CLINT

CLIMATE INTELLIGENCE

D8.4
CENTRAL DATA REPOSITORY

April, 2024

This project is part of the H2020 Programme supported by the European Union, having received funding from it under Grant Agreement No 101003876

Programme Call:	Building a low-carbon, climate resilient future: climate action in support of the Paris Agreement (H2020-LC-CLA-2018-2019-2020)
Grant agreement ID:	101003876
Project Title:	CLINT
Partners:	POLIMI (Project Coordinator), CMCC, HEREON, CSIC, SMHI, HKV, E3M, TCDF, DKRZ, IHE, ECMWF, UAH, JLU, OGC, UCM
Work-Package:	WP8
Deliverable #:	D8.4
Deliverable Type:	Document
Contractual Date of Delivery:	30 June 2023
Actual Date of Delivery:	30 April 2024 (The delay was agreed upon by the PO. Its justification is that D8.4 is an update to D8.2, but not enough new information was present at the time of the contractual data of delivery to form a new, standalone document).
Title of Document:	Central Data Repository
Responsible Partner:	DKRZ
Authors:	Carsten Ehbrecht, Étienne Plésiat, Nils Hempelmann, Stephan Kindermann
Content of this report:	Report from T8.2 updating deliverable D8.2 and describing the setup and maintenance workflow of the central data repositories
Availability:	This report is public

Document revisions		
Carsten Ehbrecht (DKRZ), Étienne Plésiat (DKRZ), Nils Hempelmann (OGC), Stephan Kindermann (DKRZ)	First draft	30 Mar 2024
Bode Gbobaniyi	Revision and comments	19 April 2024
Michael Maier-Gerber	Typos, grammar and comments on content	23 April 2024
Guido Ascenso (POLIMI), Andrea Castelletti (POLIMI)	Final revision	26 April 2024

Table of content

EXECUTIVE SUMMARY 6
1 INTRODUCTION 7
2 DATA REPOSITORY SETUP 8
2.1 SERVICES SUPPORTING FAIR DATA PRINCIPLES 8
3 TECHNICAL DESIGN OF THE DATA REPOSITORY SYSTEM 10
3.1 CLINT data repository structure 10
3.2 CLINT data sharing space 11
3.3 CLINT long term data archival 11
3.4 CLINT data repository access 12
3.5 CLINT data repository system transfer options 13
3.6 CLINT climate services 14
4 MANAGEMENT OF THE DATA REPOSITORY 15

LIST OF ACRONYMS

ACL:	Access Control List
CMIP:	Coupled Model Intercomparison Project
CORDEX:	Coordinated Regional Downscaling Experiment
CRIS:	Climate Resilience Information System
DOI:	Digital Object Identifier System
DMP:	Data Management Plan
EC:	European Commission
ERA5:	ECMWF Reanalysis
FAIR:	Findable, Accessible, Interoperable, and Re-usable
GB:	Gigabyte
GUI:	Graphical User Interface
HPC:	High Performance Computing
PID:	Persistent Identifier
POSIX:	Portable Operating System Interface
SSH:	Secure Shell
TB:	Terabyte
URL:	Uniform Resource Locator
WDCC:	World Data Centre for Climate
WPx:	Work Package (where x is the WP number)

1. EXECUTIVE SUMMARY

This document is an update of deliverable D8.2 ‘preliminary data repository’. Therefore, it is a revised version of the previous text with updates for the current status of the repository.

This document describes the mature design of the CLINT central data repository, which establishes a central data hosting and sharing platform to support the activities of the CLINT consortium by coordinated provisioning of core input and output data in support of WPs 2-8. The data repository was established by DKRZ and is now operative and managed according to the needs of the CLINT consortium. DKRZ is a data service provider, which enables the integration of existing data services, facilitating the sustainable storage and management of core data collections according to FAIR data management principles. The CLINT data repository is associated with specific data storage quotas, which are granted to the CLINT project partners yearly in accordance with the decisions of the scientific steering committee of DKRZ. This deliverable provides an overview of the core operational services supporting the usage of the data repository. The identification of core data collections to be managed in the centralised CLINT repository is done in close collaboration with continuous maintenance of the CLINT consortiums data requirements described in the CLINT data management plan (DMP D8.1 preliminary, D8.3 1st update, D8.6 2nd update, D8.8 final DMP). The specific data management procedures ensuring data provisioning based on agreed metadata, directory structure and file naming conventions are established stepwise. The first step is a lightweight data management structure outlined at the end of the document. The document also covers the connection of the CLINT data repository to long-term data archives. This points out the data lives even beyond the CLINT project funding phase. Furthermore, it addresses the technical realizations of data processing options enabling scientific analysis next to the data repositories which avoids the transportation of large data volumes. Details of data processing mechanisms will be described in the deliverables ‘Climate Resilience Information Systems Architecture’ (D8.5) and ‘Extreme Events ML in Climate Resilience Information Systems’ (D8.7).

1 INTRODUCTION

As described in the data management plan (DMP, presented in D8.1 and its updates, D8.3, D8.6, D8.8), the CLINT consortium can make use of three existing and connected data repositories. Most of the scientific input data for WP2-8 and selected output data are stored in the CLINT central data repository hosted and managed by DKRZ. The design of the different data repositories, their setup, access and usability are described in this deliverable.

Based on the principles stated in the DMP, the CLINT central data repository has been designed and set up for operational usage and has restricted access to authenticated users. The Deliverable D8.2, entitled 'Preliminary operational data repository', outlined the operational setup of the data repository, which supports the implementation of the data management plan elaborated in T8.1 and described in D8.3, followed by the next update D8.6 one month after this deliverable. A final version of the DMP will be provided by the end of the project with D8.8.

The data repository is used to collect and manage the core input and output data in support of WPs 2-8. In particular, WP8 is significantly supported by the data repository, as the automatic technical climate services rely on accessible data. The main purpose of WP8 is to establish blueprints on how to design and develop Climate Resilience Information systems. Since the term 'Climate Resilience Information Systems' (CRIS) is frequently used in ongoing internal discussion, it will hence be used in this project as well. CRIS are software architectures in HPC and cloud environments to process data close to the archives avoiding big data transfers. This setting eases the deployment of analytical algorithms to create and deploy data products and information on demand. The CLINT data repository is also a source for the technical prototypes of the CRIS (which will be described in Deliverable D8.5, entitled 'Climate Resilience Information Systems architecture') where scientific algorithms are deployed following interoperable standards. Therefore, the data repository has been designed in a way that enables the searching and retrieval of data from automatic services.

Sustainable data storage and FAIR principles are also ensured for the data repository content by the close integration of the certified DKRZ long-term archive and the additional possibility to publish data in the World Data Centre for Climate (WDCC) hosted at DKRZ. More details on the FAIR data management principles were reported in the data management plan, as well as all the rules, regulations, and provisions to deal with data security and ethical requirements that might arise with data acquisition, management, and sharing in CLINT.

Section 2 summarises the basic setup of the data repository, including the associated services that support the use of the data repository and enable FAIR data principles. The initial design of the repository (structure, quotas, access, service usage) is then described in section 3. The basic management of the data repository is characterized in section 4, and is the current final version of the data repository design according to the CLINT consortium agreements regarding the definition of the content, structure, and management conventions of the data repository. It represents a mature status of the central data repository in terms of technical usability, future minor changes in the design are still possible if needed by the consortium.

2 DATA REPOSITORY SETUP

The data repository is hosted as part of the workspace granted to the CLINT project at the DKRZ HPC system and associated data centre. This enables the exploitation of services and resources, already established at DKRZ, supporting the FAIR aspects of data management[1], as recommended by the EC. For more details on FAIR principles, see D8.1, D8.3 and soon D8.6 ('Data Management Plan').

2.1 SERVICES SUPPORTING FAIR DATA PRINCIPLES

The CLINT data repository is associated with the following supporting services:

- data catalogs to search and access existing climate data collections hosted at DKRZ (including CMIP5, CMIP6, CORDEX and ERA5 high-volume data collections);
- a JupyterHub environment[2] to interactively work on data (e.g exploiting the mentioned data catalogs), which is associated with some computing resources allocated on the DKRZ HPC system;
- a set of services to ingest external data into and export data from the data repository, including a data sharing environment based on cloud storage;
- a service to archive data into the DKRZ long-term archival system, offering different possibilities for the storage of associated metadata.

An overview of the structure and interlinkages of the services associated with the CLINT central data repository is illustrated in Figure 1. The core data repository is associated with a conventional POSIX directory hosted as part of the DKRZ HPC Lustre file system. The structure and content of this directory are organised following the specifications outlined in section 3 and are managed by the CLINT consortium as described in section 4. The specific data sharing, transfer and archiving options listed in Figure 1 are also elucidated in section 4. The overall design of the integrated architecture, including the associated data sharing and processing services consists of:

- the core CLINT central data repository implemented as a managed data pool with restricted access to authenticated users (POSIX file system space);
- a cloud-based data-sharing space (with flexible access and different restriction options);
- a long-term archival backend with the additional possibility to publish data sets as part of the WDCC[3].
- An associated HPC capacity where data can be processed and analyzed next to the CLINT data repository. This can be extended to the envisioned CRIS with deployed scientific analytics (D8.5 Climate Resilience Information Systems Architecture and D8.7 Extreme Events ML in Climate Resilience Information Systems).

[1] https://docs.dkrz.de/doc/dataservices/index.html

[2] https://jupyter.org/hub

[3] https://www.wdc-climate.de/ui/

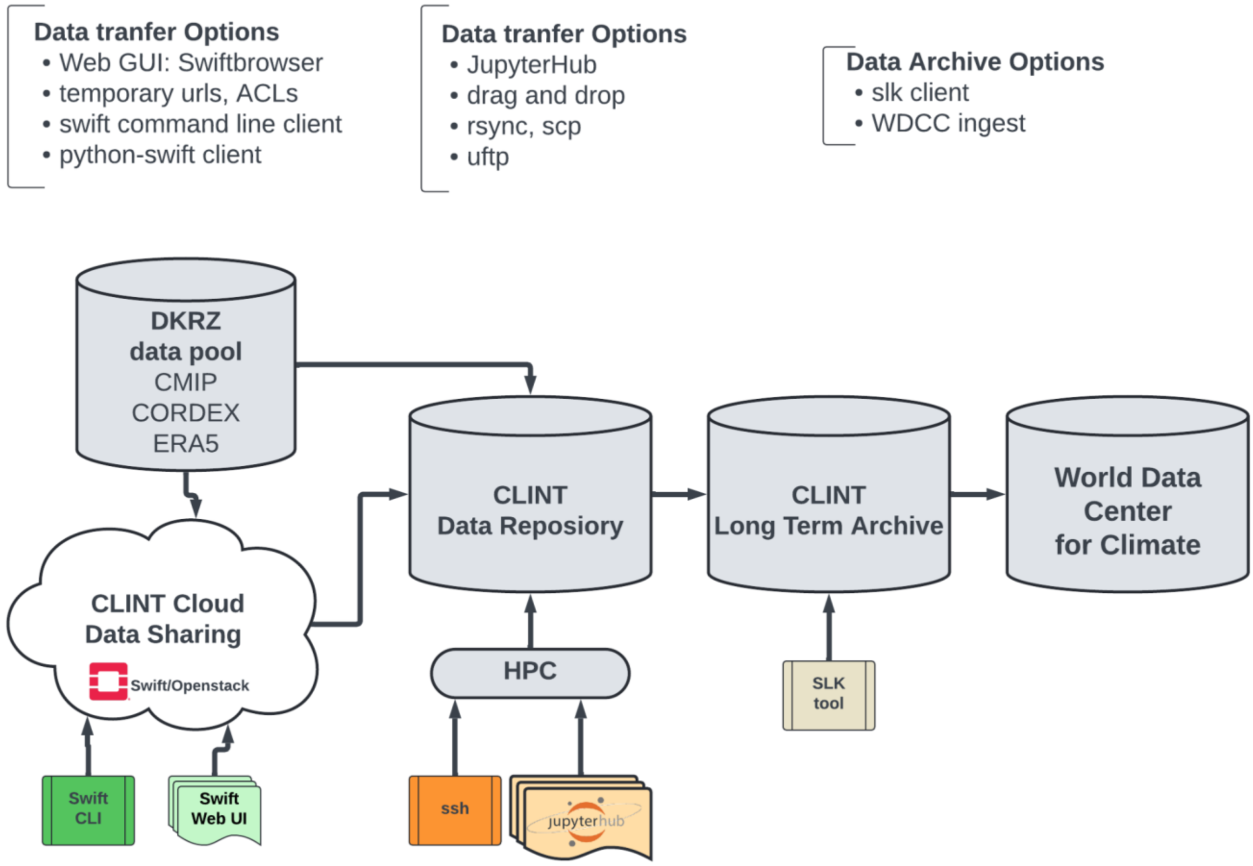


Figure 1 Overview of the data pipelines integrating the CLINT data repository with the associated data-sharing services (cloud-based data sharing and long-term archive).

In summary, the data repository setup is supporting three different stages of the CLINT data management life cycle.

1. The initial project storage and sharing of data collections based on an access in the restricted shared storage space dedicated to the project consortium.
2. Further, the data repository enables external data sharing based on a cloud storage space with freely definable access limitations. This enables data access according to the case of data sharing, e.g. completely open or open for a specific time and for a specific group of people.
3. In the third stage, storage of core data sets which are of interest beyond the lifetime of the project is possible in the certified long-term data archive of the WDCC. Here data storage is certified by the CORETrustSeal[4].

[4] https://www.dkrz.de/en/communication/news-archive/wdcc-cts-certified

Support for the FAIR data principles is important in each part but manifests itself differently depending on the specific context of data usage envisioned, e.g. from 'project internal' to 'project external' to the broader research community. Therefore, a summary is given in the following on how the FAIR data principles are addressed in the design of the CLINT data repository and enabled as part of the individual data life-cycle stages.

Within CLINT project, a stepwise definition and enforcement of conventions on data storage ensures basic data findability, access, and reusability in facilitating the work conducted in the individual WPs as well as collaboration across the CLINT work packages. These conventions include directory structure, file naming and specific metadata requirements. Metadata requirements can involve, e.g. specific NetCDF-CF[5] attributes and a short characterization of the data provenance. These conventions and their enforcement are documented as part of the evolving data management plan.

If external data sharing is needed, it can be done through the definition and enforcement of naming and metadata conventions. Further, the provisioning of data catalogs supporting the discovery and access of cloud hosted data collections is envisioned. Examples of existing catalogs and their usage are described in section 3 where the data repository is documented. They can act as a blueprint for, good practise guidance of data management following FAIR principals, designing data repositories with associated processing services and usage of certified data identifier. If data catalogs are created, they will also be documented in the CLINT data management plan.

Finally, for long-term data storage, the core FAIR data principles are supported by the data management process associated with the certified long-term data archival centre. The data ingestion process and associated quality assurance procedure is ensuring data visibility and accessibility as part of the WDCC data portal. Existing metadata harvesting procedures enable visibility of data in external data portals if needed (e.g., as part of the European Open Science Cloud[6]). The data ingestion process is also documented as part of the data management plan. If necessary, this ingestion process can include the provision of persistent identifiers (PIDs) as well as data citation references (DOIs), which are key components of FAIR data.

[5] https://cfconventions.org/

[6] https://open-science-cloud.ec.europa.eu/

3 TECHNICAL DESIGN OF THE DATA REPOSITORY SYSTEM

The overall data repository system consists of the core CLINT data repository and a set of associated services supporting the exploitation of existing data collections, hosted at DKRZ and managed therein. In this section, the initial structure and access information for the repository system are summarised.

3.1 CLINT data repository structure

The root directory of the CLINT data repository is associated to the path '**/work/bk1318/Data_Repository'** and contains the following technical elements:

- **Catalogs**: a directory containing data catalogs (metadata characterising the stored data collections), supporting the search and access of data collections;
- **Notebooks:** a directory containing example Jupyter Notebooks to demonstrate the use and exploitation of the data repository;
- **Work package storage areas** (called **WP1, …, WP8**);
- A **Service Area** associated with WP8 and intended to handle the data associated with the (pre)operational data services.

The allocation of storage resources and their associated quotas is renewed every year based on the decision of the DKRZ scientific steering committee, which regulates the HPC management and resource allocation. The current initial quota allocations for CLINT are as follows:

- The quota associated with the CLINT data storage space (**/work/bk1318**)**,** including the central data repository, is **60 TB**.
- Additionally, each user is allotted a personal **30 GB** quota for the home directory.

3.2 CLINT data sharing space

The data sharing space provides a flexible mechanism to share data within the consortium as well as externally. Data collections can be assigned direct access URLs via the web, and access rights may be granted on a flexible basis. The following facilities and resources are already implemented:

- a cloud-based data sharing space, accessible via a web interface at: https://swiftbrowser.dkrz.de
- an API and clients that can be used directly in data sharing and analysis scripts from CLINT partners, as documented at: https://docs.dkrz.de/doc/datastorage/swift/index.html

The initial quotas allocated for the cloud data sharing space are **50 GB** per individual user and **10 TB** for the entire CLINT project. The quotas can be extended if needed.

3.3 CLINT long-term data archival

The long-term archival of core data collections is managed in close cooperation with the DKRZ data management team. Interactions are managed via the request-tracker contact at data@dkrz.de.Basic data transfer to tape is supported by the SLK tool[7], yet the overall process is coordinated by the DKRZ data management team itself. There is **no specific quota** limitation for the long-term archival service. The decisions on which data collections should and can be made available in the long-term are based on agreements between the CLINT consortium and the DKRZ archival team and essentially relate to the provision of high-quality data and meaningful metadata characterize the data collections. Decisions on the external visibility of CLINT data collections that are archived as part of the WDCC in external data portals (e.g. in the European Open Science Cloud) are based on data collection-specific agreements.

3.4 CLINT data repository access

Access to the CLINT data repository is restricted to CLINT project members and requires an explicit registration process which is described in section 4. After registration, there are essentially two main options for direct access to the data repository. Options for data management after the project funding phase will be given in the final Data Management Plan (D8.8).

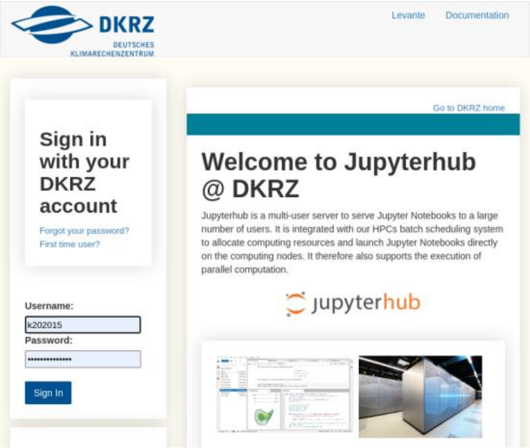
Option A: Via the SSH command line access mechanism, where two steps are involved.

- **Step 1:** In the command line, use SSH to login nodes (ssh kxxx@levante.dkrz.de).
- **Step 2:** Go to the data Repository (/work/bk1318/Data_Repository).

Option B: Via JupyterHub, which includes Option A, since interactive shells can also be spawned. It includes the following steps.:

- **Step 1:** Log into https://jupyterhub.dkrz.de with the credentials you received during the registration process outlined in section 4. This step is illustrated in Figure 2.

[7] https://docs.dkrz.de/doc/datastorage/hsm/getting_started.html



Sign in with your DKRZ account

Forgot your password?
First time user?

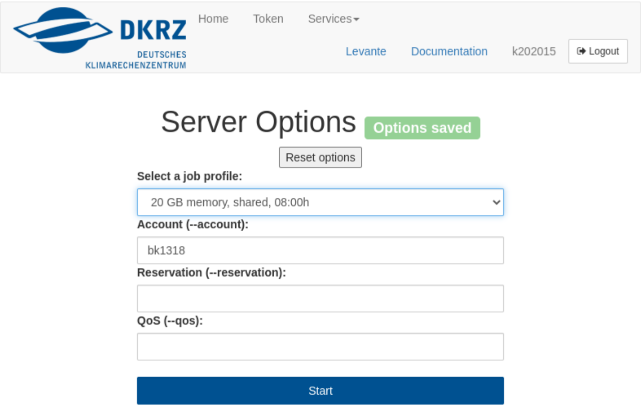
Username: k202015
Password: ••••••••••••••

Welcome to Jupyterhub @ DKRZ

Jupyterhub is a multi-user server to serve Jupyter Notebooks to a large number of users. It is integrated with our HPCs batch scheduling system to allocate computing resources and launch Jupyter Notebooks directly on the computing nodes. It therefore also supports the execution of parallel computation.

Figure 2 Login page of the CLINT JupyterHub environment associated to the CLINT data repository.

- **Step 2:** After the login, users can start an interactive Jupyter notebook. Different notebook options can be selected based on the resource requirements of the interactive session. This selection step is illustrated in Figure 3.



Server Options Options saved

Select a job profile: 20 GB memory, shared, 08:00h

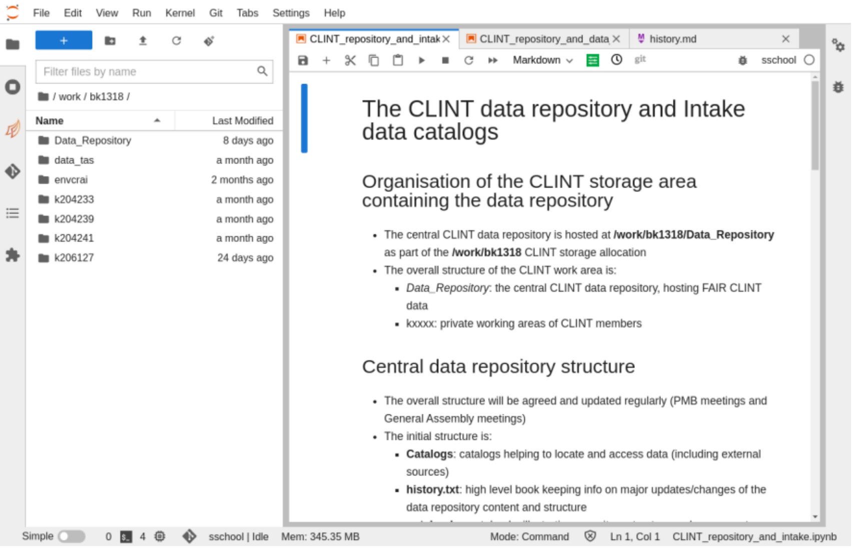
Account (--account): bk1318

Reservation (--reservation):

QoS (--qos):

Figure 3 Selection of the resource profile associated with an interactive session.

- **Step 3:** Once an interactive notebook is spawned, it is ready for input. Different types of notebooks can be selected based on the preferred programming environment (e.g. Python, Julia or R). Example notebooks are accessible as part of the CLINT data repository, as illustrated in Figure 4.



/ work / bk1318 /

Name	Last Modified
Data_Repository	8 days ago
data_tas	a month ago
envcrai	2 months ago
k204233	a month ago
k204239	a month ago
k204241	a month ago
k206127	24 days ago

The CLINT data repository and Intake data catalogs

Organisation of the CLINT storage area containing the data repository

- The central CLINT data repository is hosted at **/work/bk1318/Data_Repository** as part of the **/work/bk1318** CLINT storage allocation
- The overall structure of the CLINT work area is:
    - *Data_Repository*: the central CLINT data repository, hosting FAIR CLINT data
    - kxxxx: private working areas of CLINT members

Central data repository structure

- The overall structure will be agreed and updated regularly (PMB meetings and General Assembly meetings)
- The initial structure is:
    - **Catalogs**: catalogs helping to locate and access data (including external sources)
    - **history.txt**: high level book keeping info on major updates/changes of the data repository content and structure

Figure 4 Interactive example session based on jupyter notebooks, illustrating the data search and access based on pre-existing data catalogs.

3.5 CLINT data repository system transfer options

The data transfer options for the CLINT data repository include the standard mechanisms ftp, rsync and http and are documented at https://docs.dkrz.de/doc/levante/data-transfer/index.html.

An overview of the different types of file systems involved in data analysis activities is provided at https://docs.dkrz.de/doc/levante/file-systems.html. The different data storage components (file system, cloud, archive) are described in detail at https://docs.dkrz.de/doc/datastorage/index.html

3.6 CLINT climate services

To support the (pre-)operational deployment of climate services using the CLINT data repository, an additional associated CRIS is hosted at DKRZ. It uses software libraries to run the climate data analysis calculations required by the processing services deployed. Some selected scientific methods are going to be deployed in the CRIS and will be described in Deliverable D8.7 (‘Extreme Events ML in Climate Resilience Information Systems’).

As these computations usually need local data (stored in the CLINT data repository), the processing services can be configured with the local data access paths (see Figure 5). Therefore, the service calls the climate data analysis tools with the correct paths to local files in the CLINT data pool.

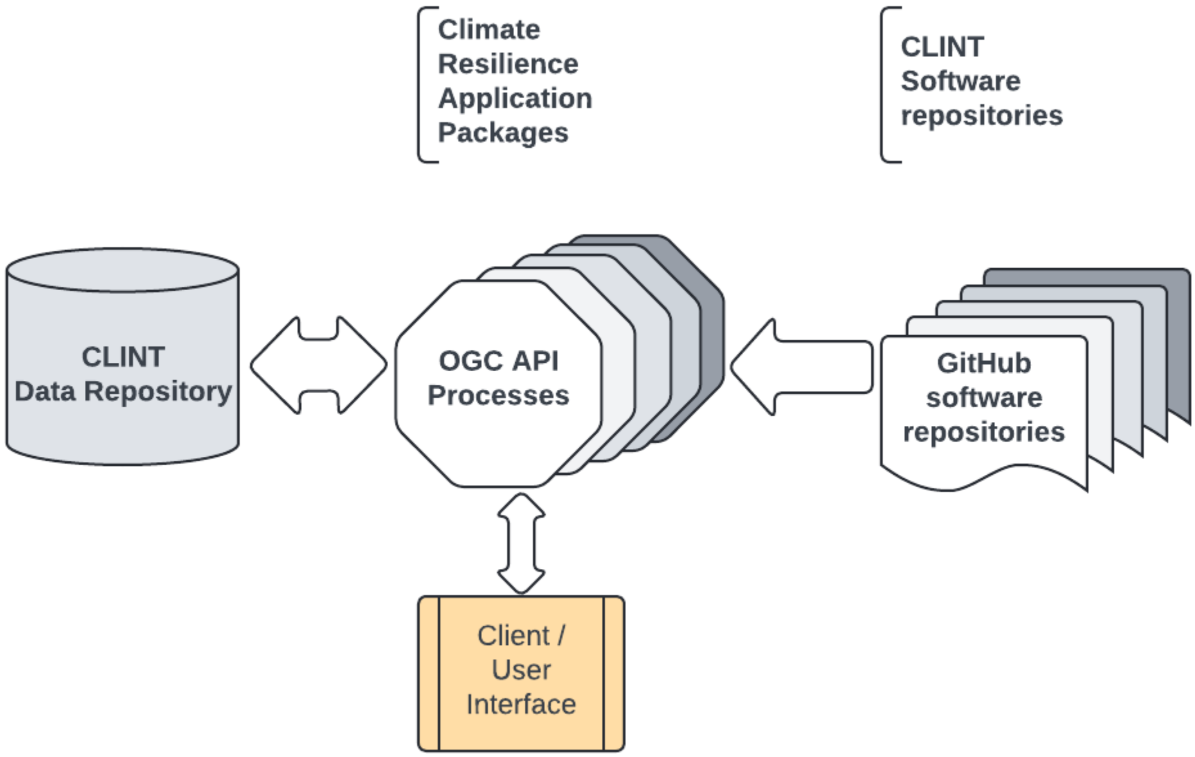

Climate Resilience Application Packages
CLINT Software repositories
CLINT Data Repository
OGC API Processes
GitHub software repositories
Client / User Interface

Figure 5 Design of the Climate Resilience Information System. The processing service is following interoperable standards and is therefore deployable in the Copernicus Climate Services (C3S) Climate Data Store (CDS).

4 MANAGEMENT OF THE DATA REPOSITORY

This section reports the basic management of the CLINT data repository and constitutes the result of the discussions and agreements on how the content, structure, and management conventions of the data repository is being used in CLINT. Future adoptions are still possible if needed and will be reported in the upcoming versions of the DMP.

User admission and access

- Users need to register an account at https://luv.dkrz.de/register/
  Then users need to log into https://luv.dkrz.de and select "Join existing project":
  - Select "CLINT - climate intelligence" (project bk1318) by scrolling down the project list or by typing into the search bar.
  - In the "message" part, shortly characterise your CLINT project context, e.g. which work package you belong to.
- After registration, users will receive a confirmation email with admission information for access to the CLINT data repository system and also to the cloud-based data sharing space.
- Access to the interactive Jupyter Hub environment is included.

- The CLINT data space is accessible on the file system path **/work/bk1318**. The data repository is located at **/work/bk1318/Data_Repository** and contains shared and curated data collections**.** A large collection of important data sets (e.g., CMIP5, CMIP6, CORDEX, ERA5) is freely accessible for CLINT project members via **/pool/data** as well as the associated data catalogs.

Management of user access and CLINT data space:

- Every CLINT project member can access the CLINT data space (/work/bk1318) including the data repository with write permissions.
- A dedicated group of people having the "data manager" role (e.g. WP and/or task leaders) have the responsibility to review the data space organisation. This encompasses a regular assessment of storage space usage, data organisation based on directory structures and file naming conventions. Problems are discussed as part of the regular management board meetings and decisions are enforced in collaboration with the DKRZ data management team (data@dkrz.de).
- The manager group also agrees on a well-defined directory structure enabling a transparent navigation of data collections in the CLINT data repository.

Management of the data sharing space and the long-term archive

- The usage of the data sharing service and associated cloud storage space is not subject to general management regulations. The individual users, task and WP leaders decide based on their data sharing requirements.
- The long-term archive is managed by the DKRZ long-term archival team. There is an explicit data handover step necessary to ingest the data. The data handover includes the agreement on the data and metadata quality requirements. The necessary communication is managed in the data support tracker associated with the email address data@dkrz.de.

Management of associated HPC processing resources

- The CLINT project can apply for the provision of HPC processing resources to support data analysis activities associated with the CLINT data repository. These allocations are provided on a yearly basis and require a written project proposal, which is reviewed by the scientific steering committee of the DKRZ. An initial allocation was granted valid until June 2024, with option to extend for the period July 2024 to June 2025. This allocation includes an additional storage allocation of 15 TB for short term data handling, as part of the /scratch file system (see documentation provided at https://docs.dkrz.de/doc/levante/file-systems.html).

CLINT

CLIMATE INTELLIGENCE

This project is part of the H2020 Programme supported by the European Union, having received funding from it under Grant Agreement No 101003876