HEALTHYCLOUD
Health Research & Innovation Cloud

D5.3 Guidelines to establish sustainable computational infrastructures for the future HRIC ecosystem
Version 0.6

Document Information

Contract Number	965345
Project Website	http://www.healthycloud.eu/
Contractual Deadline	M21, Nov 2022
Dissemination Level	PU
Nature	R
Author(s)	Sina-Victoria Barysch, Björn Grüning, Jan Korbel, Jens Krüger, Burkhard Linke, Harald Wagener (de.NBI-Cloud)
Contributor(s)	Alba Jené (BSC) Salvador Capella-Gutierrez (BSC) Juan Gonzaléz Garcia (IACS) Adrian Thorogood (UNILU) Regina Becker (UNILU) Gergely Sipos (EGI) Mark Dietrich (EGI) Anamika Chatterjee (ELIXIR) Anna Niemeyer (TMF) Irene Schlünder (TMF),
Reviewer(s)	Gergely Sipos (EGI) Pascal Derycke (Sciensano)
Keywords	Cloud, sensitive data, secure processing

Notice: *The HealthyCloud project has received funding from the European Union's Horizon 2020 research and innovation programme under the grant agreement N°965345*

Change Log

Version	Author	Date	Description of Change
V0.1	Sina-Victoria Barysch, Björn Grüning, Jan Korbel, Jens Krüger, Burkhard Linke, Harald Wagener	2022/03/24	Initial Draft
V0.2	All above plus Adrian Thorogood	2022/03/30	Initial feedback from WP2 incorporated
V0.3	All above plus Alba Jené, Salvador Capella-Gutierrez, Juan Gonzaléz Garcia, Adrian Thorogood, Regina Becker, Gergely Sipos, Mark Dietrich, Anamika Chatterjee, Anna Niemeyer	2022/05/09	Detailed feedback from colleagues (following a 2-hour workshop) partially incorporated
V0.4	All above plus Irene Schlünder	2022/06/29	Additional feedback incorporated, based on individual discussions with various stakeholders
V0.5	All above	2022/11/15	All feedback was incorporated, document revised, extended and finalized for review
V0.6	All above plus Gergely Sipos and Pascal Derycke		
			(Final Change Log entries reserved for releases to the EC)

Table of contents

Executive Summary

The objective of this deliverable D5.3 is to outline a conceptual design of a sustainable computational backbone for the future HealthyCloud ecosystem based on derived infrastructure models. For that, we have performed a gap and opportunity analysis to identify what the different computational infrastructure models (Deliverable 5.1) in Europe currently offer and what is needed to enable a European decentralised computational infrastructure for health research.

Importantly, this deliverable contains recommendations on modular and flexible architectures to facilitate the deployment of new infrastructures and/or the conversion of existing ones.

Introduction

Ideally, Computational Infrastructures (CIs) should strive to hide all technical aspects of health and life science research by providing easy services that allow submission and execution of workflows and pipelines. Computational Infrastructure in this document comprises compute, storage, and network components, as well as higher level services to facilitate research. Optimization of the resource usage SHOULD happen outside the view of the end users by the operators of Computational Infrastructures.

Effective use of current Computational Infrastructures requires training of users on both technical and operational measures to handle data and resources responsibly. Computational Infrastructures SHOULD minimise direct access to low level infrastructure to facilitate use.

Data Centric Health Research Computational Infrastructure as defined in the HealthyCloud glossary are a specific variant of the broader concept of Computational Infrastructure covered in this document.

The de.NBI Cloud is a federated cluster of research cloud infrastructure for academic research in the life sciences. It consists of multiple sites provided by 8 institutions across Germany and has seen funding of more than 50 million Euros between 2018 and 2021 by the Federal Ministry of Education and Research, Germany. Members of the de.NBI Cloud governance body created an initial version of this document, which then was revised and expanded based on feedback from other HealthyCloud representatives.

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119[1].

[1] https://datatracker.ietf.org/doc/html/rfc2119

1. Management Aspects of Computational Infrastructures

1.1. Authentication and Authorisation

Authentication and Authorisation is one of the critical cornerstones for successful European Research Computational Infrastructures. In the experience of de.NBI Cloud, the Life Science Login[2], provided by EGI[3] (CESNET) and GEANT[4] (GRNET) implementation of CESNET satisfies the requirements to enable consortia with membership of many different research institutions from across Europe to successfully and jointly use resources provided by a network of independent Computational Infrastructures. One particular strength of Life Science Login is that once a project has been set up, the responsible Principle Investigator (PI) can independently manage who will have access to the resources allocated to the project, which greatly streamlines access to the system. Another big plus is that all participants can use the credentials of their home institution rather than having to deal with yet another set of credentials.

Within de.NBI Cloud we also realised integration of Life Science Login in a way that minimises personal data of researchers being exposed to the infrastructures: The IDs used to identify individual users are masked by Life Science Login, and since the PIs manage access via the Perun system within the Life Science Login, there is no need for the infrastructure provider to know names, affiliations, or email addresses of individuals since this is handled by the PI. Since user IDs within Life Science AAI are cross-institutional, they remain stable even if an individual changes affiliation. This is an additional bonus as it reduces management overhead considerably. This way, infrastructure providers do not hold PII of the users directly: It is all encapsulated within Perun, managed by the PI as Project Owner, and responsibility and management for access is retained with a single entity.

Infrastructures SHOULD implement federated single-sign-on solutions like Life Science Login to allow researchers to use established identities and to limit exposure to personal data of researchers. These federated sign-in solutions SHOULD support multifactor authentication if used to govern access to sensitive data. Note that special considerations MUST be taken into account when working with sensitive data and/or close to clinical patient management systems[5]. Computational Infrastructures SHOULD provide hooks to allow project owners

[2] https://lifescience-ri.eu/ls-login/
[3] EGI is the federation of computing and storage resource providers united by a mission of delivering advanced computing and data analytics services for research and innovation. (www.egi.eu)
[4] The GÉANT Association is the collaboration of European National Research and Education Networks (NRENs). Together, we deliver an information ecosystem of infrastructures and services to advance research, education, and innovation on a global scale. (www.geant.org)
[5] https://edps.europa.eu/data-protection/our-work/publications/opinions-prior-check/clinical-patient-management-system-ec_en

(usually PIs) to govern access to project specific resources using additional attributes or roles, i.e. proof of training or contractual agreements.

1.2. User Management for Sensitive Data

Computational Infrastructure operators cannot provide assurances about individuals' suitability to use resources beyond a given project scope and primary investigator. It is the PI's duty to ensure that only verified identities of persons with the applicable rights to access are being given access to a projects' resources (both data and analytical tools).

For the work with sensitive data, such as health-related data, a functioning AAI as described above is essential. Since it is impossible for the Infrastructure Provider to verify all bona fide researchers, they MUST rely on PI verification and existing identities for authentication. Therefore, it is important that this can be done in a comfortable manner by the project PI or even automatized using predefined workflows. The Life Science Login is an example for an appropriate solution for PIs to add or remove collaborators from different institutions to their project. As mentioned above, technical implementation of an AAI system does not absolve PIs of their responsibility to validate and verify that only those with the right to access data are granted access via the system used.

Another important service for the work with sensitive data are so called passports and associated visas (https://www.ga4gh.org/ga4gh-passports/). The standard for these has been developed by the GA4GH to manage access to defined data sets of varying size. This standard provides each researcher with a virtual passport that proves their status as a bona fide researcher and contains information like the institution they are working for. Access to data is then granted on a so-called visa based system. In this, access to the respective data set(s) is granted by issuing a visa to the researcher. This visa can contain different constraints, such as affiliation, expiration and usage limitations.

To achieve an efficient and transparent user management for both the data consumers and data providers an infrastructure SHOULD support a solid authorization system like GA4GH passport/visa system or a functionally similar system. Any system used SHOULD provide a Data Protection Impact Assessment to allow for proper adjudication of its fitness for use.

1.3. Data Management

Beyond access control, data discovery is an important aspect of data management. With the growing amount of sensitive data that is available only access control conditions, it became clear that manual access management, such as the issuing of a GA4GH visa, becomes a bottleneck. To mitigate this problem the GA4GH

developed a data use ontology (DUO)[6][7] that allows to map the content of an informed consent to standard terms and hence allows for an automated issuing or at least pre-issuing of visas. To remove ambiguity and reduce the workload when processing data access requests, data sets containing sensitive data and offered in public clouds SHOULD be tagged with the appropriate DUO terms. Unfortunately, it is very difficult and often impossible to tag data collected with legacy informed consents. Therefore, all prospective consents SHOULD be drafted with the principles of DUO in mind. The GA4GH guidelines on machine-readable consent provide guidelines on how to draft consents that map unambiguously to the DUO. Use of DUO MUST NOT be seen as a data access control, which is covered in the "Authorisation and Authentication" and "User Management" section above. Please note that DUO is not necessarily sufficient to solve all data ontology needs and it is presented as an example option for ontology and further investigation to better solutions or a DUO extension (which is investigated within EOSC-Life) may be required.

1.4. Training and Documentation

To ensure the best resource usage and avoid security issues, Computational Infrastructures SHOULD be offered as services that preclude abuse. For those Computational Infrastructures that provide lower level access to infrastructure, a training regimen MUST be offered so new researchers can familiarise themselves with the systems and their proper use. All training MUST be accompanied by up-to-date documentation on proper use. Note that this is different to an Acceptable Use Policy, which is often a compliance requirement for any access to Computational Infrastructures. For training, providers of sensitive data SHOULD provide synthetic data sets so researchers can familiarise themselves with the use of systems without risk of data misuse.

Training and documentation is not limited to researchers, but MUST be provided to technical personnel of the Computational Infrastructures (see operational sustainability below). This kind of training is not about the effective use of Computational Infrastructures for research, but about expertise in designing, building, maintaining, and expanding, and properly decomision of Computational Infrastructures to satisfy researcher needs.

1.5. Approaches to Computational Infrastructure Design

One model that MAY be used by HRIC architects to develop effective controls that consider ethical and legal considerations is the 5 Safes model[8]. It sets the stage for

[6] https://github.com/EBISPOT/DUO
[7] https://www.ga4gh.org/genomic-data-toolkit/#:~:text=A%20GA4GH%2Dapproved%20Standard%20htsget,slow%2C%20resource%2Dintense%20process
[8] https://ukdataservice.ac.uk/help/secure-lab/what-is-the-five-safes-framework/

ethical use of data from various points of view. This facilitates a holistic approach to use of data:

- Safe data: data is treated to protect any confidentiality concerns.
- Safe projects: research projects are approved by data owners for the public good.
- Safe people: researchers are trained and authorised to use data safely.
- Safe settings: a Secure Lab environment prevents unauthorised use.
- Safe outputs: screened and approved outputs that are non-disclosive.

An extensive discussion of the 5 Safes in the context of HealthyCloud has concluded in D5.1.

Alternately, the Environmental Research Infrastructures Reference Model (ENVRI RM)[9] is geared towards the development of research infrastructures and MAY be used as a guideline to support research infrastructures (RI) define and establish their operational infrastructures.

Another aspect that SHOULD be considered is a Critical Path Analysis. Dependencies SHOULD be evaluated and their sustainability evaluated. Infrastructure providers with a high need for sustainability SHOULD consider supporting open source projects that are critical to their infrastructure.

All Infrastructures SHOULD follow a continuous improvement process, which identifies aspects that need improvement or adjustment to new technical or regulatory developments. This approach is often required by certifications as well.

2. Basic Building Blocks Of Computational Infrastructures

2.1. Network

Networks MUST be logically separated between various users of an infrastructure when it provides Virtual Machine (VM) based or container orchestration services. In the case of workflow pipeline services, pipeline services MUST limit access to the network to the functional minimum.

There are various technical approaches to achieve network separation. For example, in de.NBI Cloud's underlying OpenStack virtualization implementation, project specific networks are isolated using V(X)LAN technologies and SHOULD in general be separate and not affect other projects.

Another benefit of V(X)LAN separation are isolated network address spaces. In an infrastructure-as-a-service usecase, the users are responsible for defining the virtual infrastructure like networks, subnets and routers in their project. This is

[9] https://link.springer.com/chapter/10.1007/978-3-030-52829-4_4#Sec7

often performed automatically by configuration tools, like terraform[10], that often use the same network address ranges, resulting in conflicts in case of shared networks.

Shared or public networks SHOULD be avoided if possible, and its management MUST be restricted to administrative users. Shared network filesystems that require shared network access between independent projects SHOULD be avoided. Global shared file systems used by multiple projects MUST be read-only to protect data against unwanted changes or deletion. We are aware that legacy applications still depend on network file systems. Cloud native applications MUST NOT use networked filesystems.

Cloud implementations like OpenStack[11] usually provide mechanisms to expose selected virtual machines to the internet, e.g., to access to a web server or provide login via secure protocols. This exposure is also combined with concepts to restrict access in a firewall-like fashion (e.g., the security groups in OpenStack). The default configuration should deny all access. Users have to explicitly allow access to their instances. The current best practice is that by default projects have no externally available services. Exceptions MUST be approved, documented, and time limited.

2.2. Storage

Storage MUST provide highest security like data encryption at rest (i.e., while stored on disk and not in use), project separation, and fine-grained access control with a default to limited access.

A Computational Infrastructure MUST always provide at least one method for persistent storage that is not bound to the lifecycle of an application or virtual machine. The implementation MUST ensure that data availability and data accessibility is given in case of defined failure scenarios like a single failing hard disks or other common faults. In case of OpenStack this can be accomplished by the Cinder block storage service for attached block storage or a S3 compatible storage provider. Users SHOULD be trained to recognize the difference between local ephemeral storage (bound to the lifecycle of a virtual machine) and persistent storage, and how to use these different kinds of storage set-ups in their workloads.

Many applications can benefit from fast scratch storage, e.g., for intermediate or temporary data. It is thus recommended to provide this kind of storage to computing instances. The OpenStack compute implementation can configure extra ephemeral storage on local disks to deal with this use case. Users should be aware that this storage is not persistent and data has to be transferred to other storage locations before an instance is terminated. Indeed, any ephemeral user's data

[10] https://www.hashicorp.com/products/terraform
[11] https://www.openstack.org/

SHOULD be removed from the instance if there is the risk it can be accessed by others.

In contrast to block storage (either ephemeral or attached), object storage access is not part of the virtual machine definition; data is accessed from within the virtual machine via standard network protocols like S3 compatible storage. Modern cloud native applications SHOULD use S3 compatible APIs for storing and retrieving data. This is also the standard assumption for stateless services running in containers, see Queueing and Orchestration below. S3 compatible object storage is the *de facto* standard implemented by various storage solutions.

In a workflow oriented project the mentioned storage types can be combined to optimise processing by data staging: input data is copied from object storage to local (fast) ephemeral storage, processed, and results are copied to object storage. Users and workflow developers SHOULD be trained to this concept; most tools in the field of bioinformatics do not work well with object storage directly and require file based access. Alternately, storage management SHOULD be transparent to end users, for example by complete encapsulation of data management in a workflow engine.

User data MUST be removed from a compute node and operational storage after a project ends to avoid data leakage or misuse. Data SHOULD be archived in appropriate archives for FAIR usage.

2.3. Compute

Compute can be realised in multiple ways and abstraction layers. The simplest solution would be granting access directly to the machine e.g., via the `ssh` protocol. A big disadvantage of this approach is that one either has to grant access to multiple users to the same server or one has to reserve the entire server (and its resources) to a single user. While the single user approach is mainly uneconomical, there are multiple issues with servers shared by users such as security, stability and operating system/software constraints. For these and other reasons wherever possible one SHOULD NOT provide direct access to the bare metal machines.

A very common and vastly applied alternative to direct access is to provide virtual machines for users to work on. Again, in the simplest approach one would prepare a virtual machine for a single user and grant access e.g., via `ssh`. This is a valid approach and MAY be offered for simple tasks.

Most cloud providers have more elaborate systems, such as OpenStack, that allow to provision projects with distinct resources (e.g., #CPU, #GPU, RAM, IPs, …) for the users. The user can then, in a selfservice, provision a single or multiple virtual machines, often based on available images (with e.g., different operating systems or other software pre-installed). This approach is very common for cloud computing but it requires deeper knowledge and experience of the project members.

2.4. Queueing and Orchestration

Every analysis on sensitive data is challenging for a variety of reasons and the systems involved are most likely complex as outlined above. To reduce the security risk but also safeguard the data analysis, it is recommended to reduce human involvement and manual steps during the data analysis as much as possible. In addition a system that orchestrates analysis jobs SHOULD be traceable and reproducible. At any point in time, it SHOULD be clear what has been executed, in which environment and how can a particular step or the entire analysis be reproduced.

To solve this challenge, the first problem is to encapsulate the runtime environment from the host system as much as possible. Only doing this, it can be guaranteed that the same job on a different host yields the same results. We recommend the usage of software container technologies to address this problem.

The second problem is the executing of all analysis steps in a reproducible, strictly defined but independent manner. This is important to be traceable and reproducible. It MUST be ensured that the same inputs and outputs are passed from one step to the other in a chain of jobs (usually referred to as workflow) and that all parameters, including environment variables, stay the same. Doing this manually is very error prone, hence we recommend the usage of a workflow management system (WMS) that orchestrates all jobs and the entire analysis. State of the art WMS, as Nextflow or Galaxy, can orchestrate jobs in different container technologies and with that also address the first problem. HealthyCloud deliverable D5.2 extensively elaborates on this matter[12].

We assume that WMS are traceable and keep track of all provenance information. This provenance information MUST be made available so workflows can be reproduced on alternative workflow management systems. We do note that tracing and making available provenance data is not limited to WMS, even though manual workflow orchestration is discouraged.

WMS also addresses a different problem mentioned in the storage section and that is the dependence on different storage classes. Most tools can not natively interact with object stores, they rely on a POSIX like file system to consume and produce data. However, data can be staged from an object store to a temporary file system before the job is executed and staged back to the object store when the results are obtained. This is usually built into a WMS and should not be done manually by the user.

The underlying queueing or orchestration system (SLURM[13], HT-Condor[14], K8s[15]) used by the WMS is not important as most WMS can handle the most important

[12] https://healthycloud.eu/wp-content/uploads/2022/11/D5.2.pdf
[13] https://slurm.schedmd.com/overview.html
[14] https://htcondor.org/
[15] https://kubernetes.io/

ones transparently. Several frameworks are available for bootstrapping compute clusters with queueing systems in a cloud project, e.g., BiBiGrid or ElastiCluster.

Another, emerging approach to reproducible computational analysis is via JupyterHub[16] and Binder[17]. JupyterHub is a brower-based environment where 'data scientists' can import and analyse data through scripts written in various programming languages (Python and R are the most common in data science). The computational notebooks capture the steps of the data manipulation, and also include explanation/documentation/graphics as part of the code. Computational notebooks can be exported from JupyterHub, can be stored in open access or restricted repositories, and can be replayed by fellow researchers either in another JupyterHub environment, or in a Binder system. Binder is specifically designed to 'replay' notebooks. Both JupyterHub and Binder are using containerisation to achieve the portability and independence of applications and the underlying computer systems. EGI, one of the HealthyCloud consortium members provide JupyterHub and Binder services for research communities. These services can be either hosted on EGI servers (i.e. servers of the National e-Infrastructures that are members of the EGI federation), or on servers of the user community.

3. Sustainability

The sustainability of a Computational Infrastructure is key to its use by researchers. Researchers need a Computational Infrastructure that is reliable in the long term when they start projects, as the prospect of changing Computational Infrastructure deters users. In other words: Sustainability of a Computational Infrastructure is a major decision factor for researchers when choosing a suitable Computational Infrastructure for their projects.

Moreover, if data should be preserved and kept FAIR over a long, sustainable infrastructure to (re-)analyse research data is needed as much as sustainable data hub and collections.

Data gatekeeping, as we see it with commercial clouds at the moment, where the data is in theory accessible but behind a paywall, should be avoided as much as possible to really enable FAIR sustainable access and contribute towards the adoption of the Open Science principles, and especially the ones realted to Open Data.

3.1. Operational Sustainability

To enable sustainability we need to diversify Computational Infrastructure. This means Computational Infrastructures MUST either be run by a federation (e.g., EGI as a European example or de.NBI-Cloud as one national example) or MUST to be backed by diverse funding streams. If one funding source ceases, others have to

[16] https://jupyter.org/hub
[17] https://binderhub.readthedocs.io/en/latest/index.html

maintain the funding for a certain period of time. One example would be institutional funding plus project based funding, or funding via different streams in one institution. Sustained funding for HRICs by the EC would be another option. Moreover, operational sustainability MUST include building expertise within institutions – both on the side of technical personnel as well as for training researchers.

3.2. Strategic Sustainability

Strategic sustainability means that a Computational Infrastructure is aligned with larger, international efforts to build and use Computational Infrastructures for the future, and not just providing what is state of the art at a given time. One example is involvement with projects under European framework programmes, e.g., Horizon2020 and HorizonEurope. The European Open Science Cloud (EOSC) through cluster projects, e.g., EOSC-Life, and European Research Infrastructures, e.g., ELIXIR, represent excellent opportunities to maintain and contribute towards the technical and strategic alignment. In the context of the European Health Resarch and Innovation Cloud (HRIC), it also means ensuring that the specific requirements and challenges for health-related data processing is reflected in the strategies of those larger, overarching efforts. For example: there is no a dedicated effort in EOSC to this particular domain, e.g., EOSC-Health, and the endeavours under the EOSC-Life umbrella do not always fit the use case of working with sensitive personal health-related data.

Another aspect of strategic sustainability is accessibility of infrastructure for researchers: Under commercial models, any implemenation of the European HRIC will immediately be in competition with commercial cloud offerings, both in terms of putting additional financial burdens on research projects and the sustainability of those projects' outcomes; as well as in maturity and abundance of features.

A mixed commercial/academic-supported funding model such as CSC's MAY be an approach to retain both Operational and Strategic sustainability.

3.3. Environmental Sustainability

Computational Infrastructures do have a significant environmental impact, which is often unaccounted for. Computational Infrastructures MUST raise awareness in the community that even if compute and storage resources are free-of-charge to researchers, they do have an impact on the environment.

To enable environmental sustainability, Computational Infrastructures MUST have an *environmental cost* model including nonfinancial aspects such as CO_2 emission or surplus energy consumption for cooling and operations. This cost model MUST be applied to research activities including but not limited to cost of data storage,

workflow execution, dissemination, etc. This SHOULD allow an impact analysis by users to guide them in the choice of environmentally sustainable RIs.

Computational Infrastructures MUST apply this model to job scheduling and data archiving mechanisms to avoid redundant and inefficient compute - and data archiving. Abstracting this burden away from the end-users will be instrumental in effectively achieving environmental sustainability goals.

The challenge is to raise the awareness of the environmental impact of data analysis but at the same time keep exploratory research alive. Computational Infrastructures MUST take environmental accountability and transparency into account.

4. Security and Compliance

4.1. Ethical and legal considerations

Health research relies on large-scale molecular, phenotypic, imaging, clinical data, and a host of other data types for discovery and replication in large biomedical datasets, which includes searchable metadata, to allow for translation of fundamental research results into new clinical protocols. These methods inevitably require the processing of personal data across borders, which creates its own particular challenges within the ecosystem of the EU General Data Protection Regulation (GDPR). The GDPR harmonises rules for processing personal data across Member States and lays down rules that privilege data processing for health research purposes. The impact of the GDPR on harmonising rules for biomedical research have however been limited, due to various opening clauses that allow Member States to adopt different rules for health-related, including genetic, data, and different derogations for health research. Moreover, the current interpretation of GDPR rules for health research differs within Europe, creating persistent challenges for collaborative sharing of health-related research data. This has contributed to create barriers to scientific progress in biomedicine-related research across Europe at a time where our global community faces important world-wide health challenges. Development of the Health Research and Innovation Cloud in the European Research Area offers the opportunity to restore the value of international collaboration in health and genomics research, and bring Europe back to a position where it could become one of the leading global players in this area again in the future. Computational Infrastructures SHOULD be built in such a way that researchers are enabled to comply with existing regulatory and ethical requirements more easily. In the context of EOSC-Life, work has been done on requirements and capabilities for trusted cloud providers[18] with an accompanying

[18] https://docs.google.com/document/d/1Auuxtj7MnCCMzdkBu9stvQ_Qanyu4CK5BExzae26vPM/edit

document about Sensitive data concerning health and trusted cloud[19]. Legislative initiatives under the European Data Strategy, including the proposed Data Governance Act and the upcoming European Health Data Space (EHDS) regulation[20], which will clarify access and re-use rights as well as conditions, e.g., to process in secure processing environments, may also provide opportunities for sharing data with greater certainty. This does not preclude harmonisation between member states. If the technical specification and architecture for EHDS conformant IT infrastructure will be set out by the EC as currently mentioned in the EHDS draft, Computational Infrastructures MUST conform with these legal requirements IF they want to act as an EHDS Data Holder, Data User, or Health Data Access Point.

Computational Infrastructure offers important opportunities for biomedical research collaboration, thanks to the opportunities for secure, monitorable, scalable, and remotely-accessible data processing in a highly interoperable processing environment. Realising this opportunity will require attention to associated data protection and trust issues, including appropriate security standards, data processing agreements, and frameworks managing international transfers to third country cloud providers. Based on the data use conditions, it is important to know the physical location of data being processed, especially if the data sharing is restricted to a certain geographic location(s).

More detailed work on ethical and legal considerations was part of D2.1 (First draft on legal framework for technical safeguards with a focus on cloud usage)[21] and D2.2 (Framework of modular contract clauses for HRICs)[22]

4.2. Cybersecurity

Computational Infrastructures MUST comply with legal regulations regarding cybersecurity to allow use of sensitive data. This compliance MAY be proven by appropriate certification programmes.

4.3. Certification

A researcher as controller of sensitive data and user of a secure Computational Infrastructure MUST be aware whether the home organisation might impose particular rules, and consequently the infrastructure provider which MUST be followed by all parties involved.

19 https://docs.google.com/document/d/1ZMyv0VqQclBN_CxPGE6THOtsK9gwdQdNuGBzF5X_33U/edit#heading=h.gjdgxs
20 https://health.ec.europa.eu/publications/proposal-regulation-european-health-data-space_en
21 https://healthycloud.eu/wp-content/uploads/2022/08/D2.1-Cloud-Safeguards-v0.3_submitted.pdf
22 https://healthycloud.eu/wp-content/uploads/2022/11/D2.2-1.pdf

Certifications are used to provide verifiable proof that applied measures and regulations are compliant with recognized standards. In the IT security and cloud context many standards are available, addressing a broad range of aspects.

For an environment for the processing of sensible data in a GDPR-compliant way the existence of an information security management system (ISMS) is of central relevance. An ISMS defines, monitors and enforces requirements and measures the effectiveness and efficiency of technical and organisational security processes. It represents a basic prerequisite for all kinds of IT security certification standards. For cloud environments, the ISO 27001, BSI C5 and CSA STAR standards are well recognized. They overlap in many parts but are structured in different ways. If a cloud provider is able to present a certificate for either of these standards a user can safely consider the cloud environment to be managed in a proper and secure way. Since the certificate needs to be renewed every year, it will force the infrastructure provider to constantly adopt and renew its management of information security.

In any case, a contract between controller and processor is required to process sensible data off premise. A valid certification on processor side largely facilitates the formulation of standard operating procedures (SOPs), which SHOULD be part of GDPR-compliant processing contracts.

5. Next Steps

This version is revised and expanded, taking stakeholder feedback into account. Between now and the end of the project in August 2023, we will have further discussions on more specific issues and to get a holistic view of the different opinions and experiences of the technical stakeholders in this project. If demand is high, individual documents on the main sections of this report may be compiled with detailed views and pragmatic examples on how to satisfy the general requirements covered in this deliverable.

Acronyms and Abbreviations

- AAI – Authentication, Authorization, and Identity
- CA – Consortium Agreement
- BSI – Federal Office for Security in Informatics, Germany
- BSI C5– BSI's Cloud Computing Compliance Criteria Catalogue
- CI – Computational Infrastructure
- CPU – Central Processing Unit
- CSA – Cloud Security Alliance
- CSA STAR – Cloud Security Alliance Security, Trust, Assurance and Risk Framework
- CSC – IT Center for Science (Finland)
- D – deliverable
- de.NBI – German Network for Bioinformatic Infrastructure (Germany)
- DoA – Description of Action (Annex 1 of the Grant Agreement)
- DUO – Data Use Ontology
- EB – Executive Board
- EC – European Commission
- EHDS – European Health Data Space
- EGI – European Grid Initiative
- ELIXIR – European life-sciences Infrastructure for biological Information
- EOSC – European Open Science Cloud
- FAIR – Findable, Accessible, Interoperable, Re-Usable Data
- GA – General Assembly / Grant Agreement
- GA4GH – The Global Alliance for Genomics and Health
- GDPR – General Data Protection Regulation
- GPU – Graphics Processing Unit
- HPC – High Performance Computing
- HRIC – Health Research & Innovation Cloud
- IPR – Intellectual Property Right
- IP – Internet Protocol, IP addresses
- ISMS – information security management system
- KPI – Key Performance Indicator
- M – Month
- MS – Milestones
- PI – Primary Investigator
- POSIX – Portable Operating System Interface
- PM – Person month / Project manager
- RAM – Random Access Model
- RI – Research Infrastructure
- SSH – Secure Shell
- SOP – Standard Operating Procedures
- S3 – Simple Storage Service
- V(X)LAN – Virtual (Extensible) Local Area Network
- VM – Virtual Machine
- WMS – Workflow Management System
- WP – Work Package
- WPL – Work Package Leader