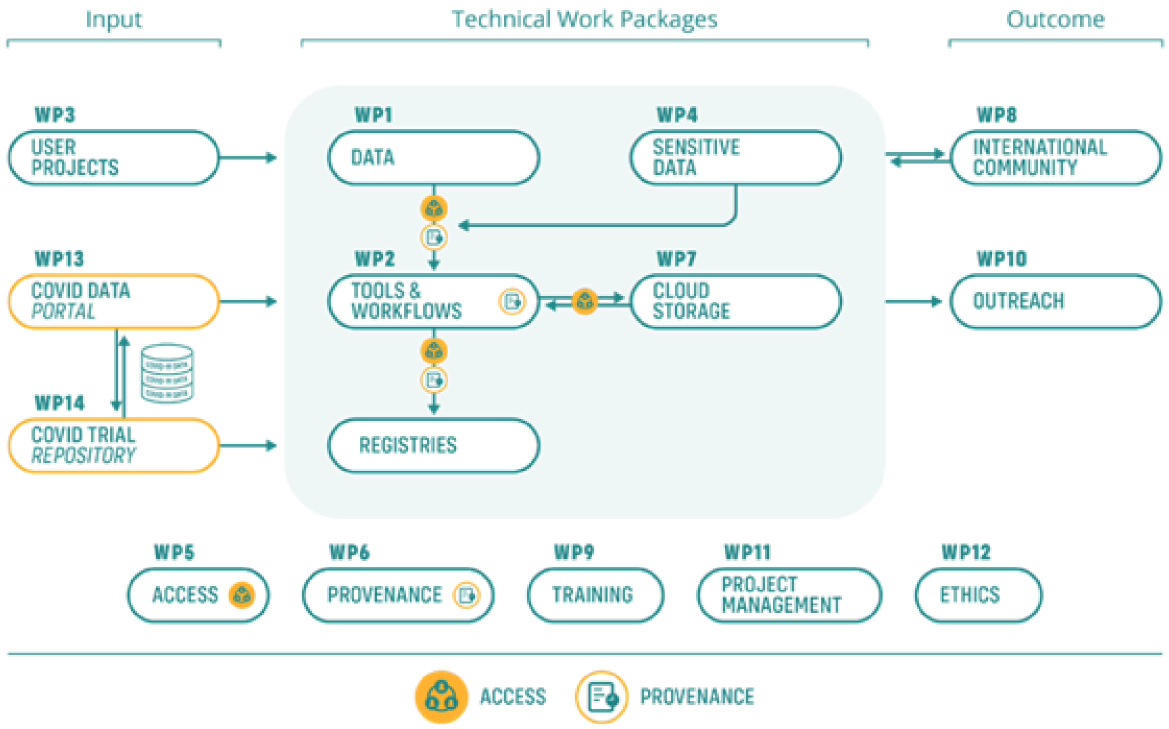


FIGURE 1: Overview of workpackages in EOSC-Life.

REFERENCES: **[1]** EOSC Partnership proposal: https://ec.europa.eu/info/sites/default/files/research_and_innovation/funding/documents/ec_rtd_he-partnership-open-science-cloud-eosc.pdf **[2]** EOSC-Life Homepage: https://www.eosc-life.eu/ **[3]** FATES technical documentation: https://fates-docs.readthedocs.io/en/latest/index.html **[4]** TIaaS reference: https://galaxyproject.eu/posts/2021/08/24/tiaas-flyer/

AUTHORS: Ivo Buchhalter[1], Alexander Goesmann[2], Björn Grüning[3], Jens Krüger[4], Alexander Sczyrba[5] and Harald Wagener[6]

[1] German Cancer Research Center (DKFZ), Im Neuenheimer Feld 280 · 69120 Heidelberg.
[2] Justus-Liebig-Universität Gießen, Bioinformatics & System Biology, Heinrich-Buff-Ring 58, 35392 Gießen
[3] University of Freiburg, Department of Computer Science, Georges-Köhler-Allee 106, 79110 Freiburg
[4] University of Tübingen, High Performance and Cloud Computing Group, IT Center, Wechterstraße 76, 72074 Tübingen
[5] Bielefeld University, Faculty of Technology, Universitätsstr. 25, 33615 Bielefeld, Germany
[6] Center for Digital Health, BIH, Charité Universitätsmedizin, Charitéplatz 1, 10117 Berlin

EXCELLENCE IN MACHINE LEARNING INSPIRED BY THE de.NBI CLOUD

The Machine Learning Cloud Tübingen

The rise of 'intelligent' technology is transforming industry, economy, medicine and society at an unprecedented scale. At the core of this revolution are breakthroughs in the field of machine learning. Developments in machine learning have the potential to transform science at an equally fundamental level. The aim of the cluster of Excellence 'Machine Learning: New Perspectives for Science' is to enable machine learning to take a central role in all aspects of scientific discovery and to understand how such a transformation will impact the scientific approach as a whole. The technical platform for this endeavor, the Machine Learning Cloud Tübingen, has been inspired by the de.NBI Cloud.

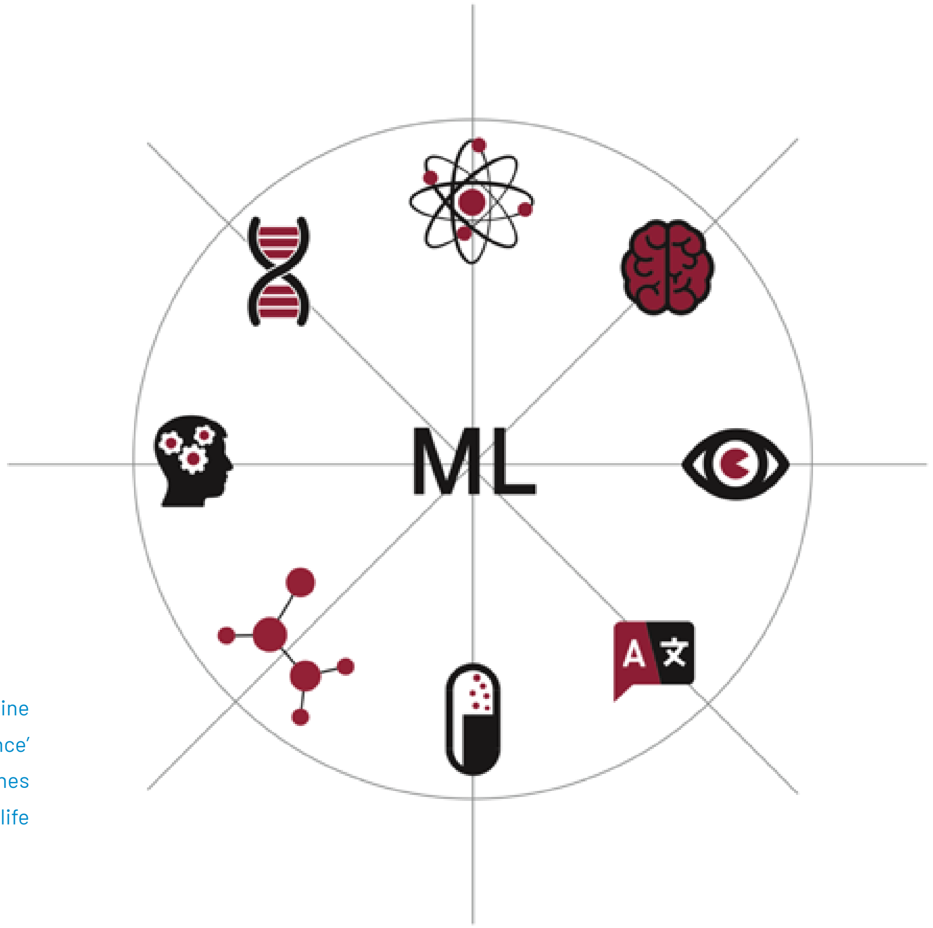
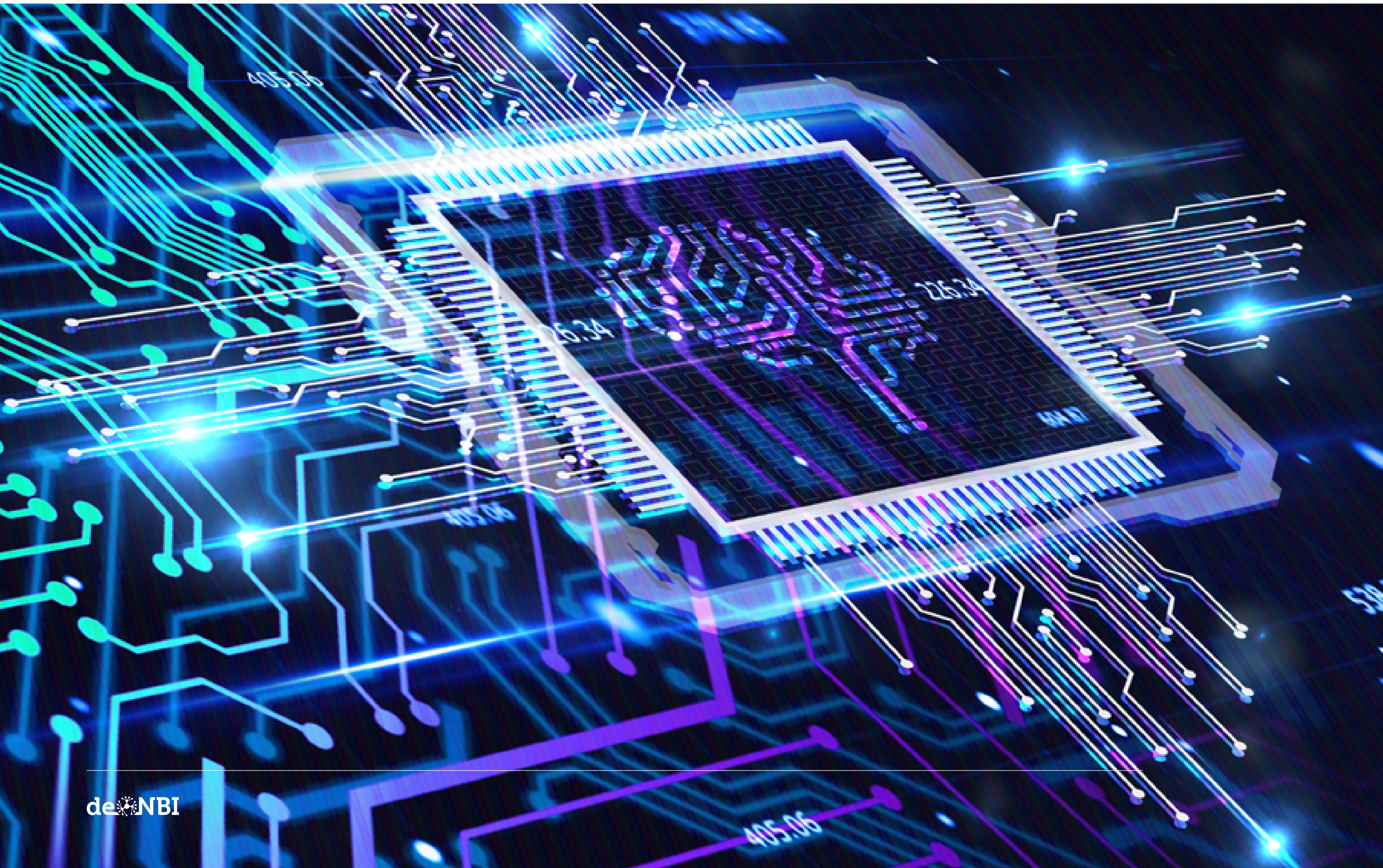


FIGURE 1 The Cluster of Excellence 'Machine Learning: New Perspectives for Science' covers a broad range of scientific disciplines ranging from computer sciences over life sciences to philosophy and ethics.



In recent years, 'intelligent' technology has been transforming engineering, industry, and the economy at an ever-increasing pace. At the core of this revolution are breakthroughs in the field of machine learning which allow machines to perform tasks that, until recently, could only be performed by humans. Most notably, advanced algorithms now allow computers to drive cars or beat humans in challenging strategy games such as Go [1].

Less visible, but at least as important, is the ongoing transformation of science through machine learning. This is driven partially by the huge data sets generated by large-scale imaging technologies and high-throughput sequencing in biomedical research and neuroscience, enormous text corpora in linguistics, all-sky surveys in physics, earth observations in geoscience or social media in social science. To leverage this flood of data for science, the challenge is not just to process these data sets or to make predictions based on them, but rather to match the process of inductive reasoning to the rapidly expanding opportunities for data generation.

In many areas of science, machine learning is already used to solve basic scientific prediction tasks, but recent breakthroughs open up an enticing new perspective: Automated inference might take a more central role in scientific discovery itself. In some disciplines, we can see the first tentative evidence for this taking place as for example in the life sciences.

Today, machine learning algorithms can fit highly complex models with millions of variables and complex dependencies to scientific data that otherwise would be impenetrable even to human experts. These new opportunities empower scientists but also present new challenges: Researchers now have to design the machine's model, understand its inherent assumptions, and monitor it for flaws. This requires new skills and techniques to distill theoretical insight from the machine's output and to relate it to previously known theories or conjectures. Where scientists used to generate data and interact with it, they now increasingly interact with algorithms.

THE CLUSTER OF EXCELLENCE

In the Cluster of Excellence, which started in January 2019, we target the following four research areas:
In Research Area A **Beyond prediction, towards understanding** we design algorithms that reveal complex structure and causal relationships from data in order to integrate machine learning into the scientific discovery process.

Research Area B **Managing uncertainty** deals with the development of tools to estimate and handle the uncertainty in data-driven scientific models and algorithms, and exploit this information for experimental design.

Research Area C **Interface between algorithms and scientists** is designed for the development of techniques to allow scientists to understand and control all stages of the machine learning process in the scientific discovery pipeline.

Research Area D **Philosophy and ethics** of machine learning in science is based on the assumption that machine learning algorithms will play a central role in the whole process of scientific discovery, which challenges our traditional understanding of the scientific process and raises fundamental questions about concepts of scientific discovery and the role of the scientists. We tackle these questions from the perspective of philosophy and ethics of science.

Our team of principal investigators consists of researchers in machine learning and its applications in various disciplines, including medicine, neuroscience, bioinformatics, vision, cognitive science, physics, geoscience, linguistics and social science, as well as experts in philosophy and ethics. Our cluster builds on the internationally renowned strength of Tübingen as a hub for machine learning as well as on the established excellence in the contributing scientific fields.

THE MACHINE LEARNING CLOUD TÜBINGEN

Establishing a productive and innovative machine learning research community requires substantial computing resources with appropriate management. Therefore, we established the 'Machine Learning Science Cloud Tübingen' in collaboration with the Center for Data Services (ZDV) of Tübingen, the Tübingen AI Center and the University. This instrument consists of a cloud environment with high-performance computing power with large-scale GPU and CPU resources as well as data storage. The layout and architecture are adapted from the de.NBI Cloud [2]. Compared to the individual management of compute resources, the facility drastically increases the peak resources available to each user, facilitates the development of standardized software tools to deploy and scale machine learning experiments, and free users from managing and administering their own hardware. Further, the cloud allows for the appropriate handling of sensitive data, for example in a medical context, due to the expertise in data handling gained within de.NBI. This way, we ourselves could design the system most useful for scientific use and modify the system for particular questions or structural needs.

This state-of-the-art cloud is comprised of several core services, hypervisors, network and storage, with a strong focus on GPU resources. The instrument was established in a server room in a new building in the 'Technologiepark Tübingen' (AI Research Building, Maria von Linden Str. 6, Tübingen), which is now also the home for all new research groups and professors established within the cluster of excellence. The 'Machine Learning Science Cloud Tübingen' supports open science best practices and advice and allows the cluster members to perform reproducible machine learning experiments. The cloud is designed to be used for both, prototyping as well as large-scale experiments and guarantees a seamless transition between the two.

FIGURE 2 The Machine Learning Cloud Tübingen provides a state-of-the-art environment, including compute and storage resources for the Machine Learning community.

It is probably needless to repeat that the construction of the 'Machine Learning Science Cloud' has heavily benefited from the experiences that were generated through the establishment of the de.NBI Cloud in Tübingen. The performance and the usage of the system were set-up in a similar way as compared with the de.NBI Cloud. As a matter of course, scientific computing in the 'Machine Learning Science Cloud Tübingen' is far away from routine operation and the whole system is in a continuous adaptation process to the most recent developments within the scientific machine learning community.

CONCLUSION

The Cluster of Excellence 'Machine Learning: New Perspectives for Science' assembles some of the worlds leading talents in machine learning. The infrastructure supporting their research activities is made available through the 'Machine Learning Science Cloud Tübingen', which has been inspired by the de.NBI Cloud. Beside providing powerful compute and storage resources, it represents a prime example for interdisciplinary collaboration and knowledge transfer among major research projects in Germany.

REFERENCES: [1] Nature 2016;529:484-489. DOI: https://doi.org/10.1038. **[2]** F1000 2019; DOI: https://doi.org/10.12688/f1000research.19013.1. **[3]** Proceedings of the bwHPC Symposium 2018, Freiburg, p 201-215, DOI: https://doi.org/10.15496/publikation-29062.

AUTHORS: Tilman Gocht[1] and Jens Krüger[2]
[1] University of Tübingen, Exzellenzcluster 'Maschinelles Lernen', Geschwister-Scholl-Platz, 72074 Tübingen
[2] University of Tübingen, High Performance and Cloud Computing Group, IT Center, Wächterstraße 76, 72074 Tübingen

GAIA-X AND THE NEW OPPORTUNITIES FOR de.NBI CLOUD

de.NBI Cloud on the way to becoming a sustainable player within large infrastructures

Launched in 2019 as a joint French-German initiative, GAIA-X is founded on a strong conviction that Europe should be determining it's own rules for digital transformation. It's core aspects are in accordance with EU data protection regulation to strengthen the data sovereign's rights and realizing economic independence from the dominant US cloud platforms. On this basis, the GAIA-X community developed the GAIA-X framework, embodied in the Policies and Rules and the Architecture of Standards. This framework, together with the Federation Services (currently under development), form the basis of cooperation between more of 300 organizations from almost all EU member states and beyond.

Domain specific data spaces (for Health, Mobility, Smart Living, etc.) are comprised of the sum of their participants. These data spaces can nest and overlap; a data provider for example may participate in several data spaces at the same time. Data sovereignty and trust are essential for their functioning and the relationship between participants. There is a huge opportunity for the de.NBI Cloud to become the premier federated provider for research infrastructures in GAIA-X.

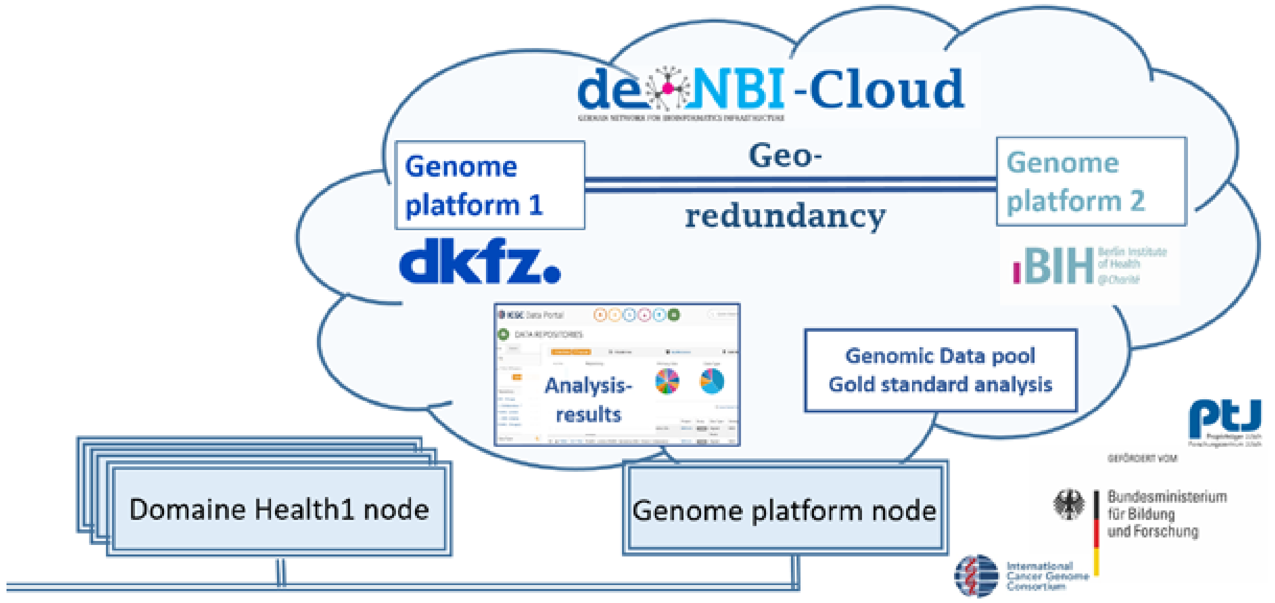


FIGURE 1 GAIA-X Research Platform Genomics: This geo-redundant cloud platform is used to store and analyse genomic data. The openness and resulting flexibility of GAIA-X enables the connection of existing (data) platforms to other research and health domains and international initiatives. For example, easier access to and greater use of the de.NBI cloud is also possible, including in connection with future funding projects based on the GAIA-X architecture. (Source: Christian Lawerenz)

de.NBI CLOUD OPTIONS IN THE LANDSCAPE OF FEDERATED INFRASTRUCTURE FOR RESEARCH AND INDUSTRY

Legally, GAIA-X is a non-profit organisation under Belgian Law (AISBL). This formal entity is complemented by an active community that reflects all participants in the ecosystem: Infrastructure and service providers, users, and various intermediaries. A number of these are ready to put the data sovereignty of a digitally emancipated Europe into practice in the coming years.

GAIA-X is based on an ecosystem of trusted and secure cloud and edge infrastructures in Europe. The focus lies here on federated cloud solutions that enable a competitive market for cloud systems. European cloud users should be free to choose local or international providers. Cloud offerings should also be transparently aligned with documented and certified compliance with GAIA-X policies and standards.

So far, GAIA-X only has commercial providers that focus on servicing small, medium and large enterprises. They are less focused on scientific communities. The portfolio includes well-known European cloud providers such as IONOS, Atos, Deutsche Telekom, OVHcloud, Scaleway, PlusServer, CISPE, 3DS OUTSCALE, who are GAIA-X founding members from the very beginning and are driving the development of GAIA-X. The de.NBI Cloud, on the other hand, is academically oriented and serves a different costumer group. This complements the offerings of cloud providers for industry with those of de.NBI, which are addressed to scientists. The close integration of the de.NBI Cloud in European projects such as ELIXIR, EOSC, EUCANCan and HealthyCloud also provides the obvious opportunity to establish service offerings in GAIA-X as a pioneer for these major European infrastructure initiatives. Also with the involvement of Charité in the role of a stakeholder of the de.NBI Cloud in GAIA-X, there are already coordination meetings between EOSC and GAIA-X to embed the cloud activities of EOSC in GAIA-X.

The de.NBI Cloud is ideally suited to contribute further to GAIA-X cloud activities. From a research and health perspective, the Charité, as a de.NBI Cloud site, has had a strong influence on the technical development of GAIA-X and has brought the requirements and potential solutions for cloud services and federation into eight technical GAIA-X working groups. This collaboration has taken place in the 'Workstream2' Self Description, Federated Catalog, Identity and Access Management, Interoperability, Interconnection / Data Broker, User Interface, Networking Minimum Viable GAIA. As an example, the ELIXIR Authorisation and Authentication (AAI) concept, which is for years fully implemented in the de.NBI Cloud, was presented as a useful model for a comprehensive GAIA-X AAI. A similar AAI concept will now be implemented in GAIA-X.

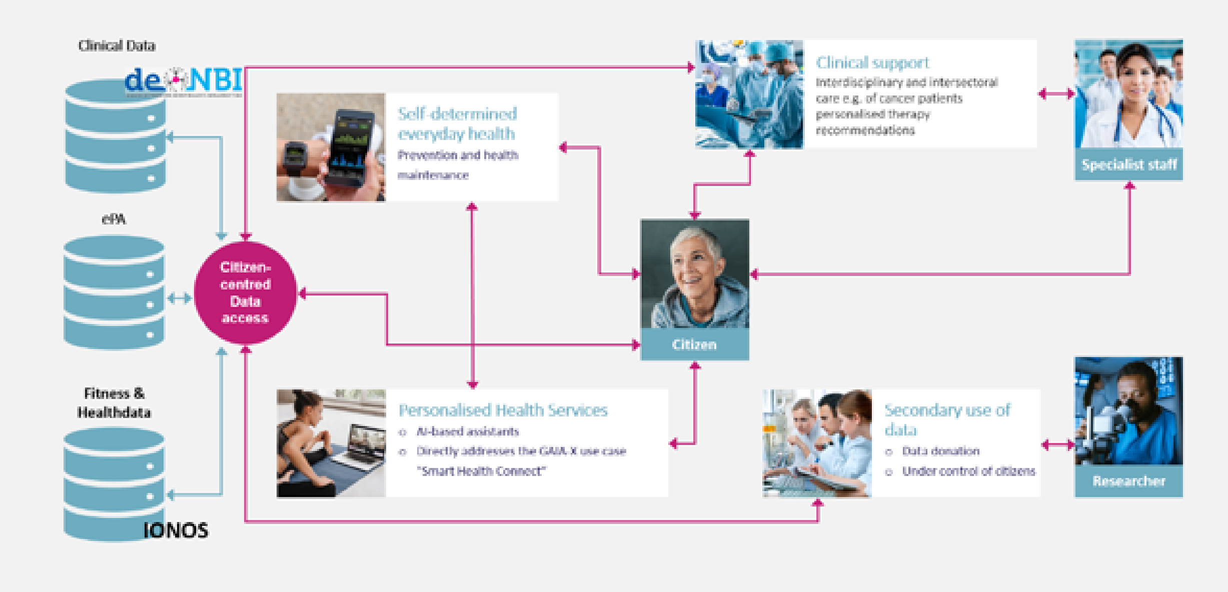

Clinical Data
de.NBI
ePA
Fitness & Healthdata
IONOS
Citizen-centred Data access
Self-determined everyday health
Prevention and health maintenance
Clinical support
Interdisciplinary and intersectoral care e.g. of cancer patients personalised therapy recommendations
Specialist staff
Citizen
Personalised Health Services
o AI-based assistants
o Directly addresses the GAIA-X use case "Smart Health Connect"
Secondary use of data
o Data donation
o Under control of citizens
Researcher

FIGURE 2 GAIA-X HEALTH-X dataLOFT: A central component of the infrastructure are the two cloud solutions de.NBI Cloud and IONOS. The use cases address socially highly relevant questions about strengthening the role of citizens, preventive health care, healthy ageing and clinical care, where the core question of who finances the 'therapy of the healthy' finds new answers in the B2P and B2C models presented. (Source: BIH, LANGE und PFLANZ, shutterstock © Alexander Raths, shutterstock © Andrey Popov, shutterstock © Arsenii Palivoda, shutterstock © ESB Professional, shutterstock © Gorodenkoff, iStock © ljubaphoto)

Without the previous experience of the federated de.NBI Cloud with its focus on research and many more active contributions would not have been possible. Examples include the active participation in 'Workstream1' planning groups of GAIA-X, the creation of various GAIA-X position papers, the development of the architecture requirements, the specification of the Federated Catalogue as an interface between providers and consumers in the GAIA-X ecosystem.

The de.NBI Cloud as a GAIA-X offering for academic users is meanwhile included in the Federated Catalogue through the 'Self Descriptions', the machine-readable and human-readable description of the de.NBI Cloud.

Another result of the positioning of the de.NBI Cloud in GAIA-X is that the 'Research Platform Genomics', a comprehensive platform of human genome data in the field of cancer, operated by the German Cancer Research Center (DKFZ) and the Charité, was validated and accepted as one of the first GAIA-X use cases in the Health domain. (https://www.data-infrastructure.eu/GAIAX/Redaktion/EN/Artikel/UseCases/research-platform-genomics.html).

It is **crucial** to stay **updated with most current information** in the field of medical research and health in order to be able to successfully position and implement innovations and ideas from Germany and Europe. The IT group around Professor Roland Eils, who has been one of the three coordinators of the GAIA-X Health domain from the beginning, has taken a very active role in the field of health. Examples are the contributions to the European Health Data Space, the German Health Hub and the regular meetings of the European GAIA-X Health Hubs. These contributions have also created essential prerequisites for a future national and European embedding of the de.NBI Cloud in the GAIA-X domain Health.

THE FUTURE GAIA-X ROLE OF de.NBI CLOUD IN HEALTH-X DATALOFT

It is important that there are concrete applications available soon. In this way, GAIA-X can prove that it creates real added value for companies and the public sector. The German Federal Ministry for Economic Affairs and Energy has for the first time selected 16 lighthouse projects from a funding competition in which around 130 applicants took part. In the Health-X Dataloft project, led by Charité-Universitätsmedizin Berlin (starting in November 2021 for three years), health data will be integrated into the Legitimate, Open and FederaTed dataLOFT platform and made accessible according to GAIA-X standards. The aim is to develop transparent, cloud-based health applications with leaders in the consumer device market, health IT and healthcare. dataLOFT will create an ecosystem of health applications built on the data sovereignty of citizens and patients as users and donors of health data. This should contribute to a broad acceptance and high economic relevance of dataLOFT. For example, although many people can be tracked via fitness apps, this data usually ends up in the clouds of Google and other commercial providers. In contrast to the usual approach so far, the paradigm shift here is that consumers, as active stakeholders, are put in a position to decide self-determined with whom they want to share their health data. In doing so, citizens are consistently placed at the centre of the action in the development and integrative use of health data from the primary and secondary health market in the Health Data Space.

The de.NBI Cloud will offer an infrastructure with the commercial cloud partner IONOS and cloud service providers as Siemens and PolyPoly to support all necessary infrastructure components and cloud stacks. The de.NBI Cloud will focus on the sharing and processing of data of the first healthcare market, e.g. the clinical specifications of patients.

de.NBI Cloud will enable secure and trustworthy data use with the certification in accordance with 27001/02 and 27017/18, the Confidential Computing Technology (e.g., based on Intel SGX). and new IDS standards to be included along with certifying services in the European Health Data Space. The IDS connectors are the GAIA-X basis of the data exchange in dataLOFT with functions and interfaces for the controlled interaction of actors, data and services. The de.NBI Cloud will thus become an integral part of the European Health Data Space with transparent and regulated data exchange.

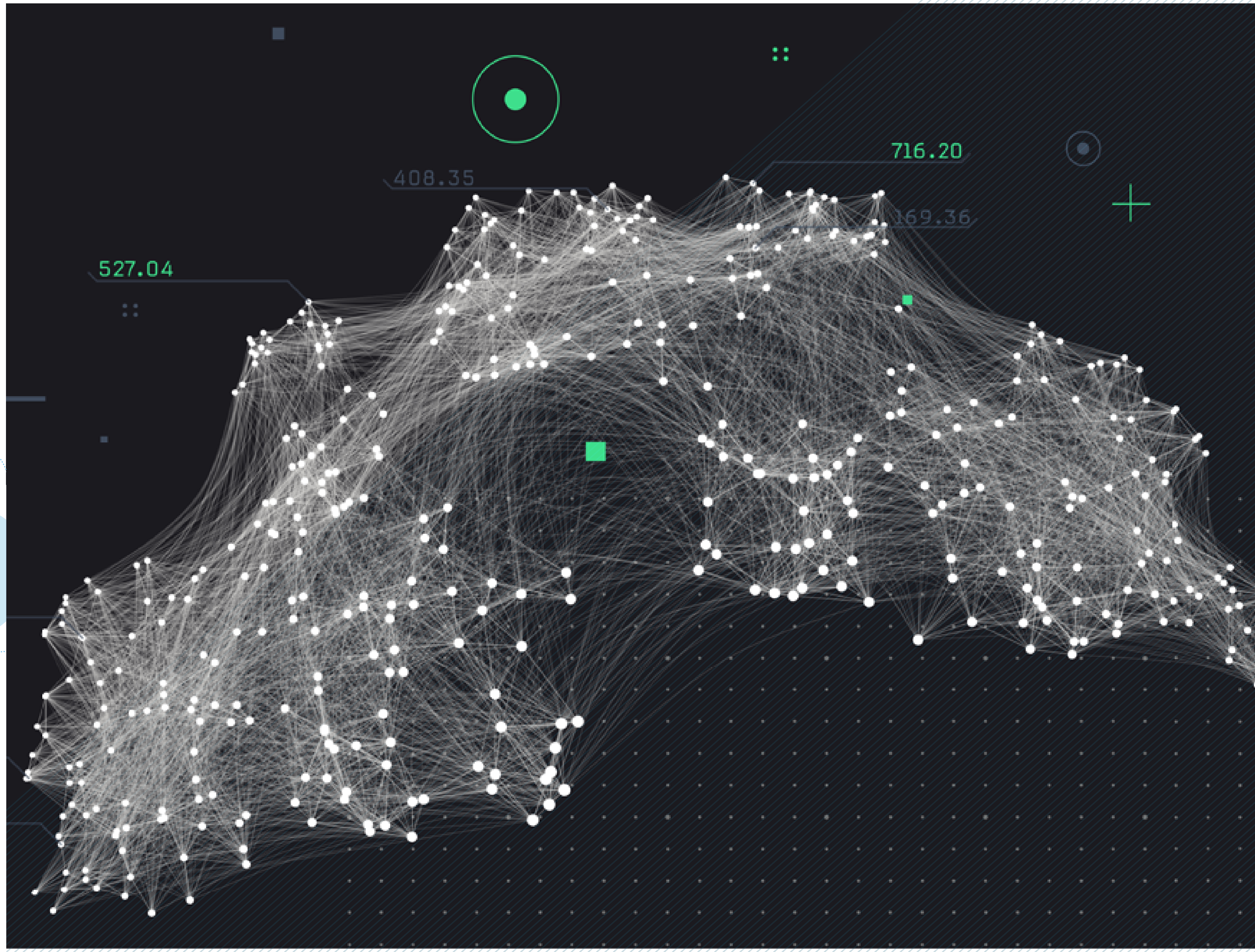
As an economic and health policy priority, sustainable business models in particular will be developed in dataLOFT, with possible extensions of new funding opportunities for the de.NBI Cloud. The implementation of solutions such as dataLOFT with a strong focus on cloud solutions is highly relevant for Europe in terms of both economic and health policy.

CONCLUSION & OUTLOOK

The de.NBI Cloud is the largest academic cloud in Germany. So far, there is no German cloud infrastrucure embedded in Europe that is comparable to and as comprehensive as the de.NBI Cloud. The integration of the de.NBI Cloud with its broad technical and scientific expertise into the GAIA-X ecosystem, which has already begun, and the establishment of the de.NBI Cloud as one of the central cloud solutions in GAIA-X, which is planned for the next few years, opens up the possibility for de.NBI to become a driver of cloud solutions for academic users on an international level. As outlined, essential preparatory work has been done and concrete applications have already been partially implemented.

This role of setting up the de.NBI Cloud in the concrete GAIA-X projects, the essential necessary groundwork for the overall positioning of the de.NBI Cloud in GAIA-X has so far been taken over by Roland Eils' group at Charité. The registered de.NBI association 'de.NBI e.V.', which will then also legally speak for the entire de.NBI Cloud and, like the Charité, could be a member of the AISBL, will be able to make even better use of the diverse possibilities of the de.NBI Cloud as one of the central academic infrastructures in GAIA-X, also in terms of national and European significance and new funding opportunities.

The opening of the de.NBI Cloud to GAIA-X with the participation in the Data Spaces and the development of the multi-layered and federated services are the best prerequisites for holding the central role as a non-commercial cloud provider in Germany and Europe in ten years' time.



AUTHORS: Christian Lawerenz[1], Harald Wagner[2]
[1] *Berlin Institute of Health, Anna-Louisa-Karsch-Str. 2, 10178 Berlin*
[2] *Center for Digital Health, BIH, Charité Universitätsmedizin, Charitéplatz 1, 10117 Berlin*

EU SIMBA PROJECT:

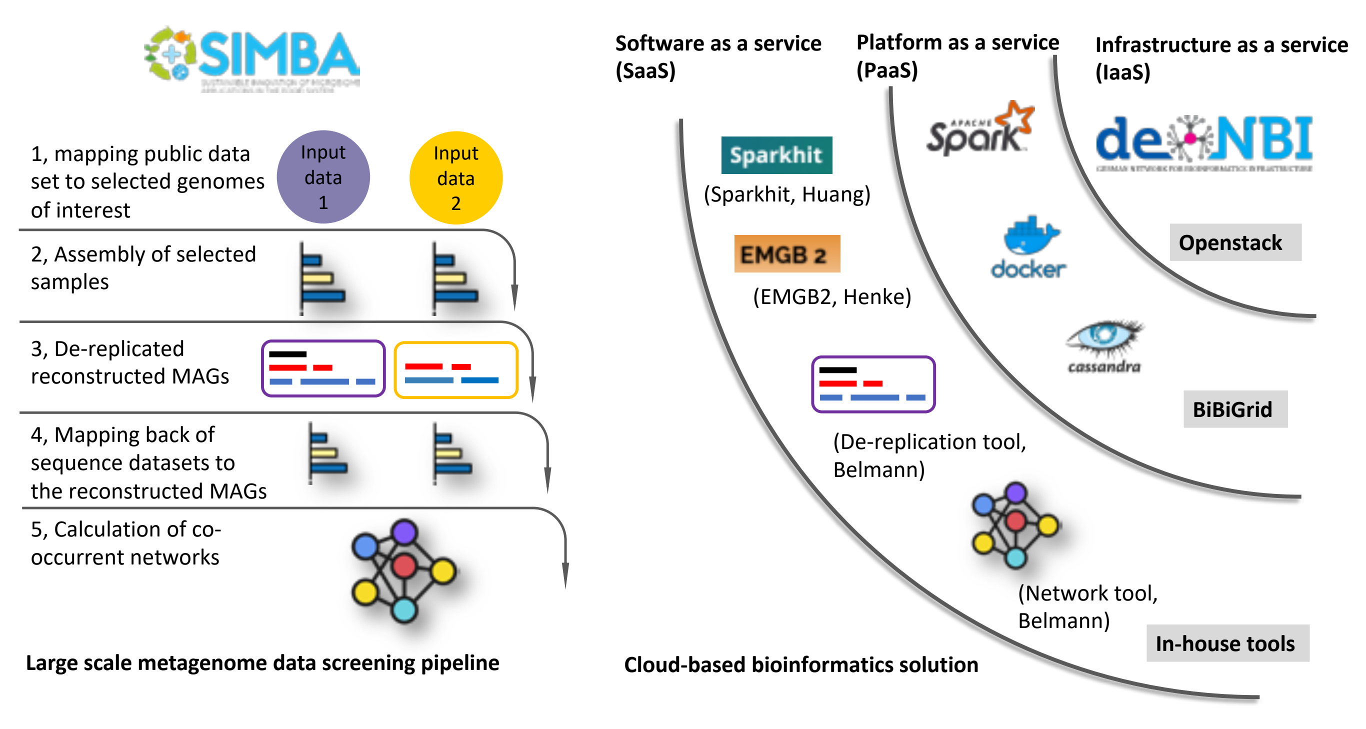
Analyzing large scale metagenomics data on the de.NBI cloud

SIMBA (Sustainable Innovation of Microbiome Applications in the Food System) is a European innovation project, funded under the EU's Horizon 2020 Funding Programme, which provides a holistic and innovative approach to the development of microbial solutions to increase food and nutrition security. SIMBA focuses in particular on the identification of viable land and aquatic microbiomes that can assist in the sustainability of European agro- and aquaculture. BIBI's (Bielefeld Institute for Bioinformatics Infrastructure) contribution to the EU Simba project is to focus on the exploration of microbial communities in large scale publicly available environmental sequencing (metagenomics) data and on association studies with plant growth promoting bacteria (PGPB). To that end, our study addresses both computationally intensive challenges of searching hundreds of terabytes of public data and sophisticated data mining (e.g. network analysis) on putative PGPB genomes.

We established a scalable bioinformatics workflow for detecting PGPB associated microbes from public data. In particular, we have developed a distributable framework, Sparkhit, that enables screening and mapping terabytes of sequencing data within hours. After preliminary screening of large datasets, EMGB (Elastic MetaGenome Browser) is used as a general purpose bioinformatics workflow for analyzing metagenomics data and visualizing annotation results. We also developed a de-replication tool that can handle large amounts of metagenomics samples and facilitate downstream co-occurrence network analysis. Most tools are containerized (e.g. Docker) and are easily accessible on the cloud.

In the case of the EU Simba project, terabytes of public soil metagenome datasets were collected and downloaded on the de.NBI Cloud object storage. Associated metadata containing detailed description of the datasets was categorised. In the first step of our analysis pipeline, Sparkhit is used to map all input sequencing data to the selected PGPB genomes. Sparkhit is an in-house fragment recruitment tool that can be scaled to hundreds of computer nodes. Once high similarity hits are found, corresponding samples are selected for assemblies or co-assemblies (multiple samples in one bio-project) using the EMGB pipeline. The EMGB pipeline also generates 'metagenome assembled genomes' (MAGs) after assembly, representing the microbes present in the samples. To remove redundancy between different samples, the generated MAGs are de-replicated and the representative MAGs are selected for further analysis. To refine our analytical pipeline, we have combined a set of existing tools for the de-replication of MAGs. By comparing and evaluating these tools, we were able to identify the best approach to de-replicate our reconstructed MAGs, and accordingly established a personalized de-replication pipeline.

FIGURE 1: Large scale metagenome data screening pipeline (left) and cloud-based bioinformatics solution (right).


SIMBA
1, mapping public data set to selected genomes of interest
Input data 1
Input data 2
2, Assembly of selected samples
3, De-replicated reconstructed MAGs
4, Mapping back of sequence datasets to the reconstructed MAGs
5, Calculation of co-occurrent networks
Large scale metagenome data screening pipeline
Software as a service (SaaS)
Platform as a service (PaaS)
Infrastructure as a service (IaaS)
Sparkhit
(Sparkhit, Huang)
EMGB 2
(EMGB2, Henke)
(De-replication tool, Belmann)
(Network tool, Belmann)
Spark
docker
cassandra
de.NBI
Openstack
BiBiGrid
In-house tools
Cloud-based bioinformatics solution

Our de-replication pipeline starts by filtering MAGs with high contaminations and low coverages. After the filtering step, Average Nucleotide Identity (ANI) methods are applied to determine species and strain level clusters. Once clusters are formed, representative MAGs are selected based on the ranking algorithm to represent each cluster. Since the core task of MAG dereplication workflows is the estimation of similarity between genomes, which can be done by calculating ANI, we collected and evaluated several ANI-based approaches. Three different datasets (unfiltered, medium, and high MIMAG) from CAMI challenge are used for species and strain level dereplication evaluation.

Once representative MAGs are selected, the pipeline re-maps the sequencing data back to the MAGs and produces MAG-abundance profiles for all samples. The abundance profiles are used to compare the PGPB diversity between different samples. It can also be used to build co-occurrence networks involving assembled MAGs and known PGPBs.

The intermediate results of the EMGB pipeline are imported into the EMGB browser. In the browser, each individual sample can be selected and its computed results can be explored in a click-button style. Users can also compare different samples by selecting multiple samples in the browser tab. Selected metrics, such as the abundance tables of de-replicated MAGs from all metagenome samples, are also accessible through the web interface.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 818431 (SIMBA). This output reflects only the author's view and the Research Executive Agency (REA) cannot be held responsible for any use that may be made of the information contained therein.

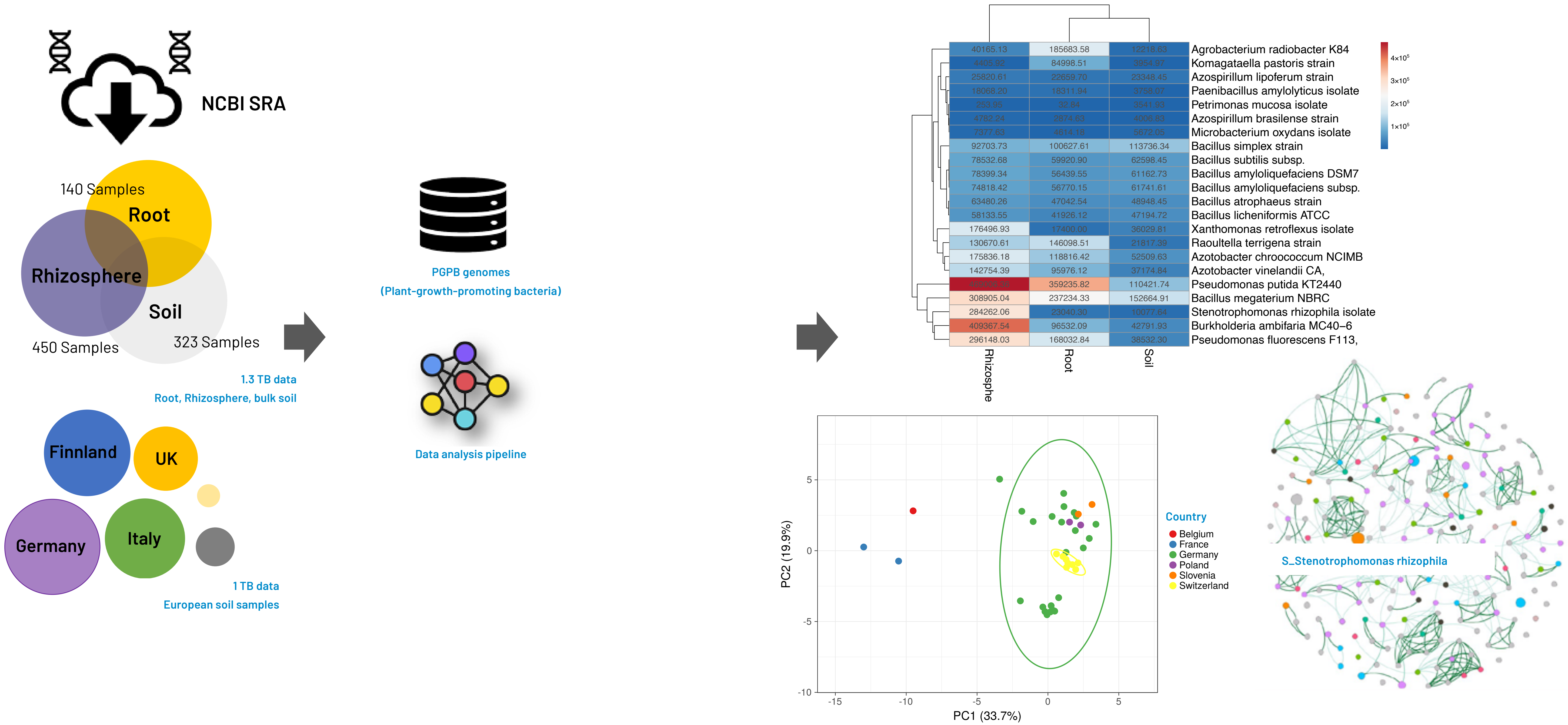

NCBI SRA
140 Samples
Root
Rhizosphere
Soil
450 Samples
323 Samples
1.3 TB data
Root, Rhizosphere, bulk soil
Finnland
UK
Germany
Italy
1 TB data
European soil samples
PGPB genomes
(Plant-growth-promoting bacteria)
Data analysis pipeline
40165.13 | 185683.58 | 12218.63 Agrobacterium radiobacter K84
4405.92 | 84998.51 | 3954.97 Komagataella pastoris strain
25820.61 | 22659.70 | 23348.45 Azospirillum lipoferum strain
18068.20 | 18311.94 | 3758.07 Paenibacillus amylolyticus isolate
253.95 | 32.84 | 3541.93 Petrimonas mucosa isolate
4782.24 | 2874.63 | 4006.83 Azospirillum brasilense strain
7377.63 | 4614.18 | 5672.05 Microbacterium oxydans isolate
92703.73 | 100627.61 | 113736.34 Bacillus simplex strain
78532.68 | 59920.90 | 62598.45 Bacillus subtilis subsp.
78399.34 | 56439.55 | 61162.73 Bacillus amyloliquefaciens DSM7
74818.42 | 56770.15 | 61741.61 Bacillus amyloliquefaciens subsp.
63480.26 | 47042.54 | 48948.45 Bacillus atrophaeus strain
58133.55 | 41926.12 | 47194.72 Bacillus licheniformis ATCC
176496.93 | 17400.00 | 36029.81 Xanthomonas retroflexus isolate
130670.61 | 146098.51 | 21817.39 Raoultella terrigena strain
175836.18 | 118816.42 | 52509.63 Azotobacter chroococcum NCIMB
142754.39 | 95976.12 | 37174.84 Azotobacter vinelandii CA,
469006.36 | 359235.82 | 110421.74 Pseudomonas putida KT2440
308905.04 | 237234.33 | 152664.91 Bacillus megaterium NBRC
284262.06 | 23040.30 | 10077.64 Stenotrophomonas rhizophila isolate
409367.54 | 96532.09 | 42791.93 Burkholderia ambifaria MC40–6
296148.03 | 168032.84 | 38532.30 Pseudomonas fluorescens F113,
Rhizosphe
Root
Soil
4×10⁵
3×10⁵
2×10⁵
1×10⁵
PC2 (19.9%)
5
0
-5
-10
-15
-10
-5
0
5
PC1 (33.7%)
Country
Belgium
France
Germany
Poland
Slovenia
Switzerland
S_Stenotrophomonas rhizophila

FIGURE 2: An overview of two use cases on the bioinformatics platform.

REFERENCES: **[1]** Bioinformatics 2018, DOI: https://doi.org/10.1093/bioinformatics/btx808 **[2]** Nature Methods 2017, DOI: https://doi.org/10.1038/nmeth.4458.

AUTHOR: Liren Huang[1]

[1] *Bielefeld University, Faculty of Technology, Universitätsstrasse 25, 33615 Bielefeld*

HEALTHYCLOUD

Defining the Strategic Agenda for the EU Health and Innovation Cloud

HEALTHYCLOUD
Health Research & Innovation Cloud

Healthy Cloud is a European consortium paving the way for an effective health-related data sharing across Europe. The goal of the consortium is to help conform the future European Health Data Space by defining the Strategic Agendenda for the European Health Research and Innovation Cloud (HRIC). In order to do so, HealthyCloud brings together experts from 21 different institutes and consortia from all over Europe. Together, they are working towards the HRIC strategy, by taking into account ethical and legal frameworks, as well as technological strategies. The de.NBI Cloud is an active partner in the consortium contributing its broad technical experience in cloud computing, which includes work package leadership for defining future cloud computational solutions.

BACKGROUND HEALTHYCLOUD

Advances in health and biomedical sciences require health research to be conducted in a timely and efficient manner [1]. To meet this need and maximize the impact of health research, the establishment of best practices for health data management is crucial. Accordingly, one of the European Commission's 2019-2025 priorities is the creation of a European Health Data Space (EHDS), to allow the sharing of health research data, which currently is present largely in a siloed way throughout Europe, including in Germany, which is considered inhibitive for the advancement of research that is of public interest.

The EHDS will serve as a vehicle for improving health research and its translation into healthcare at all levels: from public health to personalized medicine, through enabling the responsible secondary use of health data in research. In this context, HealthyCloud's mission is to establish a strategic agenda for the implementation of the European Cloud for Health Research and Innovation (HRIC), one of the cornerstones of the EHDS.

The strategic agenda will incorporate consolidated feedback from a wide range of stakeholders, including the European Commission, member states, and regional, national, European, and international relevant initiatives. These stakeholders will be invited to participate in the HealthyCloud Stakeholder Forum, designed to facilitate the dialogue among them and the consortium, and serve as an umbrella to bring together similar efforts in specific areas.

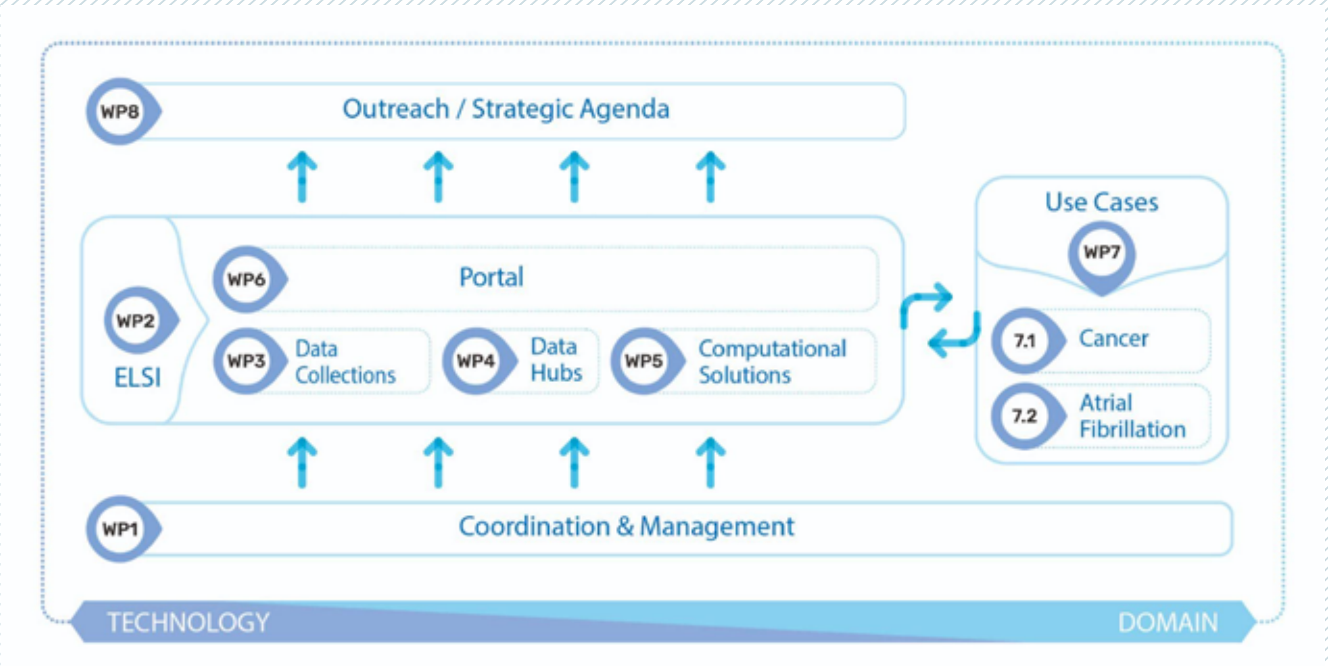


FIGURE 1: Overview of the HealthyCloud work packages (WP). de.NBI Cloud is co-leading WP5 and integrated in WP7.

HealthyCloud is organized around four fundamental objectives, which include:

1. interactions with stakeholders to ensure their voices are incorporated into the strategic agenda
2. ethical, legal, and societal issues in the design of the future HRIC ecosystem
3. sustainable access, use, and re-use of health-related data (from a diversity of health-research communities) available in various data collections distributed in data hubs, taking into account a progressive adoption of FAIR principles
4. technological solutions in the form of computing facilities and mechanisms enabling distributed analysis of health data across Europe

The project is driven by two real-world use cases on cancer and atrial fibrillation, which will ensure that propositions by domain-specific and technological experts are technically and ethically sound and legally compliant. The ultimate goal is to create an ecosystem that builds and reinforces the trust of patients and citizens in the use of their health data for research actionable through a portal that serves as an interface to interact with the cloud services.

ROLE OF de.NBI CLOUD IN HEALTHYCLOUD

Within the HealthyCloud consortium, de.NBI Cloud is present with its partners of the sites EMBL Heidelberg, BIH@Charité Berlin, University Gießen, University Freiburg, University Tübingen and DKFZ Heidelberg. de.NBI Cloud is participating in two work packages:

(1) Work package 5, which is designated to the future computational design, is co-led by de.NBI Cloud and the Barcelona Supercomputing Center (BSC, Spain), as well as partners from the EGI foundation and the Finnish IT center for science (CSC). This work package, 'Designing a decentralized cloud for health data research', focuses on exploring existing and planned computation solutions, both in terms of infrastructure (hardware) and data management and analysis (software), that will enable ethical, technically feasible, and legally compliant decentralized computation for the future HRIC ecosystem. de.NBI Cloud brings its extensive experience in successfully operating a decentralized cloud for many years.

> **"de.NBI brings to the HealthyCloud project its tremendous expertise in managing such a complex research infrastructure at the highest standards"**
>
> Juan González-García (IACS, Spain) and
> Salvador Capella-Guitierrez (BSC, Spain)
> Coordinators of HealthyCloud

(2) Work package 7 deals with specific health-related use cases, centred on research questions related to cancer and atrial fibrillation. Those use cases will drive the interactions of the other work packages by challenging them to propose specific solutions to real world scenarios.

This work package is led by IACS (Instituto Aragones de Ciencias de la Salud), Spain and de.NBI Cloud serves as a partner. Especially with its expertise in cancer-related work, de.NBI Cloud can support data storage and analysis in the context of international large-scale studies. The Pan-Cancer Analysis of Whole Genomes (PCAWG) Project is one example of such a large-scale study. It involved a collaboration of researchers from 37 countries, and it extended and advanced methods for analysing cancer genomes and applied them to a dataset of more than 2,600 genomes of different tumor types [2]. One example of an open-source framework that facilitates large-scale scientific data analysis on the cloud by providing a comprehensive toolkit for virtual infrastructure management, workflow management, monitoring and error resolution which supports a variety of academic and commercial cloud computing platforms is the de.NBI/ELIXIR-DE tool Butler [3] - which was initiated during the course of the PCAWG project and is continued, supported, and maintained by funding from de.NBI. Results from the PCAWG project confirmed important findings from previous studies, such as on the number of cancer driver mutations, and also revealed new knowledge relevant to the biology of particular cancer types, for example pertaining to patterns of chromothripsis in melanoma [2].

SYNERGIES OF de.NBI CLOUD AND HEALTHYCLOUD

German researchers are involved in, or driving, a variety of projects aiming to decipher the underlying genetic and epigenetic etiology of diseases such as cancer. However, it is so far not clear how data from these studies can be computed on a cloud, channeled and made usable for research purposes. Approaches to solving those problems have been set up in the German context, the most prominent example of which is the presently established German Human Genome Phenome Archive (GHGA[1]), which once up and running aims to be directly integrated with the de.NBI Cloud to enable processing health-related genomic data on the cloud in a secure and ethically compliant manner. But such solutions are not yet in place to enable the secure and responsible sharing of genomic data across European countries.

The de.NBI Cloud already interacts with genomics and health research oriented infrastructures (for example the German National Research Data Infrastructures, NFDI, such as NFDI4Health or GHGA, where national health-related genome data from Germany are processed in the de.NBI Cloud) to facilitate data sharing at the national level, as well as internationally by hosting a mirror for ICGC data at the de.NBI Cloud sites in Berlin and Heidelberg. The cooperation (GHGA - de.NBI Cloud) might have model character to eventually achieve solutions for working with genomic disease data across European countries. In this setting, the upcoming Federated European Genome-Phenome Archive (fEGA) - a federation of national data archives in Europe in which GHGA acts as a node for Germany as a node for Germany, may allow more similar advances in Europe in the future.

In this regard, HealthyCloud serves as a vehicle to test and explore how genomic and health data can be exchanged and made usable for research in a European framework. German de.NBI Cloud experts are also part of other initiatives relevant to these goals, such as EOSC-Life and GAIA-X (for which other articles can be found in this brochure*). Additionally, de.NBI Cloud has a long-standing expertise in running a decentralized cloud with sites at eight different institutions throughout Germany. The de.NBI Cloud operates the major service levels. (1) Infrastructure as a Service (IaaS), (2) Platform as a Service (PaaS), and (3) Software as a Service (SaaS). Through a cloud federation concept, the different de.NBI sites are integrated into a single cloud platform. Users are guided to the desired service and the appropriate cloud via the central de.NBI Cloud portal. The system is based on Single Sign-On (SSO) and Authentication and Authorization Infrastructure (AAI) through Life Science Login (formerly ELIXR AAI)

REFERENCES: [1] Genome Med. 2020;12(1):18. DOI: https://doi.org/10.1186/s13073-020-0713-z. **[2]** Nature. 2020;578(7793):82-93. DOI: https://doi.org/10.1038/s41586-020-1969-6. **[3]** Nat Biotechnol. 2020;38(3):288-292. DOI: https://doi.org/10.1038/s41587-019-0360-3.

AUTHOR: Sina Barysch[1], Nina Habermann[2], Ivo Buchhalter[3], Alexander Goesmann[4], Björn Grüning[5], Jens Krüger[6], Harald Wagener[7], Jan O. Korbel[2]

[1] *European Molecular Biology Laboratory (EMBL), Structural and Computational Biology Unit, Meyerhofstr. 1, 69117 Heidelberg*
[2] *EMBL, Genome Biology Unit, Meyerhofstraße 1, 69117 Heidelberg*
[3] *Omics IT and Data Management Core Facility, DKFZ, Im Neuenheimer Feld 280, 69120 Heidelberg*
[4] *Justus-Liebig-Universität Gießen, Bioinformatics and Systems Biology, Ludwigstraße 23, 35390 Gießen*
[5] *Albert-Ludwigs-Universität Freiburg, Bioinformatics Group, Dept. of Computer Science, Georges-Köhler-Allee 106, 79110 Freiburg*
[6] *University of Tübingen, High Performance and Cloud Computing Group, IT Center, Wächterstraße 76, 72074 Tübingen*
[7] *Charité Universitätsmedizin Berlin, Center for Digital Health, BIH, Charitépl. 1, 10117 Berlin*

**You can also learn more about the GHGA and EOSC-Life in the articles on pages 28 and 48.*

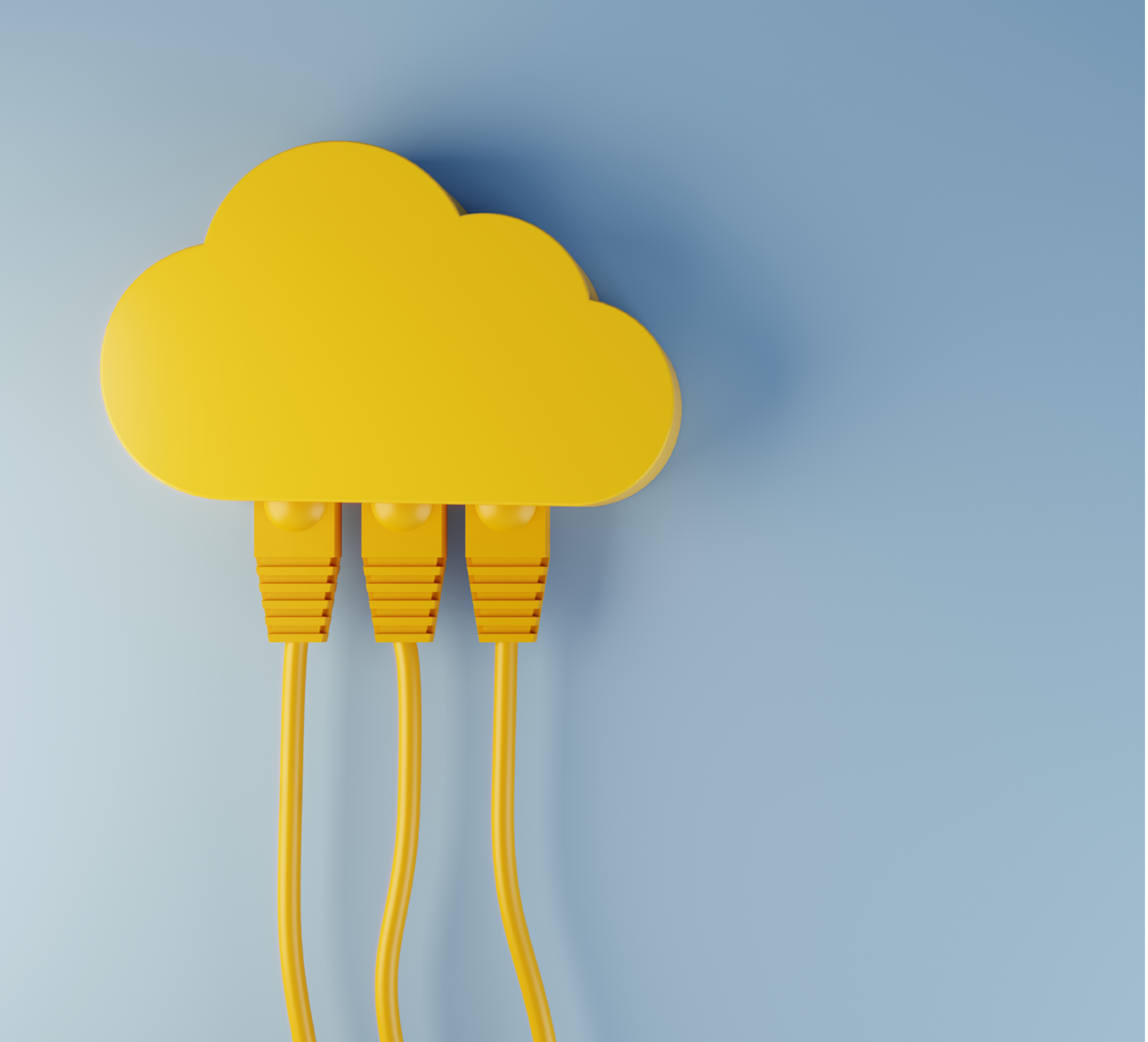


PHOTO CREDITS:

de.NBI Admistration Office
© iStock.com/Just_Super (Cover)
© S. Jonek/University Bielefeld (p. 3)
© iStock.com/monsitj (p. 4,10)
© iStock.com/MF3d (p. 5, 20)
© iStock.com/imaginima (p. 5, 46)
© shutterstock.com/Exclusively (p. 7)
© shutterstock.com/Evgeniy_D (p. 8)
© stock.adobe.com/blackboard (p. 12)
© stock.adobe.com/Funtap (p. 16)
© Steffen Neumann (p. 21)
© Carl Herrmann (p. 21)
© Matthias König (p. 21)
© stock.adobe.com/peach_adobe (p. 22)
© stock.adobe.com/j-mel (p. 24)
© stock.adobe.com/Worawut (p. 25)
© stock.adobe.com/monsitj (p. 28)
© iStock.com/Svisio (p. 33)
© iStock.com/jxfzsy (p. 34)
© iStock.com/Christoph Burgstedt (p. 38)
© iStock.com/K_E_N (p. 40)
© iStock.com/Design Cells (p. 41)
© stock.adobe.com/WrightStudio (p. 42)
© iStock.com/ThomasVogel (p. 49)
© iStock.com/Maks_Lab (p. 49)
© iStock.com/INchendio (p. 50)
© iStock.com/putilich (p. 52)
© Christoph Jäckle/ University Tübingen (p. 54)
© stock.adobe.com/ipopba (p. 55)
© stock.adobe.com/NicoElNino (p. 57)
© stock.adobe.com/Astibuag (p. 61)
© stock.adobe.com/eskay lim (p. 62)
© stock.adobe.com/ LuckyStep (p. 65)
© iStock.com/ Natali_Mis (p. 68)
© iStock.com/Eoneren (p. 72)

IMPRINT

Prof. Dr. Alexander Sczyrba
German Network for Bioinformatics Infrastructure (de.NBI)
de.NBI Cloud
Bielefeld University
Center for Biotechnology (CeBiTec)
Universitätsstraße 27
33615 Bielefeld

Tel: +49 (0) 521 106 2910
Fax: + +49 (0) 521 106-6411
E-Mail: contact@denbi.de

Editors-in-Chief:
Prof. Dr. Alexander Sczyrba (Bielefeld University, CeBiTec),
Prof. Dr. Alfred Pühler (Bielefeld University, CeBiTec)

Editorial Team:
Dr. Tanja Dammann-Kalionowski (Bielefeld University, CeBiTec),
Dr. Irena Maus (Bielefeld University, CeBiTec),
Prof. Dr. Andreas Tauch (Bielefeld University, CeBiTec)

www.denbi.de
@denbiOffice
linkedin.com/company/de-nbi

Date: July 2022

Design and Layout:
MEDIUM Werbeagentur GmbH, Bielefeld

DOI: https://doi.org/10.4119/unibi/2963745

SPONSORED BY

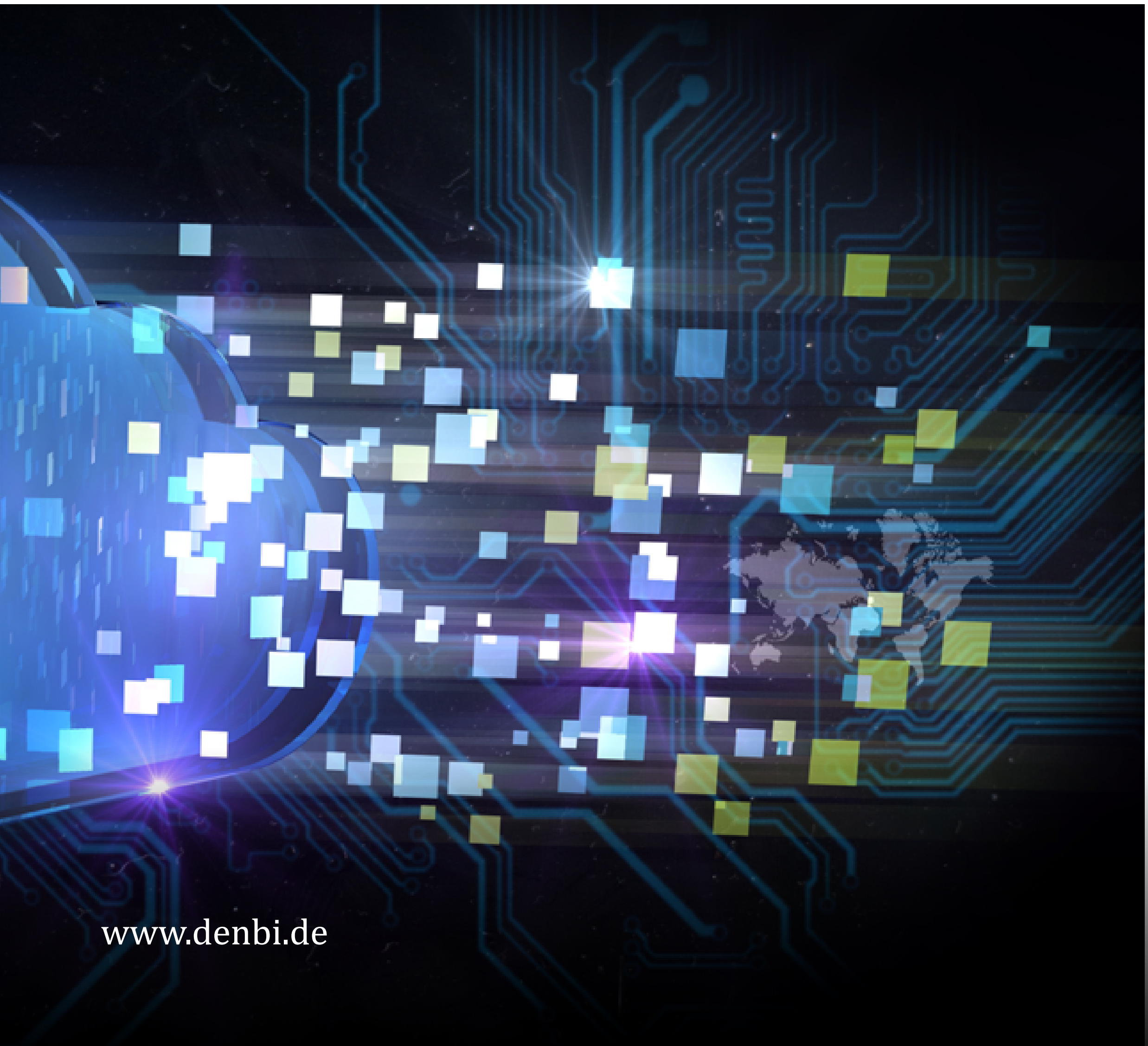
Federal Ministry
of Education
and Research

Fkz 031A532B
(de.NBI Administration Office)

UNIVERSITÄT
BIELEFELD

elixir GERMANY

de.NBI
GERMAN NETWORK FOR BIOINFORMATICS INFRASTRUCTURE



www.denbi.de