Health Data Cloud (D6.7 – SGA3)

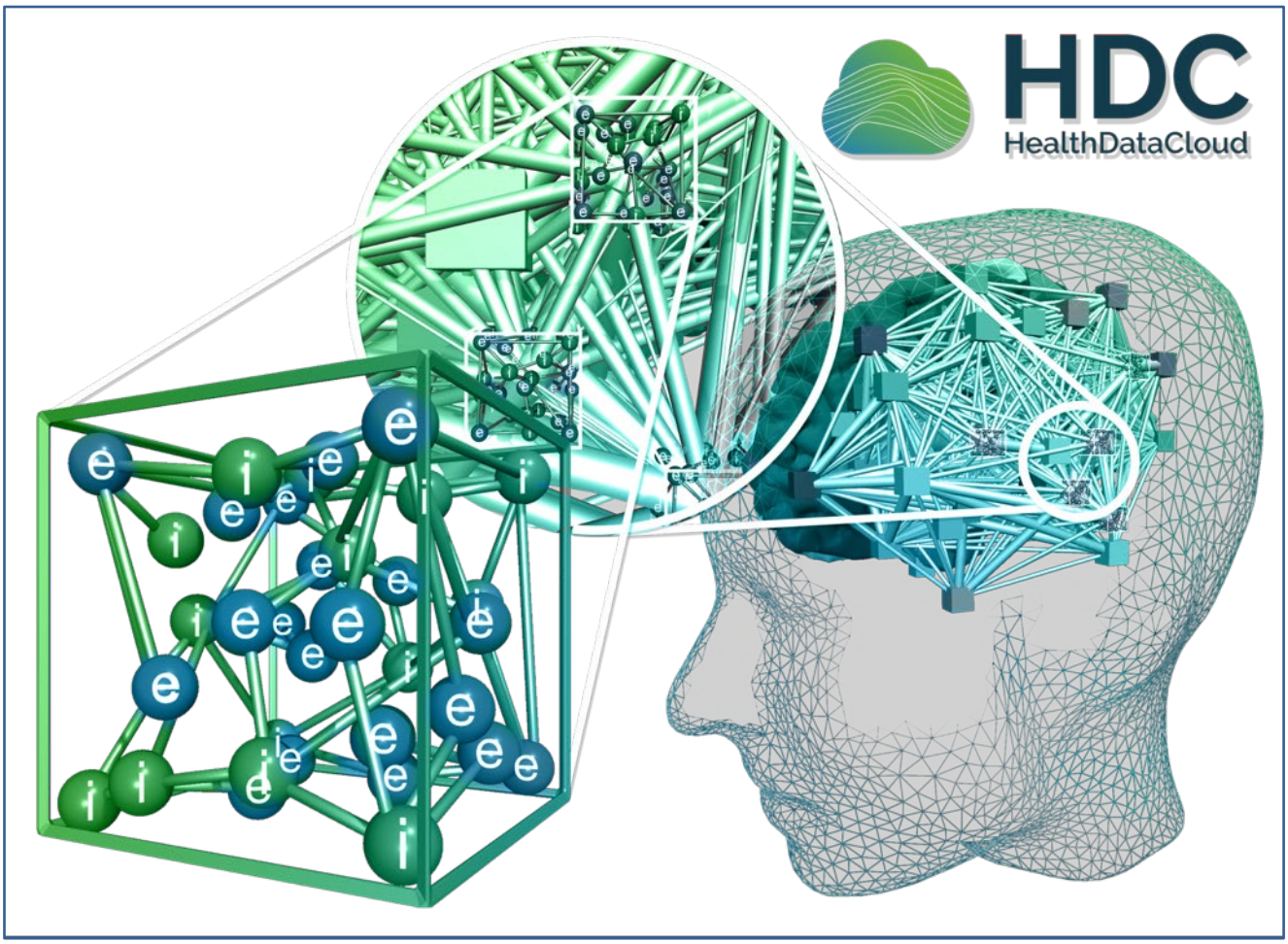

Figure 1: Health Data Cloud (HDC) delivers EBRAINS Service for Sensitive Data

Please visit our website: http://www.healthdatacloud.eu/. The Health Data Cloud Service can be reached at EBRAINS via https://www.ebrains.eu/health-research-platforms/health-platforms/work-with-health-data-2.

Project Number:	945539	Project Title:	HBP SGA3

Document Title:	Health Data Cloud
Document Filename:	D6.7 (D113) SGA3 M36 SUBMITTED 230901.docx
Deliverable Number:	SGA3 D6.7 (D113)
Deliverable Type:	Demonstrator
Dissemination Level:	PU = Public
Planned Delivery Date:	SGA3 M42 / 30 Sep 2023
Actual Delivery Date:	SGA3 M42 / 01 Sep 2023
Author(s):	Petra RITTER, CHARITE (P122)
Compiled by:	Petra RITTER, CHARITE (P122)
Contributor(s):	Michael SCHIRNER, CHARITE (P122), contributed to all sections Patrik BEY, CHARITE (P122), contributed to Section 2 Denis DOLL, INDOC (P154), contributed to Section 4 Dirk PLEITER, KTH (P39), contributed to Section 4 Ainar DREWS, OUS (P153), contributed to Section 4 Boris ORTH, JUELICH (P20), contributed to Section 4
WP QC Review:	Anna LÜHRS, JUELICH (P20)
WP Leader / Deputy Leader Sign Off:	Boris ORTH, JUELICH (P20)
T7.4 QC Review:	N/A
Description in GA:	Demonstrator of the Health Data Cloud functional platform
Abstract:	Biomedical research is currently hindered because a technical infrastructure is missing to protect the privacy of personal data. Problematically, biomedical data cannot be easily anonymized or pseudonymized such that personally identifiable information are removed, and potential re-identification is excluded. Person-related information are the explicit target in biomedical research and for personalized therapy. Precisely that information that characterizes the biomedical state of a person can also be used to re-identify a person. Those features that create a better understanding of health and illness of a person are the ones that identify a person. Neuroimaging data, and especially data related to personalized brain network modelling, like structural connectomes, functional connectomes, brain surfaces, MRI images, fMRI time series, PET loadings, simulated brain activity, parameter estimation results, etc. can potentially identify the person from whom the data was recorded and therefore EU laws require proper protection of such data. EBRAINS Service for Sensitive Data (SSD) – the Health Data Cloud (HDC) is an open-source data management and processing platform that enables medical researchers to store, process and share data in compliance with the European Union (EU) General Data Protection Regulation (GDPR). The HDC addresses the present lack of digital research data infrastructures fulfilling the need for (a) data protection for sensitive data, (b) capability to process complex data such as radiologic imaging, (c) flexibility for creating own processing workflows, (d) access to high performance

	computing. The platform promotes FAIR data principles and reduces barriers to biomedical research and innovation. It offers a web portal with graphical and command-line interfaces, segregated data zones and organizational measures for lawful data onboarding, isolated computing environments where large teams can collaboratively process sensitive data privately, analytics workbench tools for processing, analysing, and visualizing large datasets, automated ingestion of hospital data sources, project-specific data warehouses for structured storage and retrieval, graph databases to capture and query ontology-based metadata, provenance tracking, version control, and support for automated data extraction and indexing. The HDC is based on a modular and extendable state-of-the art cloud computing framework, a RESTful API, open developer meetings, hackathons, and comprehensive documentation for users, developers, and administrators. The HDC with its concerted technical and organizational measures can be adopted by other research communities and thus facilitates the development of a co-evolving interoperable platform ecosystem with an active research community. A prominent demonstrator of HDC's functional scope is a study on 650 personalized human brain digital twins published in Nature Communications.
Keywords:	Sensitive Data, Health Data, GDPR
Target Users/Readers:	Clinicians, computational neuroscience community, computer scientists, consortium members, funders, general public, HPC community, neuroimaging community, neuroinformaticians, neuroscientific community, neuroscientists, platform users, policymakers, researchers, scientific community, students

Table of Contents

Table of Figures

1. Introduction

Biomedical research is currently hindered because a technical infrastructure is missing to protect the privacy of personal data. Problematically, biomedical data cannot be easily anonymized or pseudonymized such that personally identifiable information are removed, and potential re-identification is excluded. Person-related information are the explicit target in biomedical research and for personalized therapy. Precisely that information that characterizes the biomedical state of a person can also be used to re-identify a person. Those features that create a better understanding of health and illness of a person are the ones that identify a person. Neuroimaging data, and especially data related to personalized brain network modelling, like structural connectomes, functional connectomes, brain surfaces, MRI images, fMRI time series, PET loadings, simulated brain activity, parameter estimation results, etc. can potentially identify the person from whom the data was recorded and therefore EU laws require proper protection of such data.

EBRAINS Service for Sensitive Data (SSD) – the Health Data Cloud (HDC) is an open-source data management platform that enables medical researchers to store, process and share data in compliance with the European Union (EU) General Data Protection Regulation (GDPR). The HDC addresses the present lack of digital research data infrastructures fulfilling the need for (a) data protection for sensitive data, (b) capability to process complex data such as radiologic imaging, (c) flexibility for creating own processing workflows, (d) access to high performance computing. The platform promotes FAIR data principles and reduces barriers to biomedical research and innovation. It offers a web portal with graphical and command-line interfaces, segregated data zones and organizational measures for lawful data onboarding, isolated computing environments where large teams can collaboratively process sensitive data privately, analytics workbench tools for processing, analyzing, and visualizing large datasets, automated ingestion of hospital data sources, project-specific data warehouses for structured storage and retrieval, graph databases to capture and query ontology-based metadata, provenance tracking, version control, and support for automated data extraction and indexing. The HDC is based on a modular and extendable state-of-the art cloud computing framework, a RESTful API, open developer meetings, hackathons, and comprehensive documentation for users, developers, and administrators. The HDC with its concerted technical and organizational measures can be adopted by other research communities and thus facilitates the development of a co-evolving interoperable platform ecosystem with an active research community.

A prominent demonstrator of HDC's functional scope is a study on 650 personalized human brain digital twins in Nature Communications.

The HealthDataCloud (HDC) partners provide an operational solution of EBRAINS Health Data Cloud - EBRAINS Service for Sensitive Data (SSD)[1] that draws form previous developments of an EU wide cloud infrastructure for health data with data protection by design and by default in the European Open Science Cloud (EOSC) project Virtual Brain Cloud[2] that has been led by CHARITE from 2018 to 2023.

Thus, the foundation for EBRAINS Health Data Cloud is a GDPR-compliant and EBRAINS interoperable Virtual Research Environment (VRE) located at the Charité[3] that provides a secure and scalable data platform enabling multi-institutional research teams to store, share and analyse complex multi-modal health datasets (see Figure 2).

Presently the HDC node at Charité is GDPR audited, fully functional and at the service of EBRAINS users. The central HDC node at EBRAINS is in production and undergoes user testing[4]. HDC offers its users an operational GDPR compliant cloud-based platform for medical research that helps to improve early patient-specific diagnosis and treatment of brain diseases.

While the HDC - a federated network of interoperable GDPR compliant clouds – is further extended by additional nodes iteratively - EBRAINS HDC already now provides to its users certified services for

[1] https://www.ebrains.eu/health-research-platforms/health-platforms/work-with-health-data-2
[2] https://cordis.europa.eu/project/id/826421
[3] https://vre.charite.de
[4] https://hdc.humanbrainproject.eu/login

sensitive health and associated data in compliance with GDPR. Its services serve multinational research consortia in Europe for basic, translational, and clinical research. The specific challenge that HDC is providing a solution for is that of creating an infrastructure for cross-site EU-wide collaboration on sensitive health data. This requires several layers of technical and process development, legal assessment, and definition of liabilities of all involved stakeholders. Our distributed model by which any of the EBRAINS HDC nodes can sustain the basic system, allows for an iterative extension and flexible adaption to local legal requirements.

EBRAINS HDC leverages EBRAINS where it stands out already: Integrated data, research software and computing services, and an interdisciplinary community. These features are now available also for sensitive health data and digital human twins that require proper protection by law. Thus, HDC contributes to HBP's main objectives of creating and operating a European Scientific Research Infrastructure for brain research, cognitive neuroscience, and other brain-inspired sciences; gathering, organizing and disseminating data describing the brain and its diseases; simulating the brain; building multi-scale scaffold theory and models for the brain; developing brain-inspired computing, data analytics and robotics; and ensuring that the HBP's work is undertaken responsibly and that it benefits society.

EBRAINS HDC will be actively further developed and maintained – jointly with a network of EU-wide partners – beyond the current and final HBP phase SGA3. Significant funding has been secured via the Horizon Europe project eBRAIN-Health[5] (20 partners) and the Digital Europe project TEF-Health[6] (51 partners) – both consortia being led by CHARITE. HDC will provide and further advance a GDPR-compliant federated European cloud solution for health data in EBRAINS for EBRAINS during its further evolution on the ESFRI roadmap.

HDC will pave the way for a clinical product for personalized medicine that improves the quality of life of EU citizens by enabling targeted prevention, early diagnosis, disease progression prognosis, individual treatment plans and development of novel therapies for brain diseases within a European digital Research Infrastructure. We believe that this vision will be strongly supported by EBRAINS HDC, a European cloud-based platform that is at the service of clinics, researchers, patients, and students.

[5] https://cordis.europa.eu/project/id/101058516
[6] https://www.tefhealth.eu/

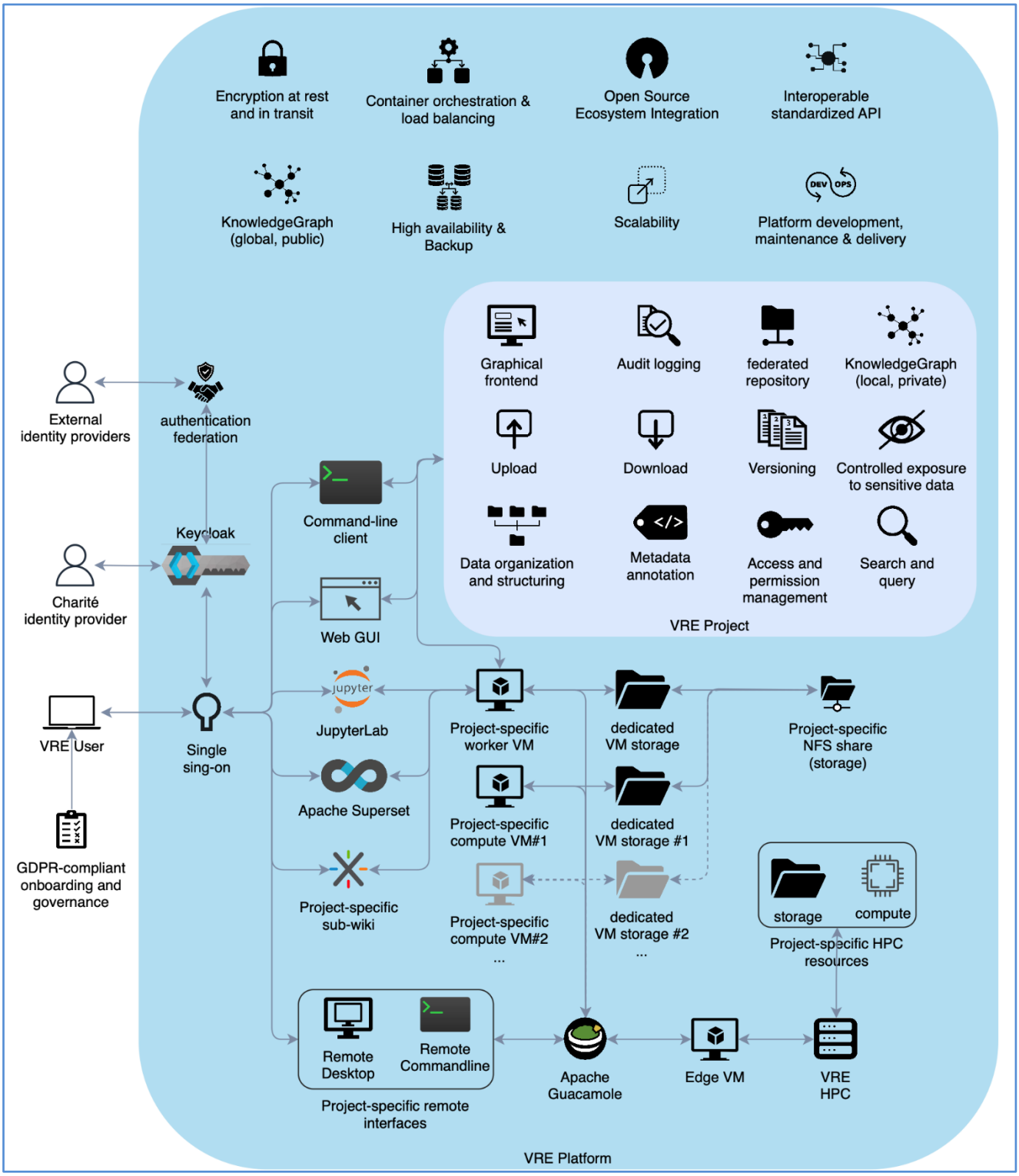

Figure 2: HDC high-level node architecture

After users have passed an onboarding procedure that includes filling out processing agreements in which the users' role under GDPR as controller or processor is fixed in case personal data are involved, users can login via the Charité identity provider. Federation with other identity providers is technically prepared. Via graphical and command line (GUI/CLI) interfaces users get access to isolated environments that offer on the one hand rich data structuring functionality and on the other hand the power of a cloud performant backend with high availability, backup, scalability, container orchestration and multiple other features. Users receive access to private isolated machines with remote desktop and remote CLI, storage backends and HPC backends where containerized workflows can be pushed and executed.

2. Health Data Cloud: GDPR Compliant Service for Sensitive Data

2.1 Sharing Health Data via EBRAINS

The HDC consortium has provided an operational solution of EBRAINS HealthDataCloud (HDC) - EBRAINS Service for Sensitive Data (SSD). The HDC enables research on sensitive personal or health data originating from human subjects that includes collaborative sharing, collaborative collection, and collaborative analysis of sensitive data as well as defined routes for public sharing of its outputs. The foundation for EBRAINS HealthDataCloud is a GDPR-compliant and EBRAINS interoperable Virtual Research Environment (VRE) located at the Charité[3] that provides a secure and scalable data platform enabling multi-institutional research teams to store, share and analyse complex multi-modal health datasets. While the HDC node at Charité University Medicine is fully operational and at the service of EBRAINS users, additional HDC nodes are presently under development. Thus, HDC will become a scalable, federated network of interoperable technology stacks (referred to as nodes) sharing a standard architecture. Presently the HDC node at Charité is GDPR audited, fully functional and at the service of EBRAINS users. The central node at EBRAINS has been delivered as Minimum Viable Product (MVP) and undergoes user testing.

With HDC Satellite VRE node at Charité, EBRAINS offers its users an operational GDPR compliant Cloud-based platform for medical research that helps to improve early patient-specific diagnosis and treatment of brain diseases. Our long-term vision for HDC is to pave the way for a clinical product for personalized medicine that improves the quality of life of EU citizens by enabling targeted prevention, early diagnosis, disease progression prognosis, individual treatment plans and development of novel therapies for brain diseases. We believe that this vision will be strongly supported by EBRAINS HDC, a European cloud-based platform that connects clinics, researchers, patients, and students.

2.2 FAIR Data and Interoperability

EBRAINS HDC satisfies the core design principles of FAIR data, co-locality of data with computing resources, support for multiple data stores at distributed sites being federated and discoverable on the EBRAINS Knowledge Graph (KG), and support for open standards at all technical levels. The HDC addresses the present lack of digital research data infrastructures fulfilling the need for (a) data protection for sensitive data, (b) capability to process complex data such as radiologic imaging, (c) flexibility for creating own processing workflows, (d) access to high-performance computing. The platform promotes FAIR data principles and reduces barriers to biomedical research and innovation.

2.2.1 Metadata

To understand the contents of the data the user can consult the data's metadata. For annotating data with metadata users can complete a GUI form or upload JSON schemas, either adhering to standardized metadata schemas or defining a custom format. For standardizing data, users have GUI and command line tools to conveniently bring data into standardized formats and to verify that a dataset adheres to a standard definition, similar to the BIDS validator.

A major design focus lies on robust and convenient re-use of intermediate processing results and variants of results. The user easily identifies data that exists at various stages and variants of processing and the full processing history of all digital objects. Version control and provenance tracking is performed with a GUI that provides a tree structure that shows all downstream processing results of a file and different processing variants. In addition, the tree has an additional dimension that allows to view the history of versions of each file and to view downstream processing results that were derived from each file.

The HDC User Journey is outlined in Figure 3.

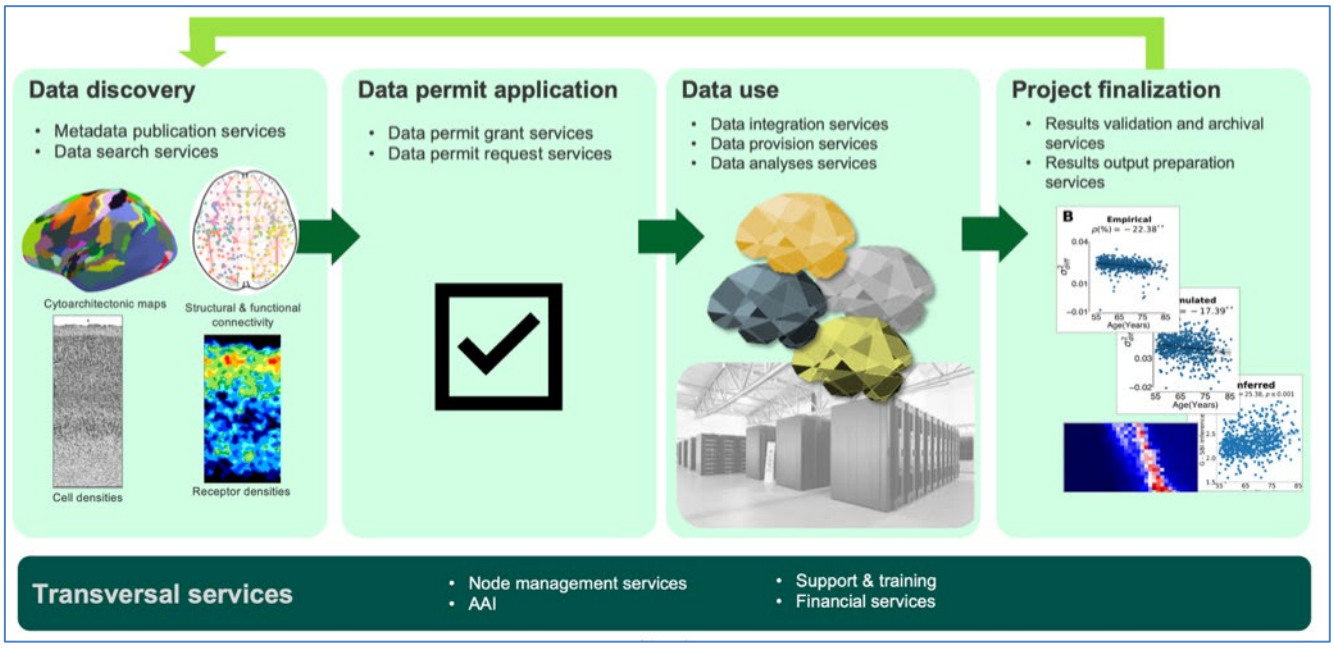

Figure 3: HDC User Journey

The HDC User Journey can be divided into four phases with corresponding services and an additional set of transversal services. Phase 1: Data discovery, Phase 2: Data permit application, Phase 3: Data use; Phase 4: Project finalization. Generated derivative data are provisioned as FAIR data for discovery and re-use. The transversal services comprise: AAI, node management services, support and training and financial services.

2.2.2 Making data findable

Data and services are being made discoverable to users via the EBRAINS web portal[7] and EBRAINS Knowledge Graph[8].

While personal and health data are not openly accessible, they are being made discoverable via the EBRAINS Knowledge Graph. While only authorized persons can access personal data in the HDC, the data sets remain discoverable through the EBRAINS Knowledge Graph. Each data set is equipped with a data descriptor that contains information on the format, data standards, the detailed format description, and compatible software that are used in the platform. Curation in EBRAINS is supported and curation requests[9] can be made by users.

2.2.3 Offer Data for Secondary Use via EBRAINS

Data in the EBRAINS HDC can be made discoverable via existing EBRAINS Data and Knowledge services. Non-sensitive metadata of the data inside the HDC can be submitted to the EBRAINS Knowledge Graph Spaces[10] for discoverability. Dedicated programmatic and GUIs support the metadata transfer to the EBRAINS Knowledge Graph.

[7] https://www.ebrains.eu/
[8] https://kg.ebrains.eu/
[9] https://nettskjema.no/a/104328#/page/1
[10] https://search.kg.ebrains.eu

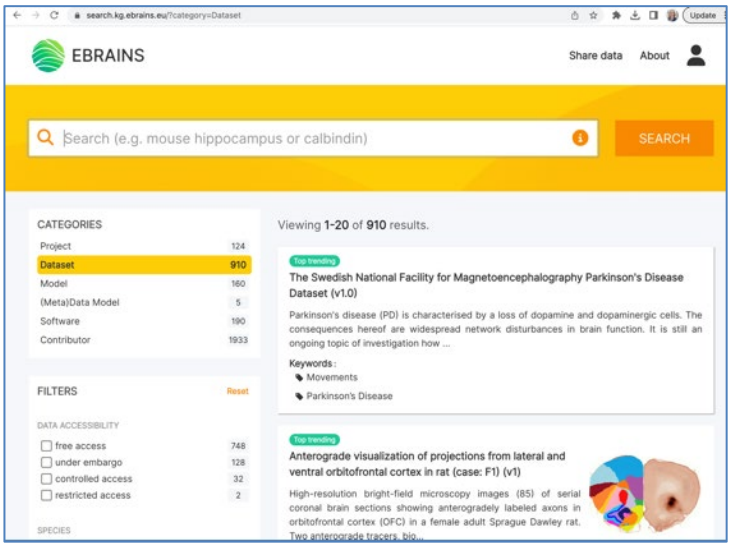

Figure 4: EBRAINS Knowledge Graph

The EBRAINS Knowledge Graph is a graph data base with a programmatic and graphical user interface for discovery of data, models, and software tools. It supports multiple metadata models. A prominent one used in EBRAINS and supported by HDC's VRE is OpenMinds[11].

In the EBRAINS Knowledge Graph (see Figure 4), each data set is equipped with a data set card (see Figure 5). For sensitive data sets stored in the HDC satellite VRE, the data set card in the "Get data" field informs: "The data is currently shared via the Virtual Research Environment at Charité – Universitätsmedizin Berlin (https://www.bihealth.org/en/translation/network/digital-medicine/bihcharite-virtual-research-environment). Please contact [data controller name and email address] for access." All datasets in the HDC receive a DOI.

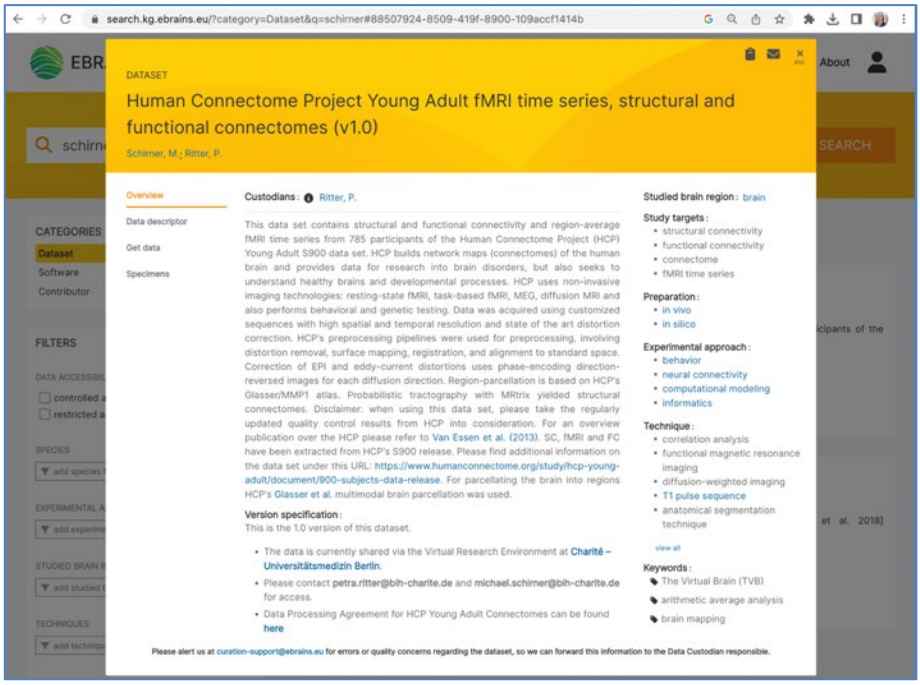

Figure 5: Example of a data set card for sensitive data

Each data set that is registered in the EBRAINS Knowledge Graph is equipped with a data set card. In case of a sensitive data set the card provides the information that the data set is stored in GDPR compliant Virtual Research Environment at the Charité[12].

Upon conclusion of a data processing agreement and a data protection impact assessment, the data are released by the data controller to a newly created "project" to which only the new user has

[11] https://github.com/HumanBrainProject/openMINDS
[12] https://search.kg.ebrains.eu/?category=Dataset&q=schirner#88507924-8509-419f-8900-109accf1414b

access and the admin right to admit authorized team members (e.g., processors with whom a processing agreement has been concluded).

2.2.4 Access Health Data via EBRAINS

The data processing agreement used for sharing sensitive data corresponds to the one recommended by the European Commission[13] (see Figure 6):

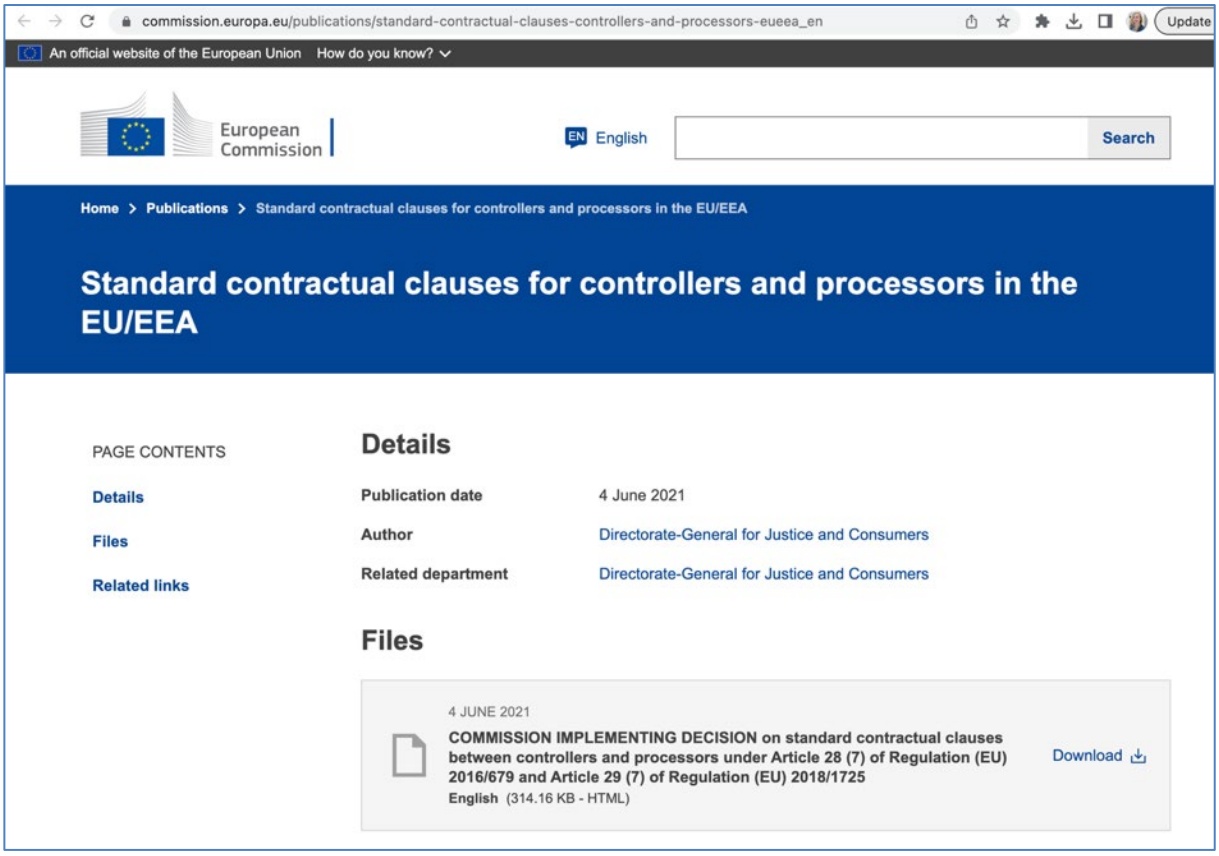

Figure 6: Standard Contractual Clauses (SCC)

SCCs provisioned by the European Commission are used as a contractual agreement between data controllers and processors when sharing sensitive data.

A link in the data set card refers to the prefilled SCC for the specific data set. The link points to an EBRAINS Collaboratory space where detailed information about the data set processing agreement and license can be found (see Figure 7).

The user fills the processing agreement (SCC) with its information (name, purpose of processing, measures to protect the data), signs and submits it to the data controller.

13 https://commission.europa.eu/publications/standard-contractual-clauses-controllers-and-processors-eueea_en

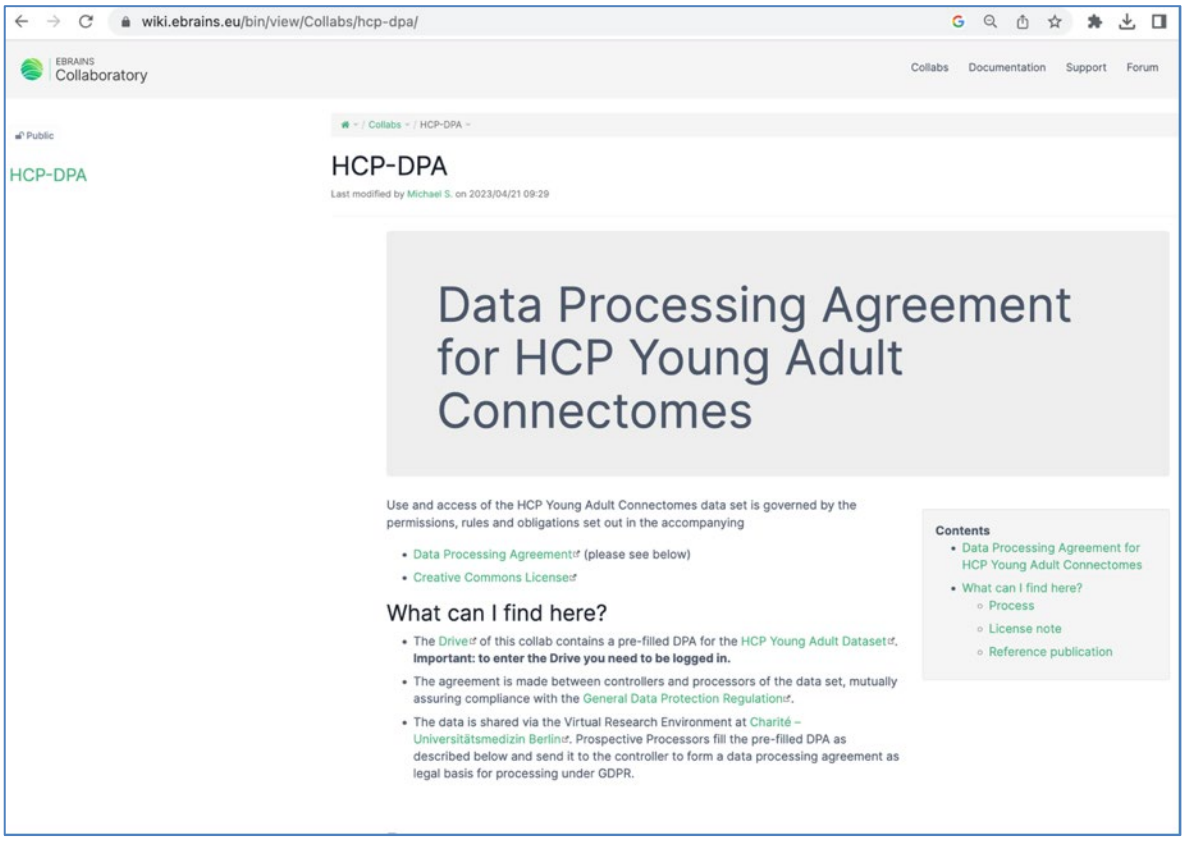

Figure 7: Collaboratory page with processing and licensing information

This exemplary Collaboratory page contains information about the data processing agreement and license that applies to the data set[14].

2.2.5 Licensing

The CC BY4.0 license on human health data could be construed as misleading due to its incompatibility with GDPR as the creator of the health information depicted, e.g., in MRI data is the data subject and it does not result from the creative process of a scientist that could give us the right to waive copyright protection. If such a license is applied it must be clear that it only pertains to very limited aspects of the data set, like for example how the files and folders are organized, but not to the health information stored inside the files. HDC suggests a License Statement similar to the following:

"We explicitly grant re-use of the overall dataset under the Creative Commons License Attribution-ShareAlike 4.0 International[15]. The license explicitly grants reuse of all non-personal aspects of the dataset. Importantly, the personal data contained within this data set is governed by the EU General Data Protection Regulation[16]. As a consequence, the license only applies to non-personal aspects of the dataset, for example, relating to the structure and organization of the dataset or the way the dataset was produced, but not to the personal information contained within the dataset."

2.2.6 Processing Health Data via EBRAINS

The HDC with its operational GDPR compliant, audited VRE node at Charité (see Figure 8) offers a web portal with graphical and command-line interfaces, segregated data zones and organizational measures for lawful data onboarding, isolated computing environments where large teams can collaboratively process sensitive data privately, analytics workbench tools for processing, analysing,

[14] https://wiki.ebrains.eu/bin/view/Collabs/hcp-dpa/
[15] http://creativecommons.org/licenses/by-sa/4.0/
[16] https://commission.europa.eu/law/law-topic/data-protection_en

and visualizing large datasets, automated ingestion of hospital data sources, project-specific data warehouses for structured storage and retrieval, graph databases to capture and query ontology-based metadata, provenance tracking, version control, and support for automated data extraction and indexing. HDC node VRE at Charité is based on a modular and extendable state-of-the-art cloud computing framework, a RESTful API, open developer meetings, hackathons, and comprehensive documentation for users, developers, and administrators. Data in HDC's VRE are always encrypted at rest. For processing, data are sent to isolated VMs with strict access control. Only in these protected VM environments the data can be decrypted and analysed by researchers, according to the predefined access policy.

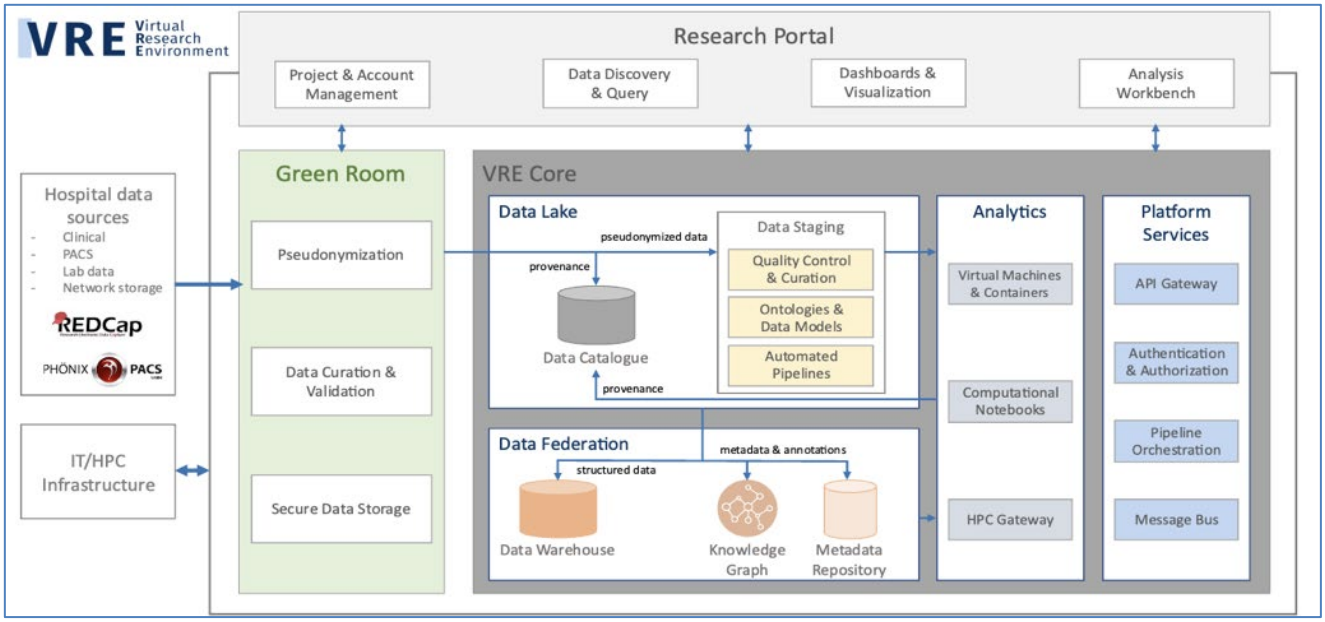

Figure 8: High-level architecture of the secure VRE at Charité

The centralized data processing involves securely transferring data to a trusted server, where the data can be analysed. By securely transferring data to HDC's trusted cloud infrastructure, clients can be confident that their data is protected from unauthorized access and potential cyber-attacks. The secure VRE at Charité data centre (also accessible for instance via the EOSC Marketplace[17]) is a full stack automated infrastructure as code (IaC) - open source under EUPL and equipped with detailed deployment and operation documentation. The secure VRE also ensures that data is stored and managed in compliance with relevant data protection regulations and standards. Datasets from different research projects and modalities, including clinical, imaging, and genomic data, can be combined into a centralized Data Warehouse. Importantly only authorized data controllers and processors can access the respective data. Data controllers stay in control of the data they are liable for. GDPR defines the roles of data subjects, data controllers (determine the purpose of processing) and data processors (process data on behalf of the controller). GDPR explicitly does not foresee another authority beyond these defined roles. Therefore, HDC's secure VRE offers controllers full control over their data. For each data set the controller may differ. Controllers can be individual researchers or controllership can be executed by use and access committees e.g., in case of hospital data. Metadata is captured centrally in a Metadata Repository, which helps researchers find and extract their data for downstream analysis and visualization.

[17] https://marketplace.eosc-portal.eu/services/secure-virtual-research-environment-vre-for-sensitive-data

2.2.6.1 Collaborative workbench: Health Data Cloud and Virtual Research Environment

The first step is to access the HDC from the EBRAINS portal URL[18]. Only registered users can get access to the portal.

The user is invited by a Project Administrator to access a project, in this example case the project Lesion2TVB. We here describe exemplarily the containerized data processing and handling as implemented for the current Lesion2TVB use case project set up inside the VRE portal.

A dataset from a subject (sub-0000), composed by a Nifti image and the corresponding openMINDS metadata schema are uploaded into the HDC Green Room storage (see Figure 9).

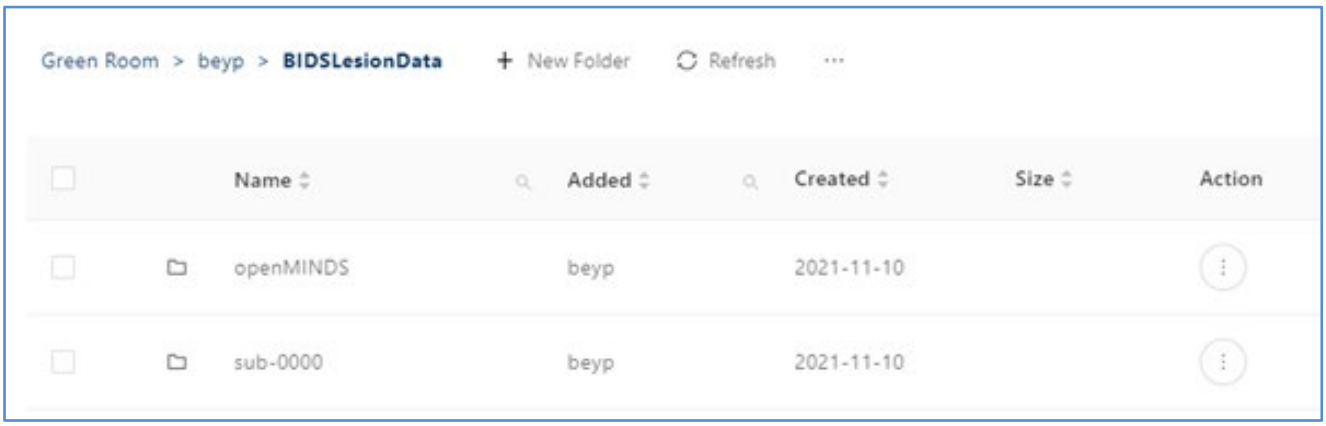

Figure 9: HDC data upload

Upon upload, the data are stored in the HDC Green Room storage area of the Lesion2TVB project. After integrating the data using the HDC CLI tool (a command line client that provides access to HDC managed services and data), we run a defacing pipeline for the given anatomical input modalities and can visually check the results using a Nifti visualization tool (MRIcroGL[19]) inside the HDC Green Room VM (see Figure 10).

[18] https://www.ebrains.eu/health-research-platforms/health-platforms/work-with-health-data-2
[19] https://www.nitrc.org/projects/mricrogl/

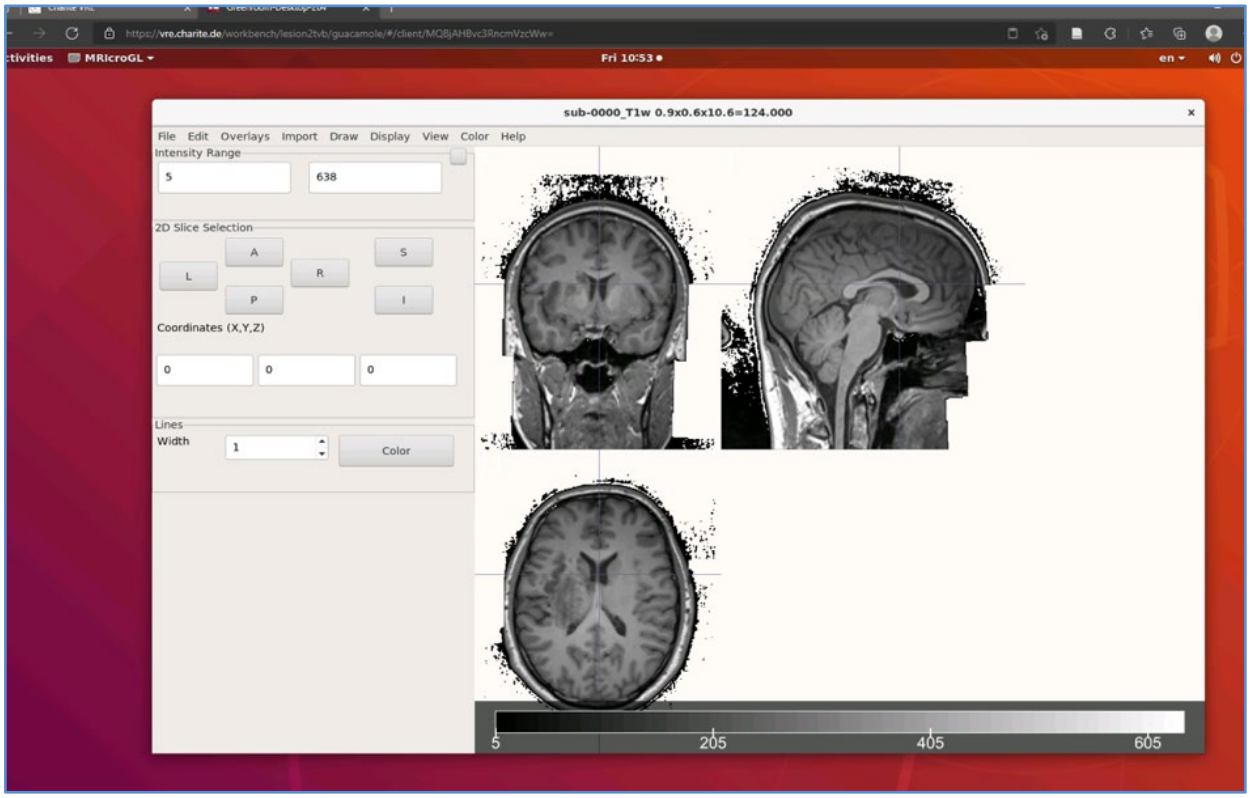

Figure 10: Visualization of results in the HDC

The results are then uploaded back into the HDC Green Room storage to be approved and copied by the Project Administrator to the project's HDC Core storage (see Figure 11).

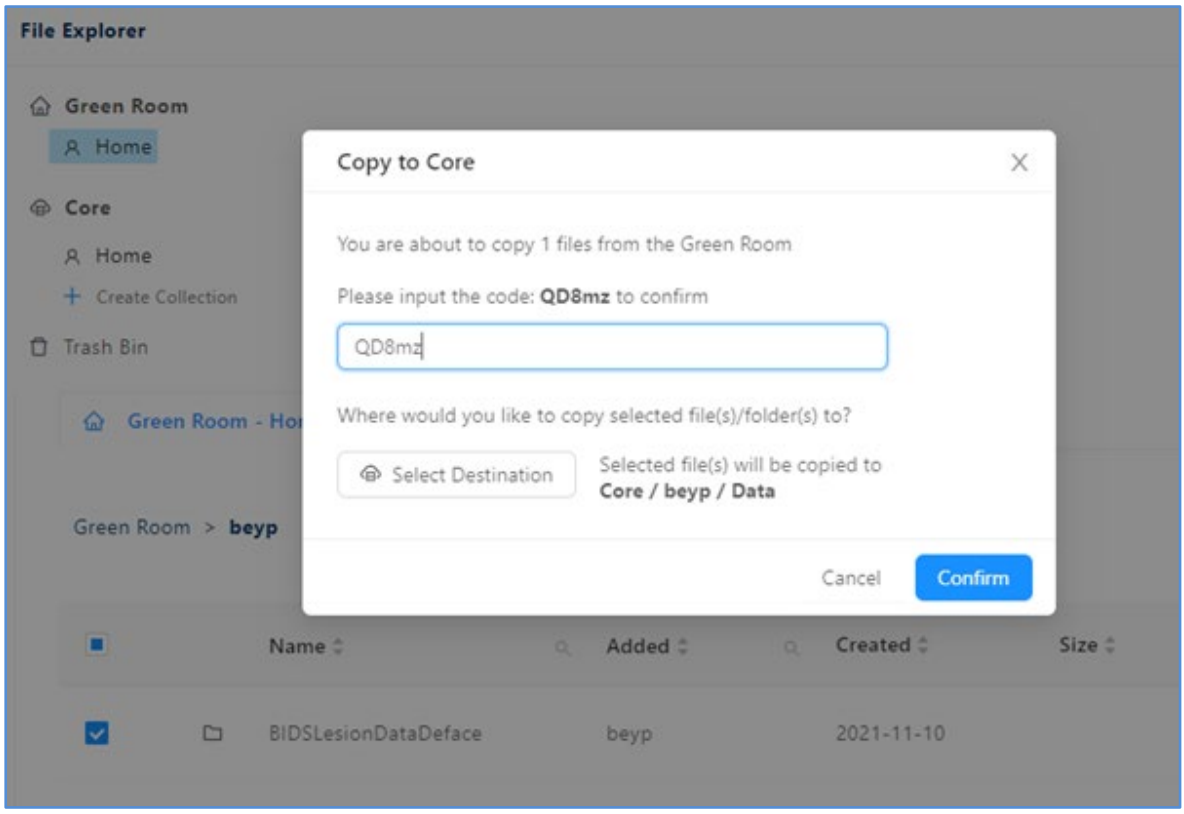

Figure 11: Project admins can copy data from Green Room to HDC Core

Once the upload to the HDC Core storage is completed, users can access an HDC Core VM to further process the data. First, data are downloaded inside the VM using the HDC CLI tool (Figure 12 Data load inside the HDC VM via CLI).

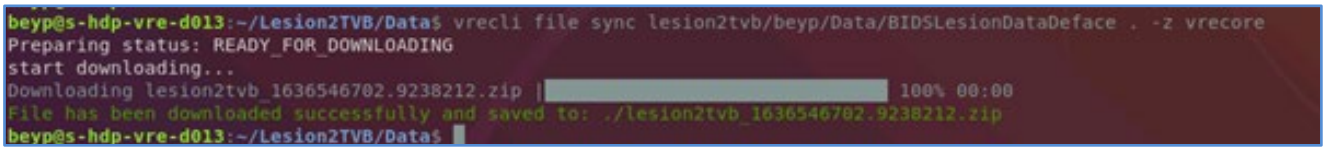

Figure 12: Data load inside the HDC VM via CLI

In a first processing step we perform lesion correction using a developed container image performing enantiomorphic brain correction for all input modalities using the following code snippet:

```
cd $HOME/Lesion2TVB/Container


singularity run \
    -bind “${HOME}/Lesion2TVB/Data/BIDSLesionDataDeface”:/data
    -env Input="/data/sub-0000/ses-01/anat/sub-0000_T1w.nii.gz" \
    -env Mask="/data/sub-0000/ses-01/anat/T1w_lesion_mask.nii.gz” \
    Ebc
```

To validate the results, we use the GUI enabled HDC Core VM and visually validate the performance (see Figure 13).

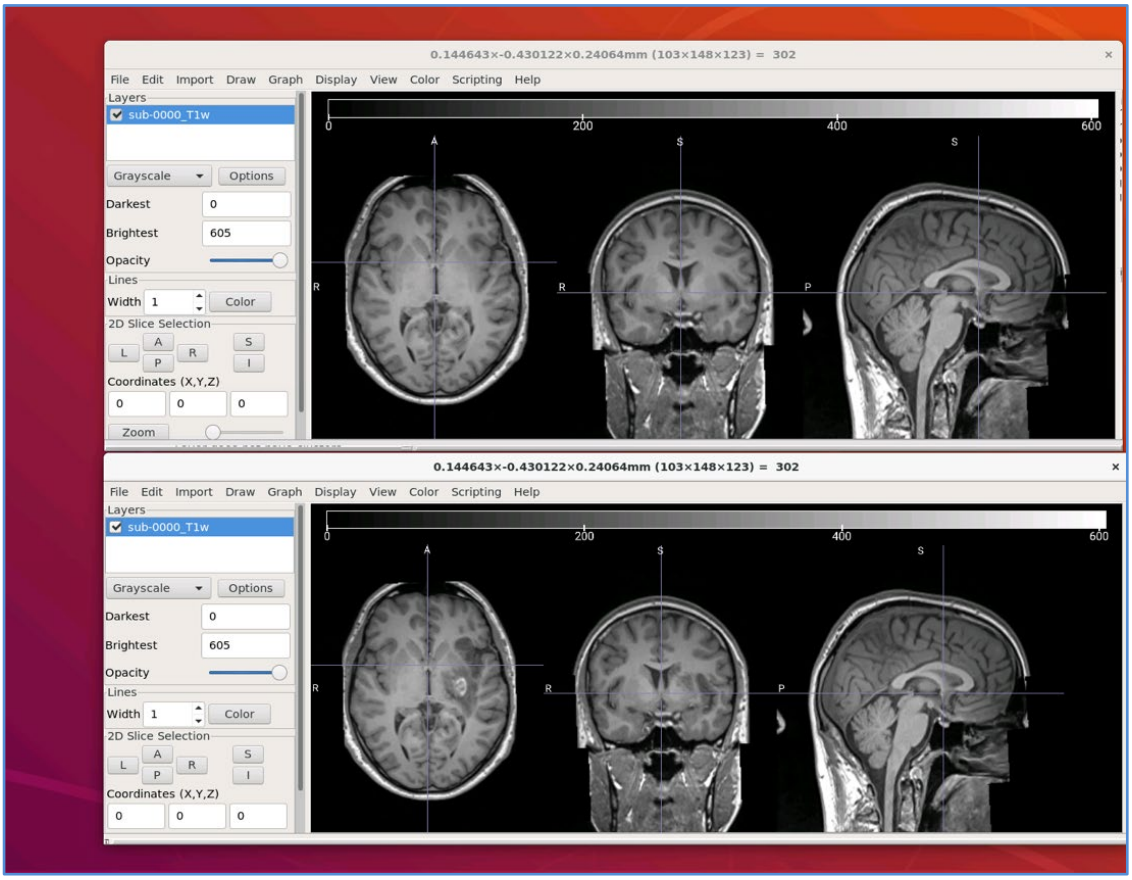

Figure 13: Visualization of a pipeline image correction step

Using the corrected data, users can start the computationally expensive processing using the Human Connectome Project (HCP) minimal processing pipeline inside the HPC environment. To this end users connect using ssh to the cluster (see Figure 14).

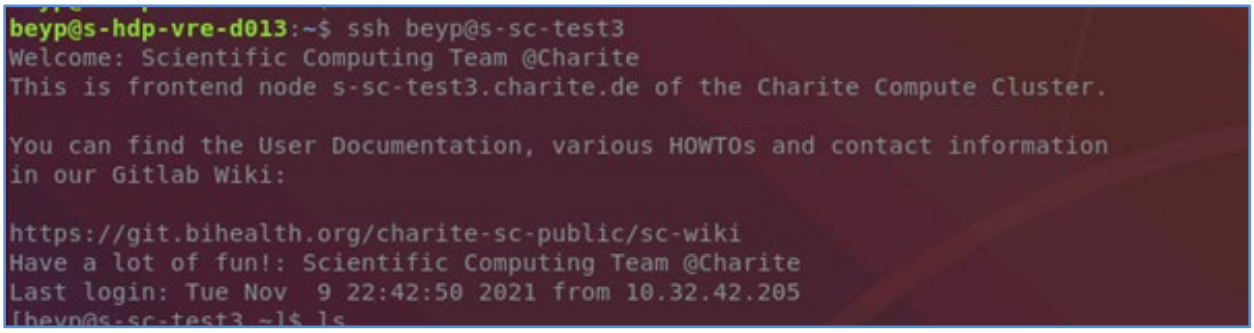

Figure 14: Connecting to HDC HPC backend

Inside the cluster users can use HDC CLI again to retrieve the created data and submit a batch script to the SLURM job management tool of the cluster, calling the corresponding HCP container image for processing:

```
cd Lesion2TVB/Container


Path=”home/beyp/Lesion2TVB/Data/BIDSLesionDataDefaceCorrected”
OutDir=”/home/beyp/Lesion2TVB/Data/Processed”
SubID=”0000”
Steps=”PreFreeSurfer FreeSurfer PostFreeSurfer”


sbatch \
HCPPipelinesBatch.sh \
--path=”${Path}” \
--subject=”${SubID}” \
--outdir=”${OutDir}” \
--stage=”${Steps}”
```

In a similar fashion, users can process the diffusion imaging data to create structural connectomes for the patient and create TVB-ready data that is then pushed back into project-specific VRE Core storage for subsequent retrieval by the TVB backend.

2.2.6.2 The Virtual Brain (TVB) Web GUI

To continue the workflow, we can perform analysis and simulations with the data uploaded in the previous step using the TVB simulation software installed in the VRE. To access TVB, the user can launch a Core VM from the "Guacamole" option (the remote desktop gateway in Figure 15).

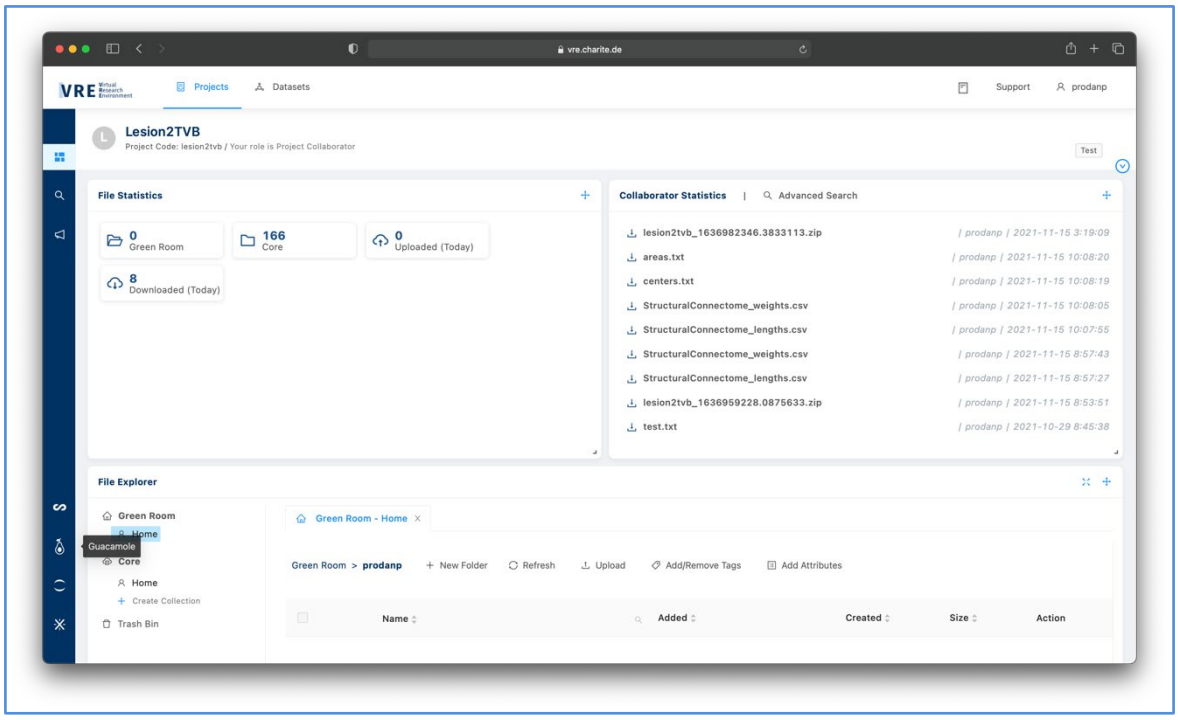

Figure 15: HDC offers a remote desktop gateway via "Guacamole"

In the Core VM, from the Terminal, the user can download the Lesion2TVB TVB-ready data via the CLI tool. Below is an example of using the CLI to download a connectome from the Lesion2TVB project (see Figure 16).

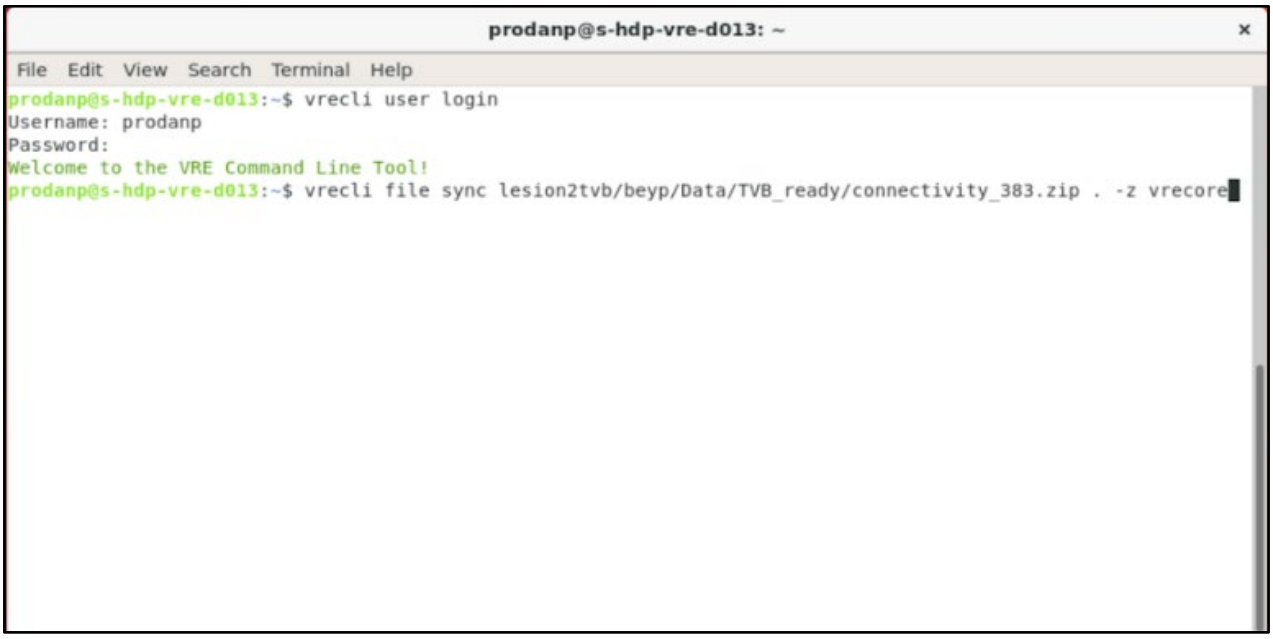

Figure 16: Using CLI to load results from HPC to the HDC Core VM

To access the TVB Web GUI available in the HDC, the user opens the following URL in a local web browser: https://vre.charite.de/tvb. This launches the TVB Web GUI, first redirecting the user to the login page. The user signs on using their VRE credentials. Only users with a VRE account can log into this instance of TVB Web GUI (see Figure 17).

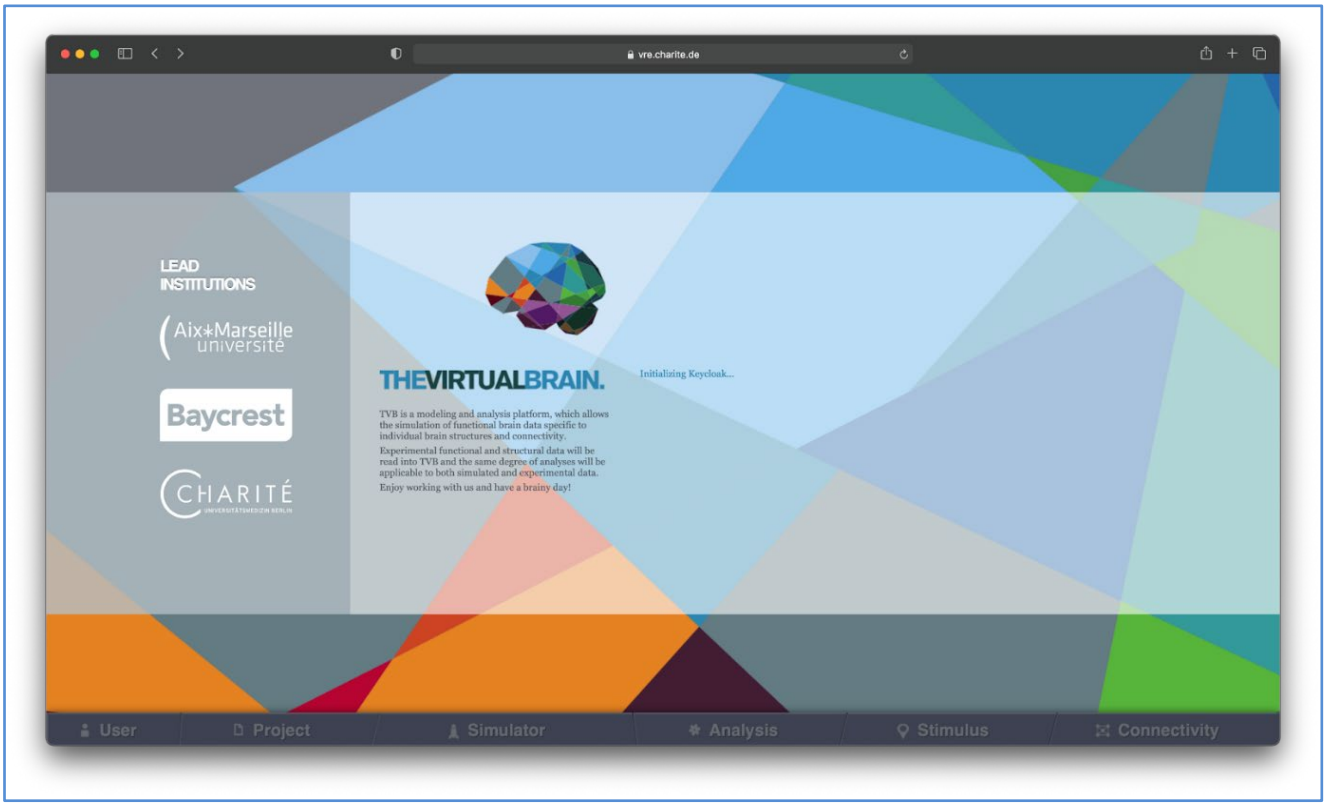

Figure 17: Accessing TVB login page via HDC

2.2.6.3 Jupyter Notebooks

Within the VRE portal, users can also access the JupyterHub service, to access TVB via its command line interface (see Figure 18).

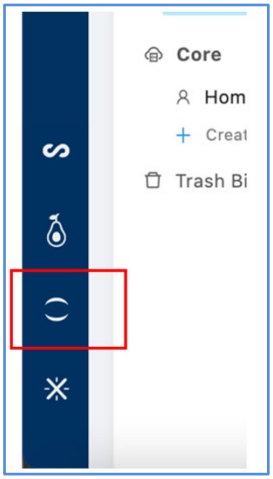

Figure 18: JupyterHub can be launched from within secure HDC

It is possible for users to create their own Jupyter kernels that are persisted in the VRE for later usage. This way, the user has full control of upgrading/adding libraries in the long term. Libraries installed using the 'pip' Python package installer in a Python virtual environment (venv) can persist in JupyterHub and can be accessed in future sessions without requiring re-installation. Following the instructions provided on VRE User Manual wiki[20], the user can carry out all the steps needed to create a virtual environment on their JupyterHub workbench.

[20] https://vre.charite.de/xwiki/wiki/vrepublic/view/Main/user_guide/Analyzing Data/JupyterHub/

In the image below (see Figure 19), the user has already created a “tvb-env” kernel where TVB is already installed. Thus, the next step is to create a Jupyter notebook that uses that kernel.

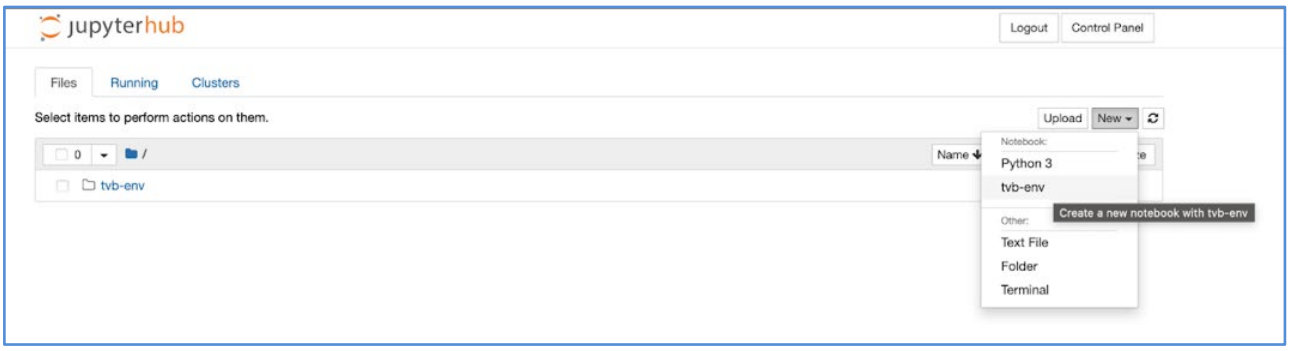

Figure 19: Running pre-installed software via JupyterHub

The user can check whether TVB works as expected, by doing a simple import (see Figure 20).

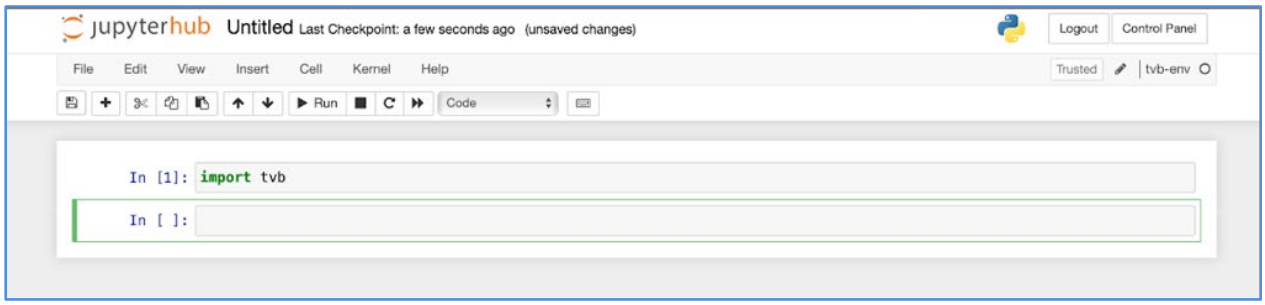

Figure 20: Importing TVB

The VRE CLI tool can be used to bring the TVB-ready data from VRE Core storage into the current user folder of JupyterHub. Then, the user can start configuring and running simulations (Figure 21).

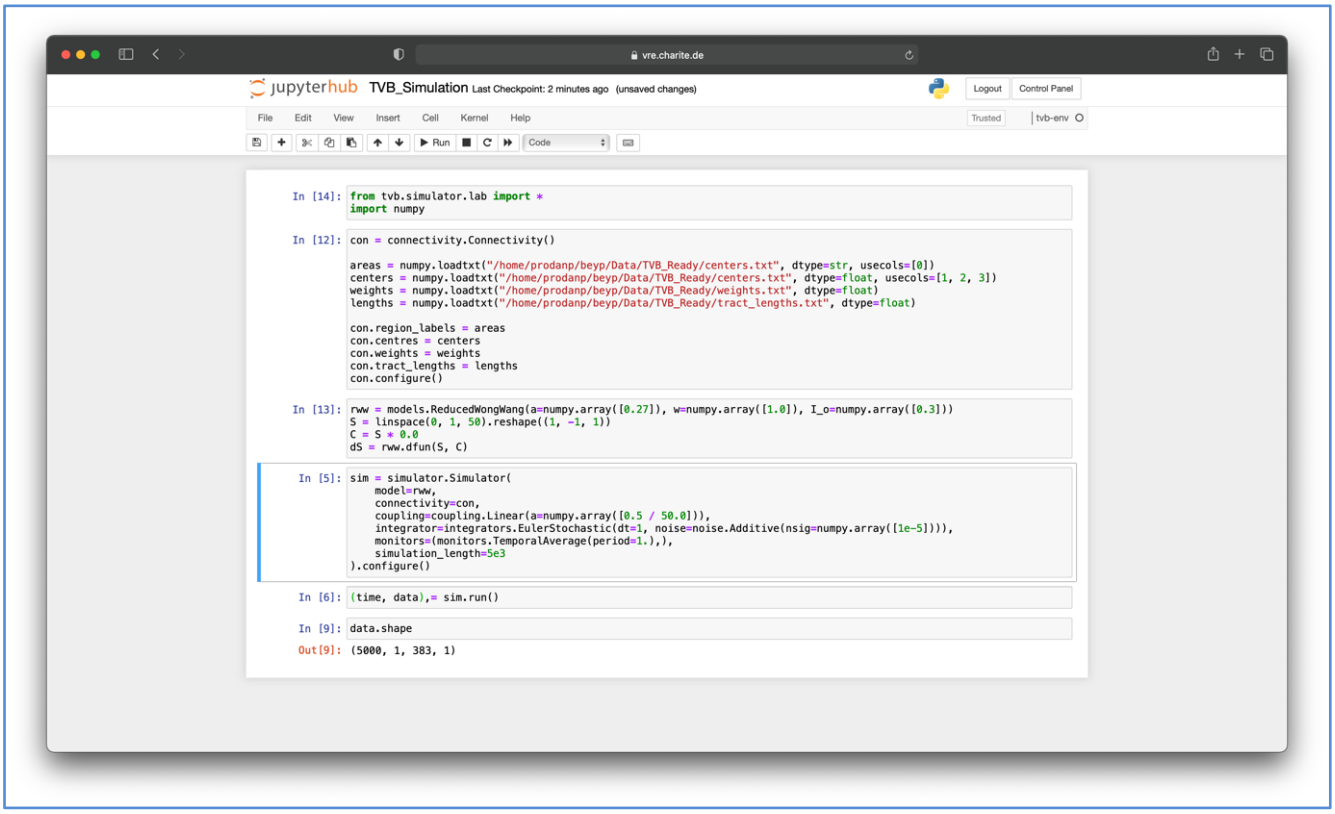

Figure 21: Configuring simulations inside HDC

2.2.7 Onboarding HDC Node at Charité Berlin

The secure Virtual Research Environment at Charité is discoverable via EBRAINS but also via other official research data repository registries such as the EOSC Marketplace[17] (see Figure 22), German Research Foundation Research Infrastructure resources registry (see Figure 23), Registry of Research Data Repositories (see Figure 24) and FAIRsharing.org (see Figure 25).

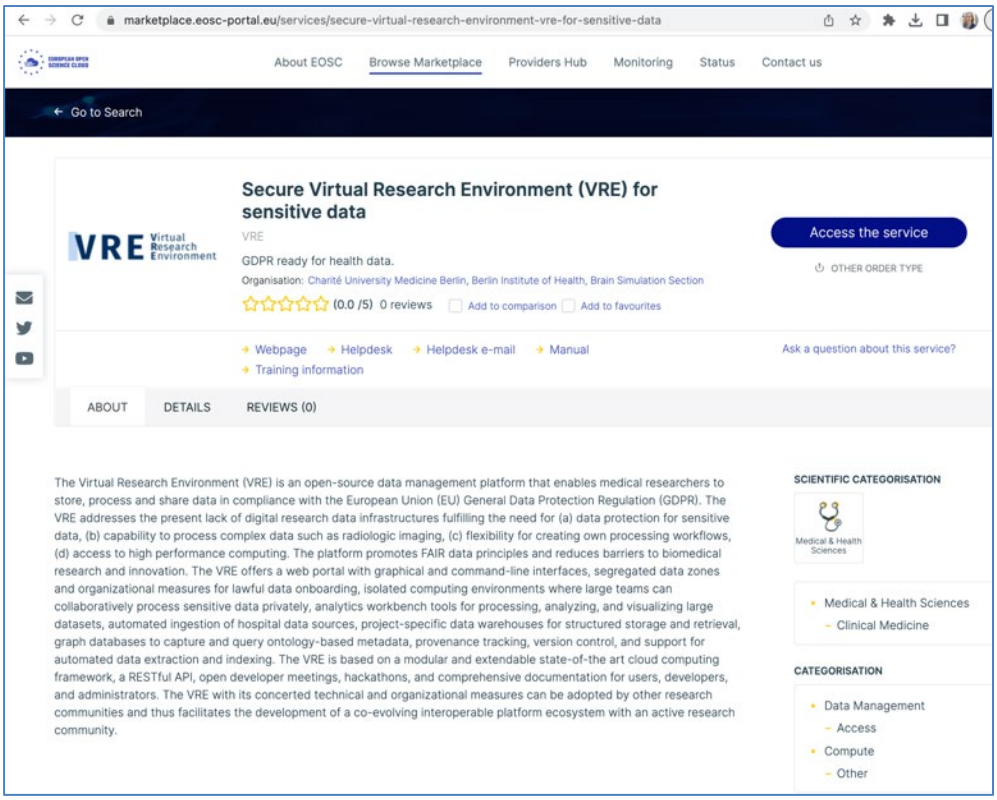

Figure 22: EOSC marketplace promotes HDC's VRE at Charité

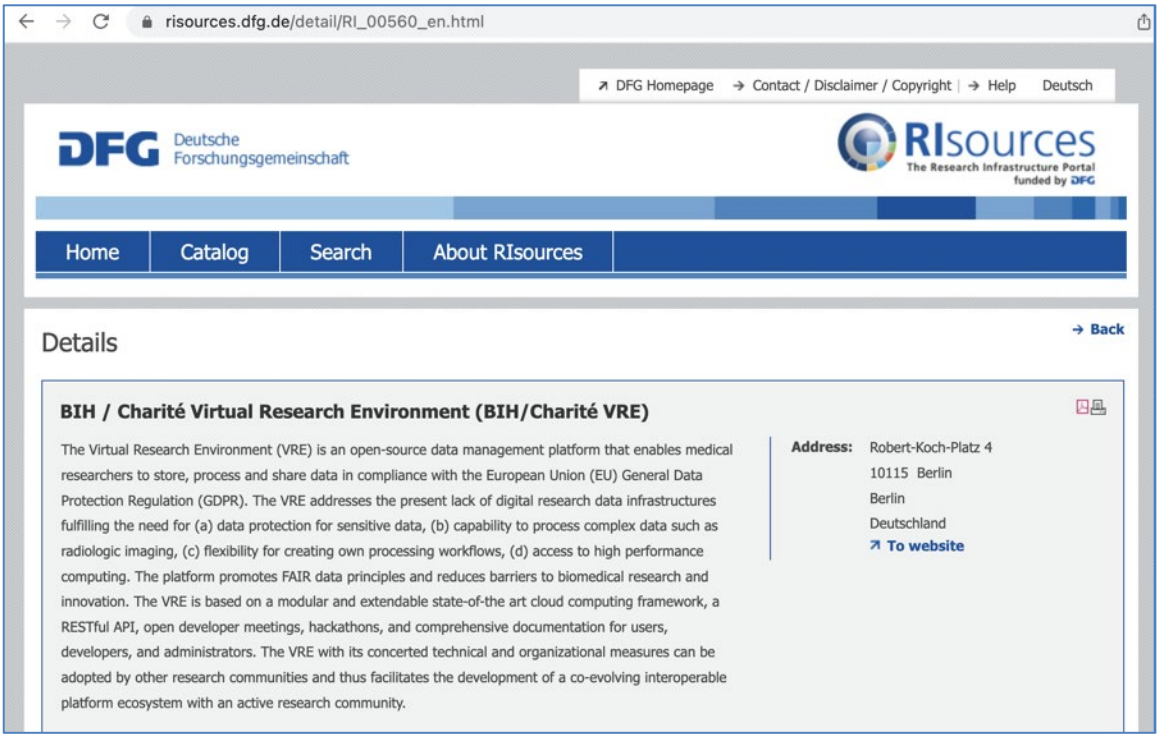

Figure 23: German Research Foundation (DFG) lists HDC's VRE

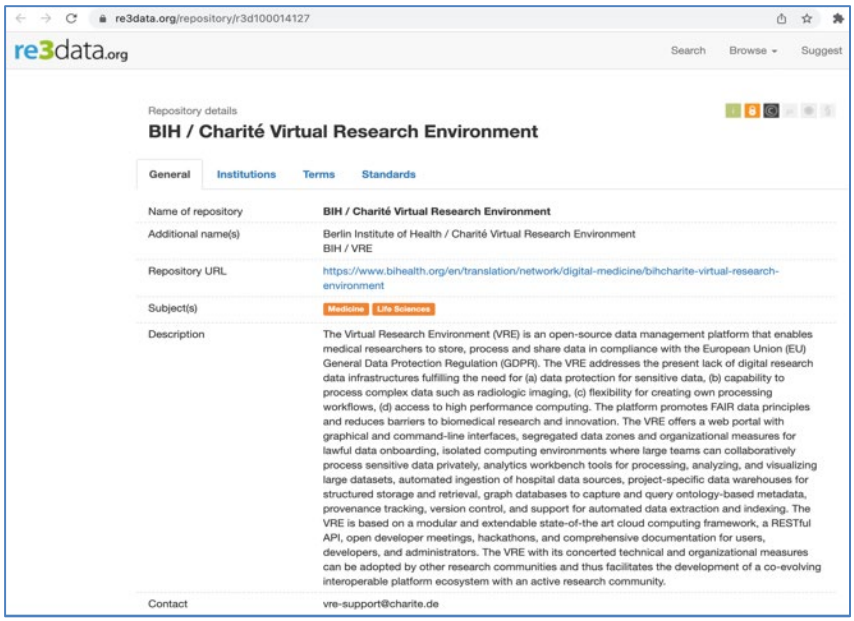

Figure 24: Registry of Research Data Repositories

Registry of Research Data Repositories re3data[21] lists the VRE

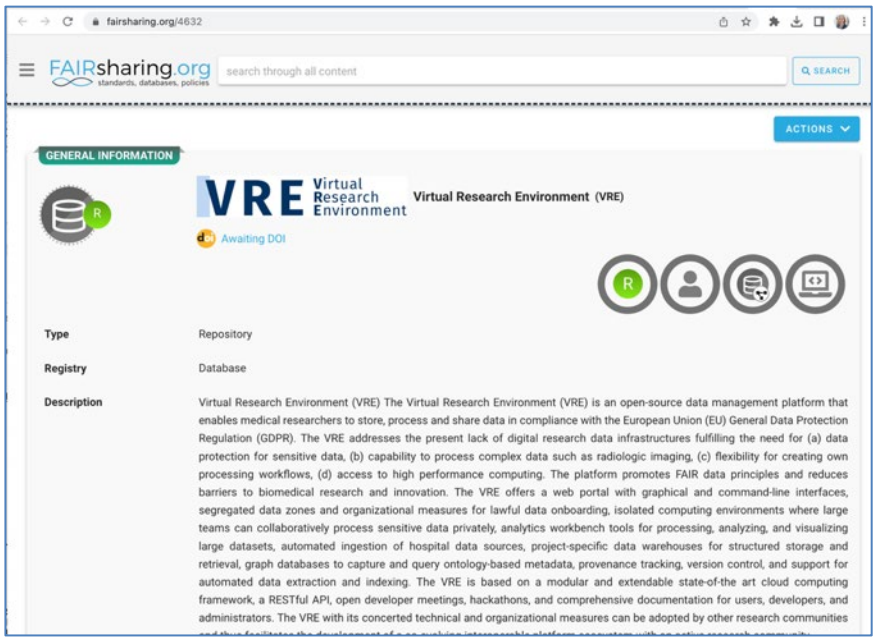

Figure 25: FAIRsharing.org lists[22] HDC's secure VRE for its users

With respect to Nature Scientific Data repositories[23] we presently prepare the inclusion of HDC's secure VRE under the category "health repositories".

HDC's VRE User onboarding includes:

- Signing of a data processing agreement
- Registering an Active Directory entry
- Creating VRE Platform account

[21] https://www.re3data.org/repository/r3d100014127
[22] https://fairsharing.org/4632
[23] https://www.nature.com/sdata/policies/repositories#healthsci

2.3 Governance and Processes

EBRAINS HDC as part of its data protection concept defines organisational and administrative processes. Liabilities of all stakeholders (IT centres, service providers, users) are defined by access policies, and agreements such as terms of use, data processing agreements, and service level agreements. All data sets discoverable in the EBRAINS HDC have their specific terms of use and license policies. The HDC Use and Access Committee is chaired by HDC Project Lead Prof. Petra Ritter (Charité). GDPR requires a data protection impact assessment and consultation of the responsible Data Protection Officers (DPOs) for automated processing activities of sensitive health data. This implies that each use case involving health data requires their own risk assessment and data protection concept evaluation. Since the lawful basis for processing needs to be assessed for each individual use case, EBRAINS HDC provides templates (see this example[14]) for data processing agreements and supports users by also providing templates for the comprehensive documentation involved in conducting a data protection impact assessment. HDC establishes efficient processes that make it simpler for researchers to comply with GDPR and to demonstrate GDPR compliance.

2.4 GDPR Compliance Assessment

The HDC node VRE at Charité has been assessed with respect to GDPR compliance. This involved the preparation of a risk assessment, a data protection impact assessment (DPIA) and consultation with the DPO to ensure that GDPR requirements are met to allow EBRAINS HDC users of the satellite node at Charité to demonstrate compliance with GDPR. The same type of GDPR compliance assessment is planned for additional nodes of HDC. In preparation of this evaluation, the HDC team collaborates with the EBRAINS Information Security Officer who is attending the weekly HDC Technical Coordination meetings.

Hospital IT centres such as at Charité adhere to specific national regulations and security standards. Certifications according to ISO standards, such as ISO 9001 for quality management and ISO 27001 for managing information security, facilitate GDPR compliance assessments considerably, as these confirm adherence to certain standards.

We are in the process of evaluating the certification status of ETHZ involved IT centres (CSCS, LeoMed). The GDPR assessment involves the relevant DPOs and established law firms with proven expertise in this domain.

The maintenance of IT infrastructures with high protection requirements involves processing agreements, and data access will be strictly regulated. Any access to systems will be recorded and thus made auditable. While high protection standards are in place or will be achieved at some of the involved IT centres, others will conclude service and processing agreements to process data at highly protected centres as required. This will allow less protected facilities to limit data holdings to encrypted data with decryption allowed only in sandboxed, fully isolated environments.

The HDC node VRE operated at the Charité underwent a comprehensive external independent GDPR compliance audit which confirmed the GDPR readiness of the infrastructure.

In October 2022 the European Data Protection Board (EDPB) endorsed a GDPR certification scheme for the first time: The Europrivacy certification now hands out the European Data Protection Seal to certify GDPR compliance (see Figure 26). The certification criteria were drafted by the European Center for Certification and Privacy. Europrivacy certification is registered in the public register of certification mechanisms and data protection seals and marks.

We presently are in process of achieving additionally to the already successful external GDPR audit a European Data Protection Seal to facilitate researchers/users to demonstrate compliance when using the HDC VRE node for their processing of sensitive data. Similarly, other HDC nodes will undergo Europrivacy certification.

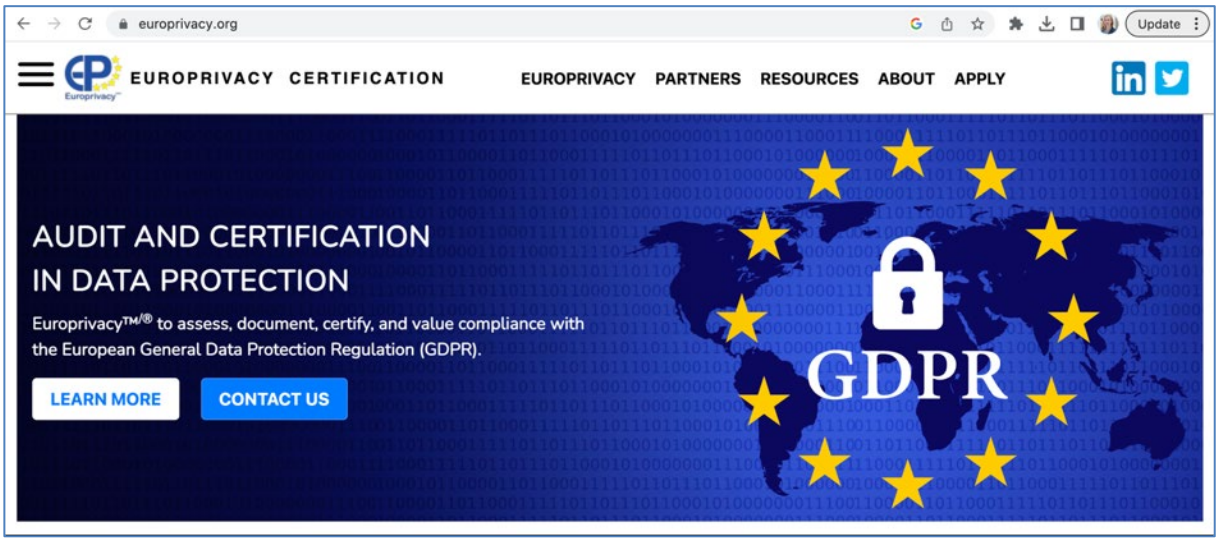

Figure 26: Europrivacy certification hands out the European Data Protection Seal

2.5 GDPR Use Cases

2.5.1 Human Connectome Project Data

A first Use Case of the Health Data Cloud has been published in the journal Nature Communications (P4047)[24] (see Figure 27). The title of the study is "Learning how network structure shapes decision-making for bio-inspired computing". Health data from 650 persons provided by the Human Connectome Project were used for this study. The processed data derivatives can be discovered now via the EBARINS KG and they can be re-used after the legally obligatory HDC onboarding process has been successfully concluded.

In this article the use of the HDC infrastructure is acknowledged as follows: "This work was supported by the Virtual Research Environment at the Charité Berlin – a node of EBRAINS Health Data Cloud."

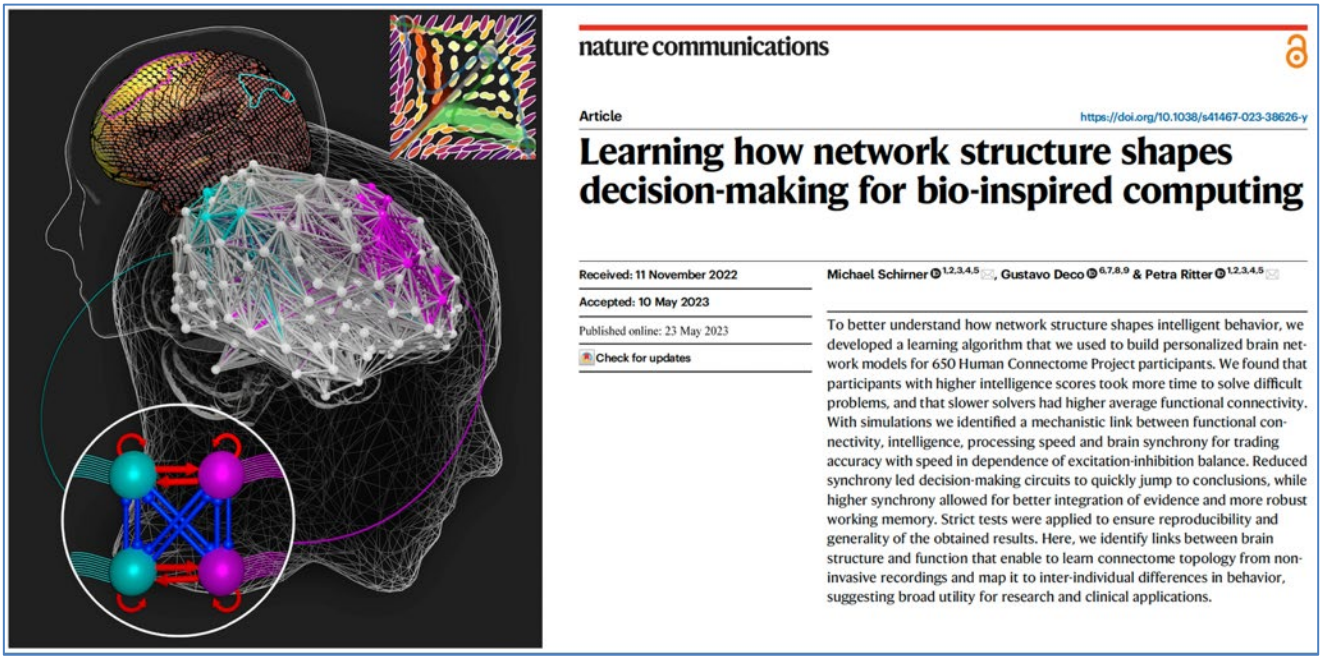

Figure 27: Publication of an HDC Use Case in the journal Nature Communications

[24] https://www.nature.com/articles/s41467-023-38626-y; https://doi.org/10.1038/s41467-023-38626-y

2.5.2 Human Longitudinal Stroke Cohort Data

Workflows for batch processing of stroke imaging data

Processing ischemic stroke magnetic resonance imaging (MRI) data can be susceptible to lesion-based abnormalities. CHARITE developed and validated the Lesion Aware automated Processing Pipeline (LeAPP; see Figure 28) incorporating appropriate correction methods and significantly improving resulting volumetric and connectomics measures compared to current standards in automated processing pipelines (Bey et al. 2023). The pipeline is containerized and runs fully automated within the HDC.

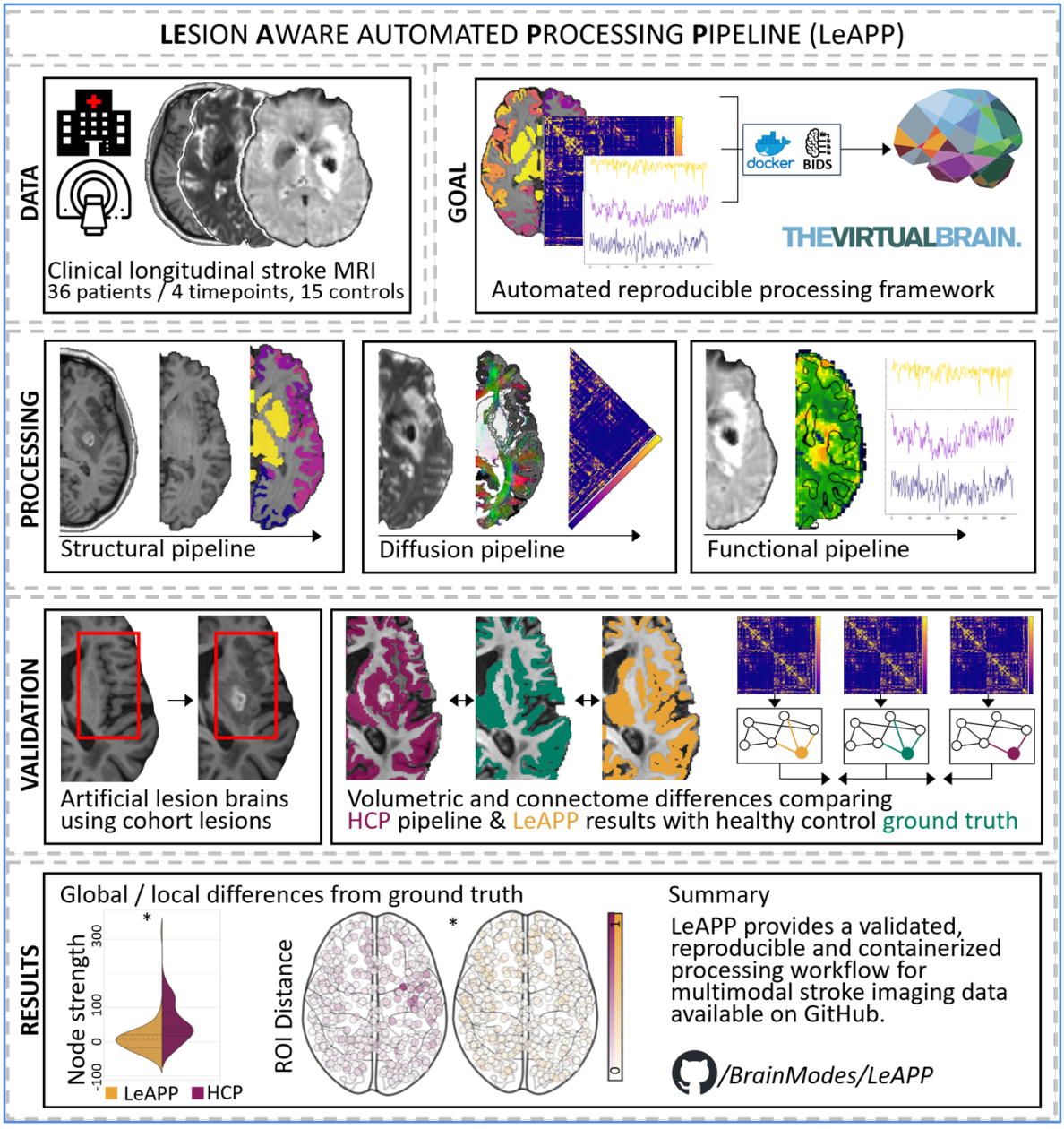

Figure 28: Automated processing pipeline for stroke imaging data

The containerized pipeline runs in the secure HDC and generates data derivatives for statistical and computational modelling (Figure from Bey et al. 2023)

In this article the use of the HDC infrastructure is acknowledged as follows: "This work was supported by the Virtual Research Environment at the Charité Berlin - a node of EBRAINS Health Data Cloud."

3. Health Data Cloud Community

HDC's VRE node at the Charité has not only a rapidly growing user community but importantly also supports the engagement of users and research software developers in the co-design and co-development of the platform. This model is to be adopted for the HDC which is to follow the principles of a community project. Various support mechanisms guarantee this continuous user support and troubleshooting, amongst others, regular bi-weekly meetings with the development and operations teams, which are open to the community to guarantee the most direct support and user feedback collection. The HDC/VRE with its concerted technical and organisational measures can be adopted by other research communities and thus facilitates the development of a co-evolving interoperable platform ecosystem with an active research community. The already operational and audited secure VRE node at Charité data centre is a full stack automated *infrastructure as code* (IaC) equipped with detailed deployment and operation documentation.

4. Health Data Cloud Functionality Under Development

4.1 HDC Federated Architecture

HDC's full services are partially still under development. The HDC development started in January 2022 after the HDC consortium led by Charité had won an open call. The full HDC services will comprise a federated peer-to-peer (P2P) network, which is a network of interoperable equally privileged nodes that all exist autonomously (see Figure 29). Each node is autonomous, e.g., CHARITE VRE, UiO TSD, and HBP's EBRAINS RI all serve their individual user groups with an individual user facing frontend. Nevertheless, the HDC P2P framework connects these nodes and enables the exchange of data, metadata, and code between nodes. Thus, metadata from data stored at CHARITE VRE or at TSD can be sent to the EBRAINS Knowledge Graph and published after passing EBRAINS quality control and curation process. Data in distributed data centres thus become discoverable centrally via a single search interface. After discovery users can send access requests which are delivered to the respective data controllers. Upon approval by data controllers, data (minimized for the specific processing purpose in accordance with GDPR) are made accessible to the new users who conclude a sharing/processing agreement, conduct a DPIA and thus become lawful controllers or processors.

The EBRAINS HDC will consist of a scalable, federated network of interoperable technology stacks (referred to as nodes) sharing a standard architecture, including a central node deployed on the EBRAINS primary infrastructure and remote nodes deployed at Charité and Oslo University Hospital. The EBRAINS HDC builds on an already established GDPR-compliant audited service at the Charité University Medicine Berlin, the Virtual Research Environment (VRE), which is open-source licensed under EUPL and by design developed to be EBRAINS interoperable.

The conceptual architecture of the EBRAINS HealthDataCloud is a federation of interoperable nodes including a central node and an expandable set of satellite nodes (see Figure 30). A node represents a location where research data are collected or processed. Nodes communicate over REST-based API endpoints to exchange data and metadata and to access services. The central node is deployed at the EBRAINS RI primary data centre (ETHZ-CSCS) alongside but isolated from other EBRAINS services. This deployment replicates the functionality of the Charité VRE, enabling research consortia to manage the complete lifecycle of their studies, from data collection to analysis and sharing, in accordance with data protection regulations. Other researchers will use the central node for specific needs such as accessing HPC resources or the Knowledge Graph. Satellite nodes represent hospitals, research institutions and computing centres, typically hosting a digital environment for collecting and processing sensitive data. Research teams store, curate, analyse and share data initially within this environment, while publishing metadata to the EBRAINS Knowledge Graph to make their projects and datasets discoverable and shareable, and transferring permitted data to other nodes for additional processing and broader sharing.

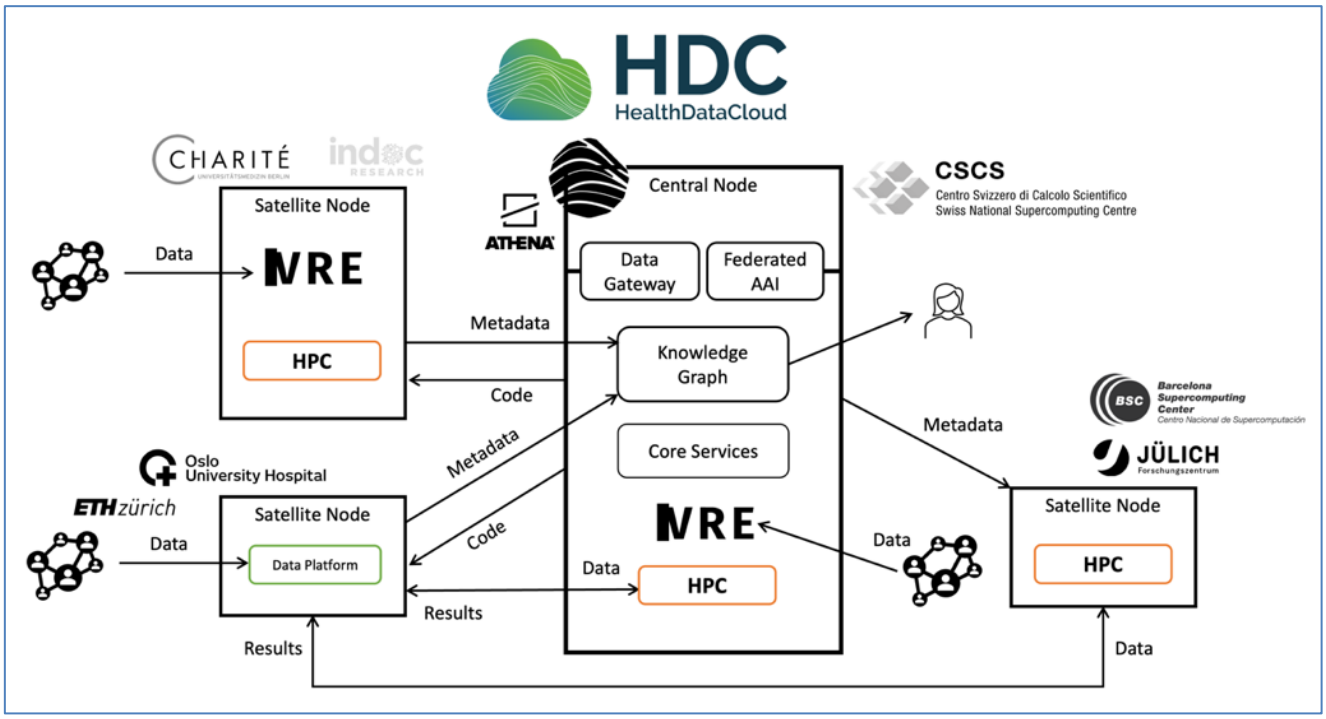

Figure 29: EBRAINS Federated Health Data Cloud

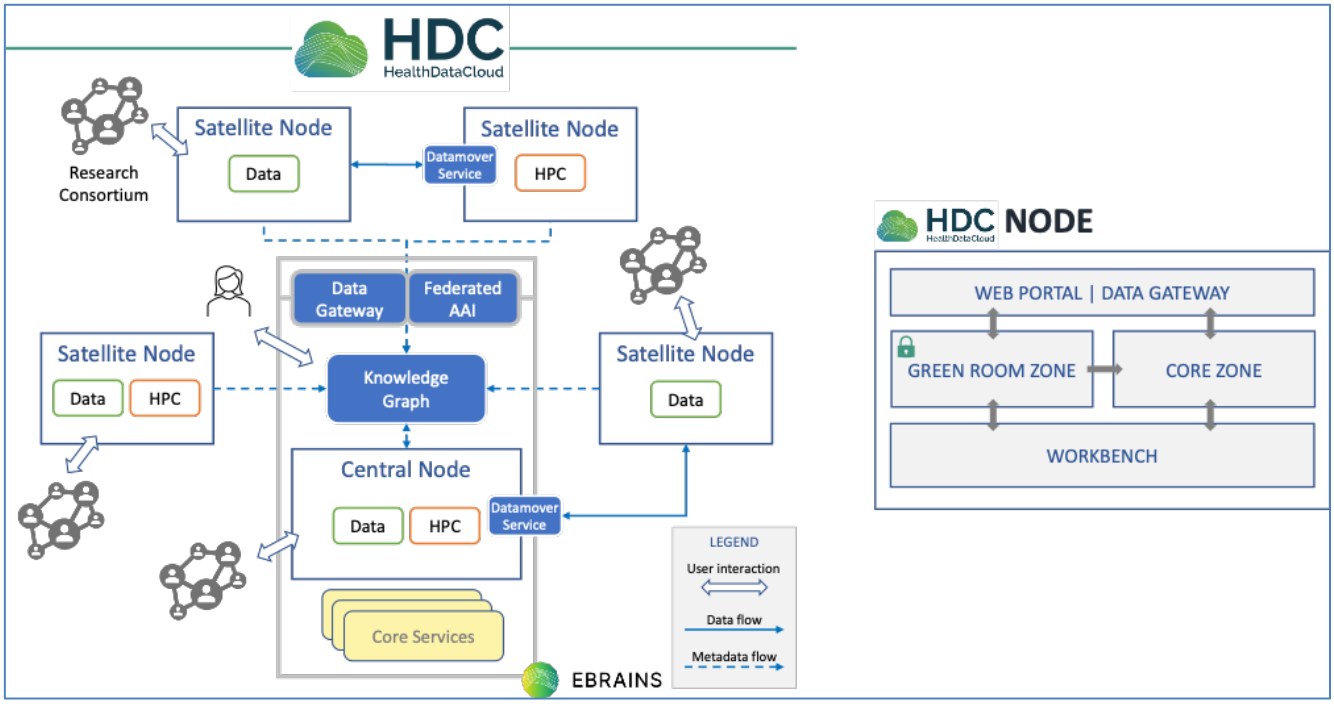

Figure 30: HDC network (left) and architecture (right) of a node

4.2 Central Node at EBRAINS

The HDC central node is deployed at the ETHZ supercomputing infrastructure at CSCS and provides a complete data management platform founded upon the VRE at Charité[25] and Pilot technology[26]. Therefore, it shares most of its core features with the GDPR-compliant satellite node VRE deployed at Charité. In addition, the HDC central node comprises additional features that enable the full integration with EBRAINS infrastructure and services. This concerns the physical architecture deployment at the Fenix centre CSCS (network, server/virtualization, storage, security controls), where the systems of the central node have been deployed in a dedicated subnetwork within the CSCS network, using virtual machines (VMs) and storage provisioned and managed with OpenStack. The HDC's central node also provides Authentication and Authorization Infrastructure (AAI) integration, including Fenix federated AAI, services integration including Knowledge Graph, and HPC integration at CSCS. For instance, users can login to the central HDC node with their EBRAINS credentials, or upload metadata to dedicated and managed spaces in the EBRAINS Knowledge Graph directly via the user interface of the HDC platform (see Figure 31). The HDC central node has been delivered as Minimum Viable Product (MVP) that presently undergoes user acceptance testing and is accessible in the HDC production environment: https://hdc.humanbrainproject.eu/login

The deployment on the EBRAINS RI and the development of the HDC central node software are aligned and compliant with the EBRAINS Software Delivery Guidelines. Similar to Charité's GDPR audited VRE, the HDC central node at CSCS includes strategies like automation, multi-environment, and monitoring. It utilizes microservice architectural pattern paired with Kubernetes-based container orchestration, as well as VM-based and CI/CD delivery technologies.

At a high level, the architecture of the central node comprises a web portal, platform services, and infrastructure elements (see Figure 8). The web portal enables access to data and resources, and administration of projects and users. Platform services provide access control, account management, data operations, data lifecycle management, provenance, processing pipelines, notification, and workbench tools. Infrastructure elements include middleware services that support, manage, and monitor platform services and interface with computing and storage resources. A key feature of the VRE and the HDC central node is the Green Room, a data processing zone protected by network and access control policies designed for data minimisation of sensitive data in compliance with GDPR. In both the VRE at Charité and the HDC at CSCS authorized researchers can access Green Room data and use Workbench tools to securely curate and prepare datasets (e.g., by applying pseudonymization) for further analysis and sharing in the Core Zone that is also protected and ready for managing still sensitive data containing identifying and health information.

Sensitive data processed on HPC will be encrypted at rest and sandboxed before they are decrypted as described in Schirner et al. 2022 (P2973)[27], and implemented in the TVB workflow on EBRAINS[28]. HDC VRE and HDC enhance the abstraction of data centre infrastructure from the application software layer. The deployments at Charité and ETHZ-CSCS serve as models for possible future deployments at other data centres. Importantly, while the VRE node at Charité has an elaborated information security and data protection concept that has been audited and that contains continuous vulnerability monitoring and maintenance, the present deployment of HDC at CSCS lacks currently the demonstration of compliance with information security standards such as ISO 27001 and Data Protection laws such as GDPR. In contrast to the Charité Hospital Data Center, which is a critical infrastructure undergoing certification renewal every two years, CSCS is a Supercomputing Center with lower information security and privacy standards. Building on their previous successful development of GDPR compliant processes at CSCS (Schirner et al. 2022), the teams at Charité, ETHZ, CSCS and LeoMed work jointly on developing protection measures with a shared liability model. For instance, one solution currently being explored is that data will be encrypted permanently when stored at CSCS, while processing (and thus decryption) takes place exclusively at

[25] https://github.com/vre-charite
[26] https://www.indocresearch.org/pilot
[27] https://doi.org/10.1016/j.neuroimage.2022.118973
[28] https://www.ebrains.eu/modelling-simulation-and-computing/simulation/whole-brain-simulation-2/

secure HPC centres certified for GDPR compliance such as the ones at Charité or LeoMed that can ensure sufficient isolation and protection from unauthorised access during processing.

Alternatively, in cases of the need for HPC resources outside a critical infrastructure as provided for HDC's VRE at Charité, we explore solutions where sensitive data are transferred to an isolated sandbox for processing as described in Schirner et al. 2022. The sandbox is a temporary environment that outputs only encrypted results using the authorised user's private key, which deletes all non-encrypted data when expiring.

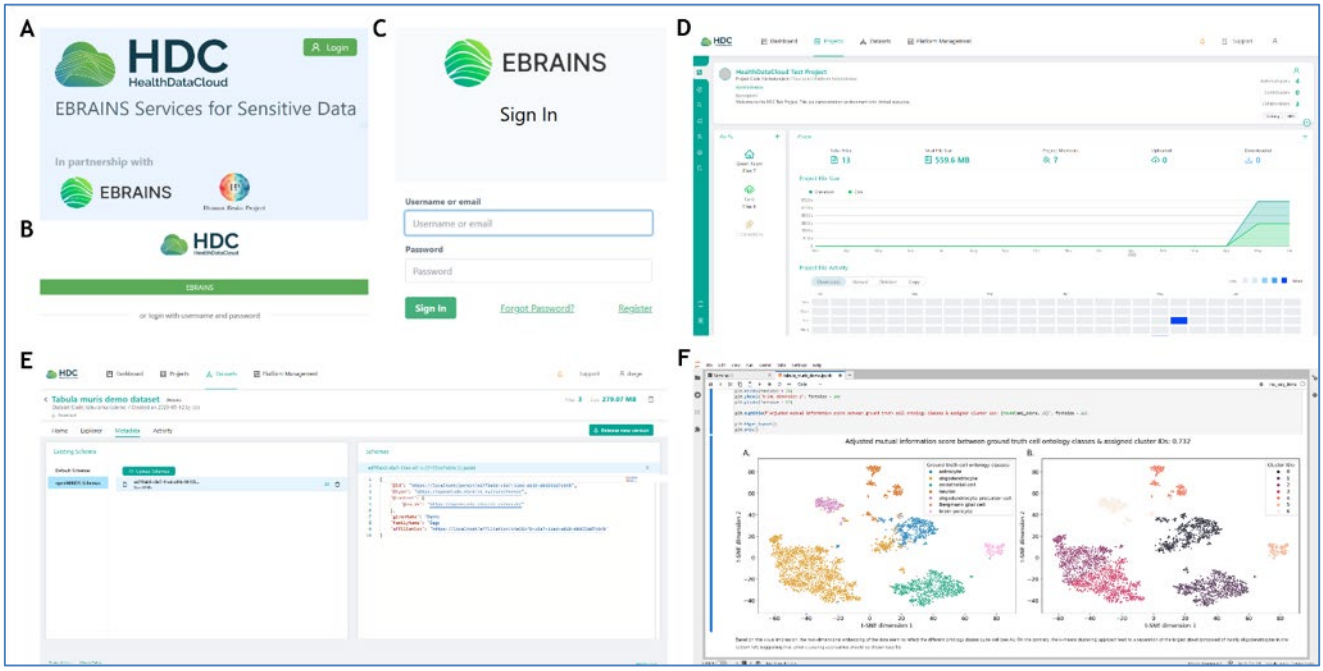

Figure 31: Impressions of HDC central node MVP deployed at CSCS

A-C: Login process to the HDC central node via the landing page (A), which is integrated with the EBRAINS Identity and Access Management (B and C).

D: After logging into the Platform using an EBRAINS account, users can, for instance, upload and share data with other project members, or gain a general overview of the current Project status using the Project Canvas.

E: Users can utilize the dedicated Datasets feature of the HDC central node to prepare Datasets for sharing and publication and future functionality of HDC will enable users to upload metadata as openMINDS JSON-LD files to the centralized EBRAINS Knowledge Graph.

F: As in the VRE at Charité, users in the HDC central node also have direct access to Workspace tools, such as JupyterHub, to leverage computing resources to analyse their data within the dedicated platform environment.

4.2.1 AAI and Data gateway

As part of the data protection concept and to comply with GDPR, account approval for the EBRAINS HDC SSD will need to be more stringent than current processes, by requiring applicants, including existing EBRAINS users, to provide additional information to validate their identities (as is the case for the Charité VRE). The authentication service has been implemented by federating the SSD identity provider (IdP) based on Keycloak with the EBRAINS AAI (also based on Keycloak), providing single sign-on (SSO) across EBRAINS HealthDataCloud and other EBRAINS services. This federated AAI will also support multi-factor authentication using the common practices of One-time Password (OTP). Should EBRAINS implement AAI federation with external identification services with a suitably high level of assurance, users could potentially authenticate with institutional credentials.

The federated AAI will also be used to authorise access to HDC projects and datasets, and to associated metadata stored in the EBRAINS Knowledge Graph. The API gateway will broker all service requests, including requests for upload/download of datasets. The API gateway first authenticates each request through the Keycloak IdP over OIDC, then routes the request to the target service. In the case of data access requests, upon successful authentication, requests are routed to a platform-

level encryption/decryption service. The combination of these services (API Gateway, IdP, encryption/decryption service) essentially represents an SSD Data Gateway service.

4.2.2 Integration with EBRAINS Core Services

Researchers publish original data, derived data, models, and software in the EBRAINS Knowledge Graph to make them readily discoverable and accessible. This is achieved by curating datasets based on EBRAINS guidelines and submitting curated datasets and related metadata (e.g., based on the openMINDS JSON-LD format) for Knowledge Graph ingestion. Web tools facilitate this process and researchers can engage the EBRAINS curation team for support and for involved needs such as atlas registration. In the context of the EBRAINS HDC SSD, personal data must remain in protected environments (e.g., hospitals), and even sharing of metadata must be controlled. To support discoverability of sensitive data stored on the EBRAINS HDC, we will coordinate with the EBRAINS Data and Knowledge services development team to extend the metadata schema template with new elements pertinent to the protection of sensitive data, including sensitivity (e.g., direct identifiers, pseudonymous), visibility (open versus access-controlled), residency (permitted storage/processing locations), and others as needed. As it is done currently on EBRAINS, SSD users will submit metadata through the Knowledge Graph API or web interface (serving different user preferences and use cases) for ingestion into the Knowledge Graph.

To enable HDC users to submit metadata directly to the Knowledge Graph from the HDC central node, we have developed a workflow to create a Collab in the EBRAINS Collaboratory and a space in the Knowledge Graph corresponding to each HDC project. Currently, user permissions to access their metadata in the Knowledge Graph as administrator, editor or viewer are managed through the Collaboratory. Since the EBRAINS and HDC AAI systems are federated, only authenticated users who are authorized members of a project are allowed to submit metadata and to discover datasets associated with that project.

We have developed a new set of HDC APIs that work in concert with the Collaboratory and Knowledge Graph APIs[29,30] to programmatically execute the above workflow. In addition, new components have been added to the user interface of the platform to support the interactive submission of metadata from the HDC to the Knowledge Graph. Alternatively, users also have access to a Command Line Interface within the HDC Workbench (e.g., in Linux terminals) to submit their metadata over the Knowledge Graph API.

The Knowledge Graph serves HDC users as a discovery tool for finding data generated within their research projects or consortia, and it provides an avenue for public discovery of health data sets via non-sensitive metadata and eventual sharing of sensitive data for research purposes according to prevailing law. Furthermore, it enables researchers to launch HPC computations on remotely stored data, as explained in Section 2.2.6.

4.2.3 Co-Location and Federation of Data and Computing

Among the fundamental design principles of the EBRAINS HDC is the co-location of storage and computing resources and the ability for “code to visit data”. This is first enabled by the network architecture of the HDC, by which data can be collected and computed at multiple locations according to data residency rules. As discussed above, the EBRAINS Knowledge Graph is critical to enabling this principle, acting as an index for the location of all datasets distributed across the network. Within the Knowledge Graph each dataset stored in an HDC node will be annotated with metadata elements related to data sensitivity, data residency, storage location, and visibility. These elements not only make the dataset discoverable but can also enable EBRAINS HDC users to analyse the dataset by dispatching a containerized pipeline to the data centre holding the data. Invocation

[29] https://wiki.ebrains.eu/bin/view/Collabs/the-collaboratory/Documentation%20Wiki/API/
[30] https://core.kg.ebrains.eu/swagger-ui/index.html

of remote processing jobs has already been implemented for the TVB workflow on EBRAINS, which uses the UNICORE API to submit data and code to supercomputers at HPC centres.

However, this workflow requires that data be stored locally within the EBRAINS RI. We will adapt this workflow for remote processing of sensitive data stored in SSD nodes, so that SSD users will be able to programmatically run and manage processing jobs from within SSD Workbench tools (e.g., within a Jupyter notebook). Processing is preferentially co-located with the data if adequate computing resources are available. If data must be transported to another storage system at the same location or different location, a request will be dispatched to a Data Mover Service or File Transfer Service, as implemented in the HDC (see Section 2.2.6). Permissible computing locations associated with the input data will be identified from corresponding Knowledge Graph entries. The request will typically include a dataset DOI or another appropriate identifier, the location of the input data, and compute job parameters. The Data Mover Service, protected by the SSD API Gateway, which connects to Keycloak through OIDC for authentication, will call a Data Retrieval Service - also protected by the API Gateway - deployed at the SSD node hosting the requested data. The Data Mover Service is self-contained and includes a group of microservices and an object storage backend to enable encryption at rest. Data are immediately encrypted by an object encryption key (OEK) once it is received by the Data Mover Service. Data in-transit between the SSD node and the Data Mover service is protected by TLS. After the data is landed by the Data Mover server, another Data Mover API will leverage the UNICORE client API to initiate a job submission on the user's behalf on the HPC resource, following the already established key exchange process of the TVB workflow on EBRAINS in which generated keys are automatically encrypted at rest and only decrypted for decrypting processing results returned from the HPC.

This object-level encryption solution is highly performant and infrastructure-agnostic as it does not rely on proprietary encryption technology on filesystem or lower-level blocks. It uses an industry-grade key management system (KMS), e.g., HashiCorp Vault, to manage the master key, from which a key for encryption key (KEK) is derived to en/decrypt the final OEK. The KEK always resides in memory and the OEK is always saved in its encrypted form as part of the object metadata. This feature will build on the implementation already in place at HBP, including the use of encryption at rest and sandboxing techniques coupled with secure key exchange to protect data being processed in shared HPC environments. This also provides the motivation to scale the EBRAINS HDC network to new satellite nodes holding data or HPC resources (or both), with Knowledge Graph integration enabling authorized researchers to seamlessly find and act on data across the network, subject to data residency constraints.

4.3 HDC Satellite Nodes

4.3.1 Charité VRE

Charité serves as an HDC satellite node with the full functionality of the existing VRE platform. VRE architecture and deployment is described in the VRE Whitepaper[25]. There are equivalent services between VRE and EBRAINS including VRE Portal, XWiki, JupyterLab, Knowledge Graph, and HPC integration. Charité provides storage, computing resources and IT personnel for VRE operation and expansion, including approximately 500 dedicated cores for platform VMs, and over 700 dedicated HPC cores.

The VRE platform at Charité is open source licensed under EUPL v1.2, a free software licence that was written and approved by the European Commission and consistent with copyright law in 27 EU member states. The licence is available in 23 official languages of the European Union. All linguistic versions have the same validity. It removes legal uncertainty as exists for instance for GNU by explicitly taking into account EU law, and hence products under EUPL license are accepted in the EU for Interoperable Delivery of European eGovernment Services to public Administrations, Businesses and Citizens[31] as is relevant when VRE and its derived products become infrastructure

[31] https://de.wikipedia.org/wiki/European_Union_Public_Licence

components of the European Health Data Space (EHDS), an emerging European framework for the management of health data in the EU. HDC coordinator Charité is further developing HDC's VRE technology stack with various partners as part of the novel Digital Europe co-funded (60 million EUR) world-class reference centre "Testing and Experimentation Facility for Health AI and Robotics (TEF-Health)"[6]. Thus, VRE HDC technology is in full compliance with Digital Europe regulations strictly enforced by the European Commission that ensure digital, technical, and data sovereignty of Europe.

4.3.2 LeoMed

The ID/SIS team of ETHZ supports the HealthDataCloud by creating a dedicated isolated tenant on the Leonhard Med platform[32].

The platform and its components (processes, applications and systems) have a very high level of protection (as defined in the Article 19 of the ETHZ Directive Information Security[33]) and is provided by the Scientific IT Services, at the IT Services ETH Zurich.

The dedicated HDC isolated tenant guarantees isolation at the levels of data, network and compute, and supports the necessary security controls to properly store, access and process confidential and strictly confidential data within the context of the HealthDataCloud project. The HDC tenant provides a dedicated login node, a SLURM-based HPC cluster system, a dedicated distributed filesystem with automatic backup and a collection of pre-installed application modules. Access to the HDC tenant has been provided to selected users to deploy and successfully validate the first HDC data analysis pipeline. Data transfer from the HDC's VRE will be performed as a pull mechanism from within the HDC's tenant using the VRE CLI.

Priority has been given to provide HDC users with an interactive access to the HDC tenant to facilitate development of code, installation and configuration of the data analysis pipelines as well as a supervised execution on selected data. A model where unsupervised data analysis pipelines execution is supported is still in discussion.

4.3.3 Oslo

AI-MIND is a large-scale EU project with a focus on integrating EEG data for dementia research. AI-MIND is using the University of Oslo (UiO) TSD[34] as a local SSD solution, to provide data storage and processing to clinicians and researchers in a secure environment. Figure 32 shows the architecture of the AI-MIND platform, which includes:

- File Staging Server, responsible for temporarily storing data files that enter and leave the platform,
- Data Pre-processing Server, hosting data extraction, transformation and loading (ETL) processes and feature extraction processes,
- ML Model Training Server, which hosts the Model Training Subsystem where ML models are trained, validated, and tested by data scientists,
- HPC Service, the high-performance computing facility (Colossus) hosted at SD and used to train complex deep neural network (DNN) models,
- AI-MIND Clinical Platform (Testing), a replica of the external Clinical Platform for integration and acceptance testing,
- Project File System, used for permanent storage of data and code,
- Database Service, the AI-MIND object-oriented PostgreSQL database, and

[32] https://ethz.ch/staffnet/de/it-services/katalog/server-cluster/leonhard-med-secure-scientific-platform.html
[33] https://rechtssammlung.sp.ethz.ch/Dokumente/203.25en.pdf
[34] https://www.uio.no/english/services/it/research/sensitive-data/about/index.html

- Project Management Server, a web server for managing the AI-MIND platform and project.

It is planned to enable AI-MIND interaction with EBRAINS through a data exchange mechanism thus making UiO TSD another satellite node of the EBRAINS HDC.

AI-MIND represents an important and common use case of a satellite node that is already equipped with a research data platform, but that requires the use of EBRAINS services to share data, access computing resources, and collaborate more broadly. This effort will thus demonstrate that existing data infrastructure such as AI-MIND can be integrated with the EBRAINS HDC, despite differences in platform architecture from the typical satellite node shown in Figure 30.

The AI-Mind/Oslo University Hospital team together with the Oslo-University-based EBRAINS team have performed a preliminary mapping of the needed adaptions and extension of the openMINDS metadata schema to support their AI-Mind use case. The plan is to formulate metadata in the openMINDS standard for the various AI-Mind data, which range from EEG and MEG recordings, socio-demographic-, cognition- and biomarker-data, which was and is collected through AI-Mind's clinical study. Furthermore, the AI-Mind team is in the process of formulating distinctions between sensitive and non-sensitive metadata from a data-sharing perspective inside the HDC. The work of expanding the openMINDS schema to the description of machine learning models (and deep learning models in particular) is also in the initial planning phase. Also, the AI-Mind team is in the planning phase of a "metadata translator" that would take AI-Mind metadata and convert it to the openMINDS schema.

Lastly, the IT department at the University of Oslo is currently in the planning phase of implementing a general meta-data sharing-service for the TSD infrastructure, which will allow for sharing such data with other services such as the HDC. The sharing of AI-Mind meta-data is planned to be a first use case. This solution will also provide the opportunity for a more robust and general way of exchanging (meta) data between the TSD infrastructure situated at the University of Oslo, any other relevant projects it hosts, and the HDC.

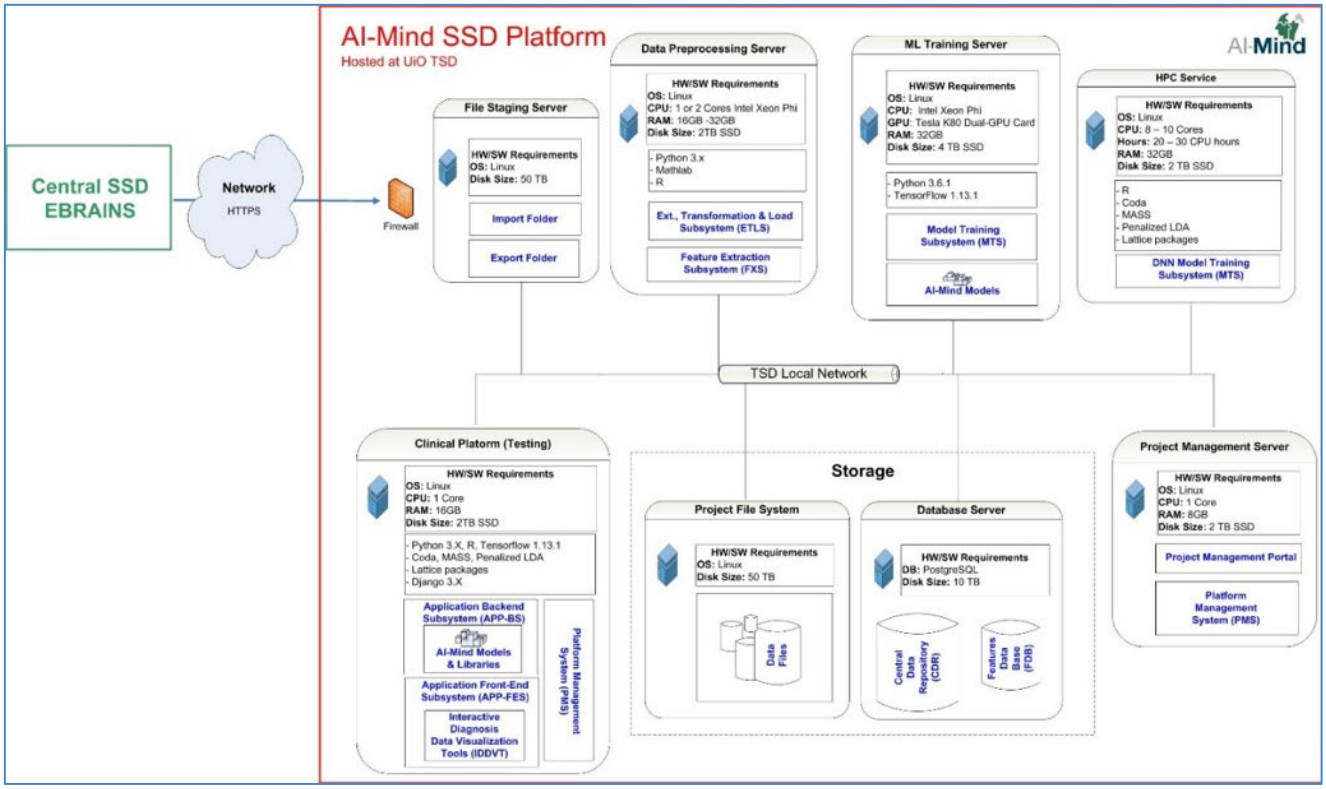

Figure 32: OUH Satellite Node physical architecture

4.3.4 KTH

The KTH team supported the alignment of the HealthDataCloud and the Fenix architecture as well as further evolution of the latter, e.g. integration of Key Management Systems. Furthermore, work has been performed on the evolution of the File Transfer Service (a first version of the service is presently being implemented) specifications to take requirements of HDC into account. This concerns the aspect of data encryption as well as the scheduling of the data flow when transferring data between different sites.

4.3.5 JUELICH

The JUELICH team at the Jülich Supercomputing Centre (JSC) has been closely engaged in the ETHZ-CSCS deployment of the HDC central node, participating in particular in the discussions of the data protection and security aspects. The technical and organisational measures in place for the JSC infrastructure have been evaluated and documented in preparation of the HDC deployment at JUELICH and a GDPR compliance assessment. HPC, cloud computing and storage resources including support are provided by JUELICH for the HDC deployment and the implementation of a backup solution for the HDC central node.

4.4 Sensitive Metadata

In HDC sensitive values of metadata will always be entered encrypted. A sensitive datum is never written in cleartext into e.g. an openMINDS JSON file, from the first moment the datum is entered into a computer. The data controller controls who can decrypt sensitive values via authentication. Decryption happens via public-key infrastructure and any client (like KG, VRE, other data platforms, etc.) having implemented a routine for decryption.

4.5 Use Cases under Development

The HDC development is driven by the requirements of use cases where mechanistic insight and predictive power is achieved through multimodal, multi-scale data integration through brain simulation and machine learning.

Despite HDC developments started only in 2022 – after an open call procedure – two of our use cases that build on EBRAINS simulation workflows of The Virtual Brain (TVB) platform and that used the EBRAINS HDC VRE node at Charité have recently been published[27].

Computational models and data of these use cases are discoverable via EBRAINS Knowledge Graph. The sensitive health data used in these publications are stored in HDCs VRE node.

4.5.1 Learning how network structure shapes decision making

785 processed imaging data sets of the Human Connectome Project with simulation ready data outputs discoverable via EBRAINS Knowledge Graph:

https://search.kg.ebrains.eu/?category=Dataset&q=schirner#88507924-8509-419f-8900-109accf1414b

These data have been processed and analysed for a study that has been published in Nature Communications (P4047)[35].

[35] https://doi.org/10.1038/s41467-023-38626-y

4.5.2 In silico DBS

Additionally, we used human digital twins to simulate the effects of deep brain stimulation (P2974, P3920)[36]. The corresponding computational model can be discovered via the EBRAINS Knowledge Graph:

https://search.kg.ebrains.eu/?category=Model&q=ritter#4efb127d-8393-4c97-b955-90f2c492b526

4.6 Key Performance Indicators

We provide validation of HDC, a dedicated cloud-based environment that leverages the potential of big data and HPC for understanding brain function and dysfunction and precision medicine.

In summary our use cases demonstrate unprecedented knowledge gain linking behaviour and clinics to large scale brain activity based on big data analytics. These success stories may serve as a strong motivator for the community to use EBARAINS Health Data Cloud for processing sensitive data.

Key Performance Indicators (KPIs) of the EBRAINS HDC are use- and output-centred:

1) Number of community members co-designing EBRAINS HealthDataCloud (weekly attendance of developer meetings): Goal N=15 by M1, gradually increasing to N=30 by M10, stabilizing around that number.
   We have met and exceeded this KPI from project start on (see Figure 33).
2) Number of beta users at the different nodes:
   Charité: N=15 by M3, N=25 by M10;
   Oslo: N=5 by M3, N=10 by M10;
   ETHZ: N=15 by M10.
   We have presently about 100 registered users at HDC's Charité VRE node.
   We currently onboard test users to HDCs central node that is now in production but lacks GDPR certification[4].
3) Number of returning users: MVP goes into production at M10, N=50; full product by M15, N=100. MVP of central node went in production in M16. HDC as a whole presently has 100+ users.
4) Use case published by M15: achieved[37]
5) Demonstrator: multi-media digital demonstration of the Prototype[38] (M10): achieved

[36] https://doi.org/10.1016/j.expneurol.2022.114111
[37] https://www.nature.com/articles/s41467-023-38626-y
[38] https://wiki.ebrains.eu/bin/view/Collabs/health-data-cloud

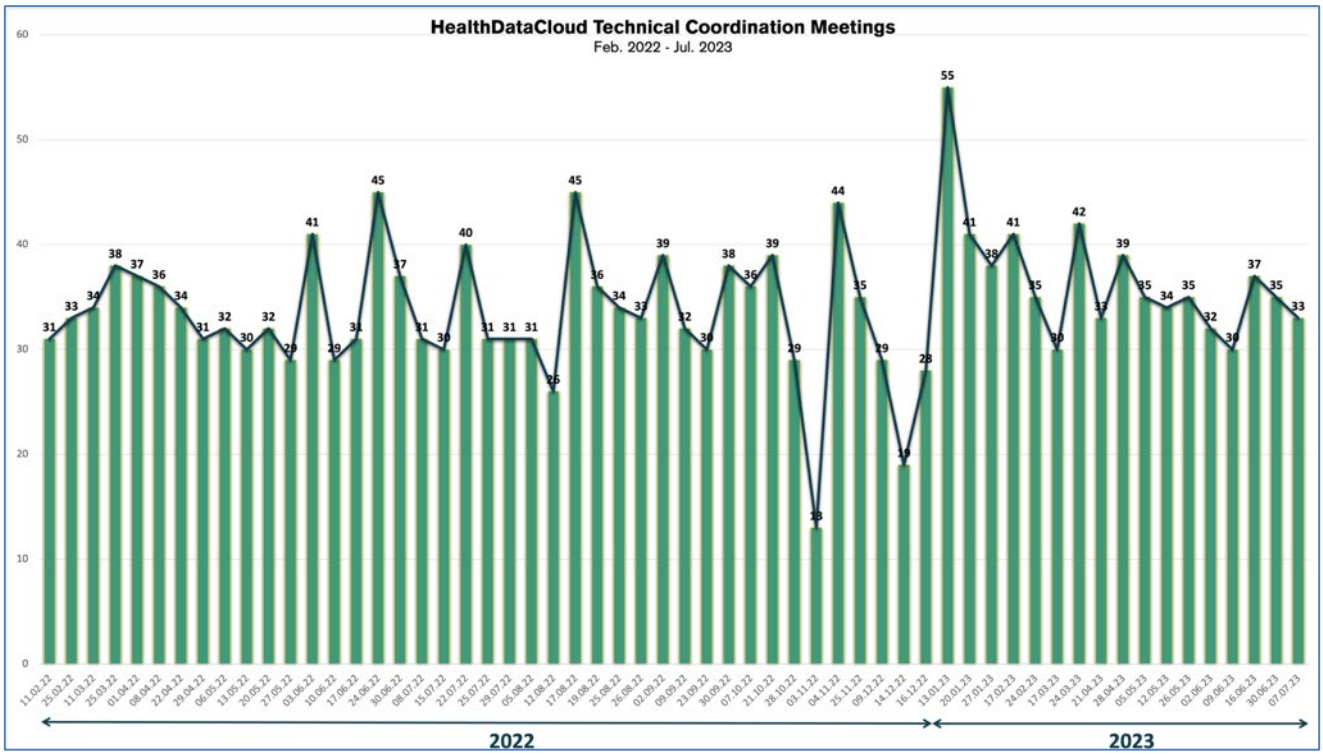

Figure 33: Weekly HDC Technical Coordination meetings led by Charité: Participation

5. Outlook

EBRAINS HealthDataCloud enables scientists to perform research with health data, including with personalized human digital twins, while complying to EU law and protecting the rights of the data subjects.

While we have a fully functional and certified service in place for EBRAINS users, we will continue developing the functional scope of the HDC and we will extend the number of nodes in the federated HDC peer-to-peer network.

The HDC plans to contribute its advanced health data platform technologies to build the European Health Data Space.

It also is taking steps to become an official repository for managing sensitive data in a FAIR and lawful manner for open science journals such as Nature Scientific Data.

The Health Data Cloud is being further developed as an EU-wide resource for the public and for the private sector as part of the EU Reference Center Testing and Experimentation Facility for Health AI and Robotics.

In the future we plan to build on, advance and integrate existing technologies in continued joint collaboration between the MIP/HIP and HDC teams towards an interoperable network of platforms with different functional scopes tailored to their user communities. We plan to also package and offer the entire suite of these technologies in a way compatible with the EOSC Interoperability Framework and the EHDS. Compliant with data protection legislation and other local, national, and European regulations, we will jointly develop a complete accountability toolset, e.g., automated contract services and logging, thereby supporting cloud hosted workflows across computer centres in different countries.