PHIRI

Population Health Information Research Infrastructure

Upgrade option for PHIRI federated infrastructure extension

Deliverable 7.4, 23.10.2023

IACS Instituto Aragonés de Ciencias de la Salud

Francisco Estupiñán-Romero

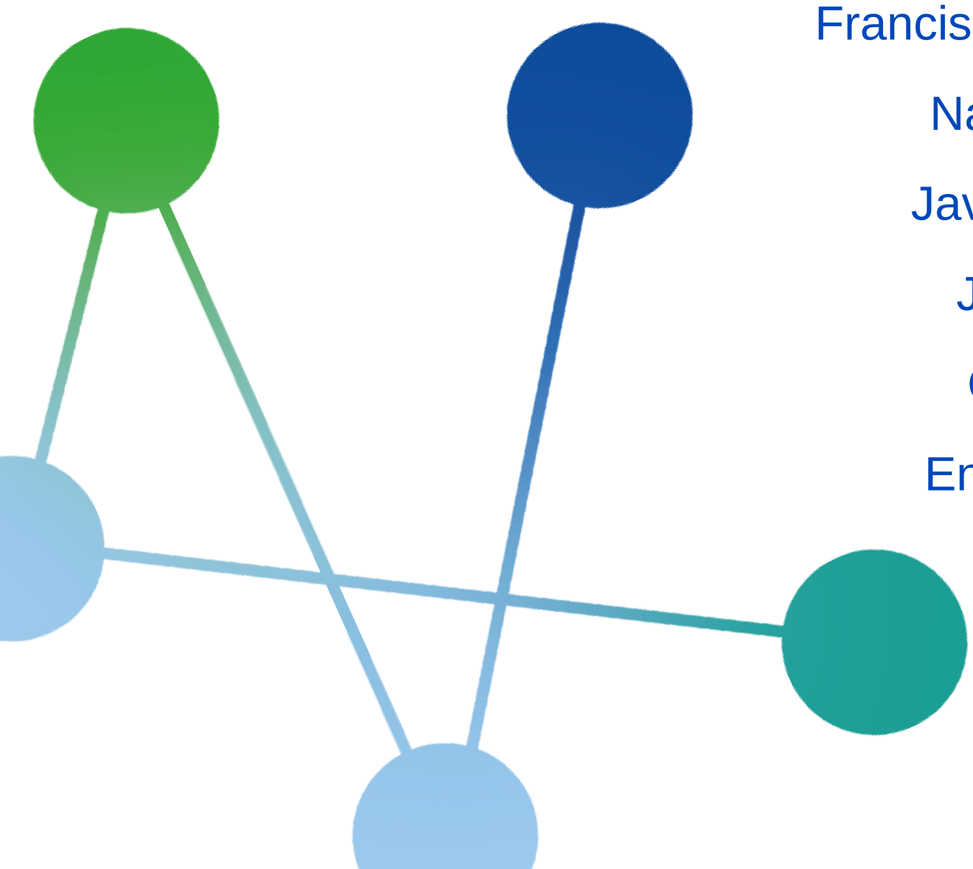
Natalia Martínez-Lizaga

Javier González-Galindo

Juan González-García

Carlos Tellería-Orriols

Enrique Bernal-Delgado



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101018317

Contributors

WP6 & WP7 co-leads	Institution
Ronan Lyons	Swansea University, UK
Simon Thompson	Swansea University, UK
Pascal Derycke	Sciensano, BE
Charles-Andrew Vande Catsyne	Sciensano, BE
Martin Thissen	Robert Koch Institute, DE

Acknowledgment

We are indebted with the contributions from the participants in the developers' forum within the work package for the technological development of PHIRI.

Table of Contents

Executive summary

PHIRI aims to facilitate and support open, interconnected, and data-driven research by sharing cross-country COVID-19 population health information and exchanging best practices related to data collection, curation, processing, use, and reuse following ELSI and FAIR principles. It has the objective: 1) to provide a Health Information portal for COVID-19 with FAIR catalogues on health and healthcare data, 2) to provide structured exchange between countries on COVID-19 best practices and expertise, and 3) to promote interoperability and tackle health information inequalities.

Within PHIRI, WP7 has developed the technological substrate for implementing a federated research infrastructure that allows mobilising potentially sensitive data to respond to multiple research queries in multiple sites, while preserving GDPR principles.

Task 7.1 demonstrated the suitability of this federated research approach in the production of research outputs for a rapid policy response to COVID-19; in particular, this task established the governance mechanism and implemented the technological solutions to respond to four use cases carried out in multiple sites (countries). The following use cases took place:

1. Impact of COVID-19 on health care in more vulnerable populations (Use case A).
2. COVID-19-related delayed care in breast cancer patients (Use case B).
3. Effects of the COVID-19 pandemic on maternal and newborn health (Use case C).
4. COVID-19 related changes in mental health care (Use case D).

Deliverable 7.1 (13/05/22) already contained features of the mid-size prototype of the PHIRI Federated Infrastructure to serve use cases that were already implemented at that point in time. Previously, we described the building blocks of developing the medium-scale (mid-size) PHIRI federated research infrastructure. This entails the development of a common data model for each use case, and the implementation of scripts for data quality assessment and analyses, all of them contained within a Docker Image. The Docker Image is a technological solution that is a secure environment for the implementation of the Federated Research Infrastructure (FRI) across nodes.

In Deliverable 7.2 (14/04/23) we complemented the development of the mid-size prototype of PHIRI Federated Research Infrastructure with the development of a demonstrator piloting the human-to-machine interface. This interface automatically builds an updated version of an international comparable report for use case B. This report is generated each time a partner completes the local analysis and shares their aggregated outputs.

In Deliverable 7.3 (29/09/23), we provided the final developments made for a fully operable FRI. Specifically, the implementation of the technological stack to enable users' authentication using the EGI AAI platform, thus guaranteeing secure access to the use of the user-to-machine interface referred to in the deliverable 7.2.

In this Deliverable we present the analyses done under Task 7.4 and Task 7.5, around the future directions that should be taken in the PHIRI FRI development, taking into consideration not only the technical aspects but also the regulatory requirements, semantic elements and its location in a rapidly changing environment, especially regarding those elements of interaction with the European Commission's European Health Data Space regulation as well as with other data sharing infrastructures currently under development, such as the Genomics Data Infrastructure or the European Cancer Image Infrastructure.

1. The need for a federated research infrastructure in population health research

The extensive and continuous reuse of sensitive health data (e.g., clinical data, electronic health or clinical records, administrative and claims data) could enhance the role of population health research on public decisions[1].

Nowadays, sensitive health data reuse is still scarce. Common arguments to explain this paucity are multiple: data privacy and safety issues; the difficulty in discovering data sources of value; complex access rules; uneven data quality; or, limited computational capacities[2],[3].

In the Joint Action on Health Information (InfAct), the implementation of a very small-scale federated network demonstrated how to mobilise sensitive data and yield the expected research outputs while complying with legal and ethical requirements[4].

Building on InfAct's achievements, PHIRI has implemented a larger-scale federated multipurpose research infrastructure involving multiple population health researchers and multiple data sources hosted in various European countries.

2. The PHIRI Federated Research Infrastructure

a. Research infrastructure architecture

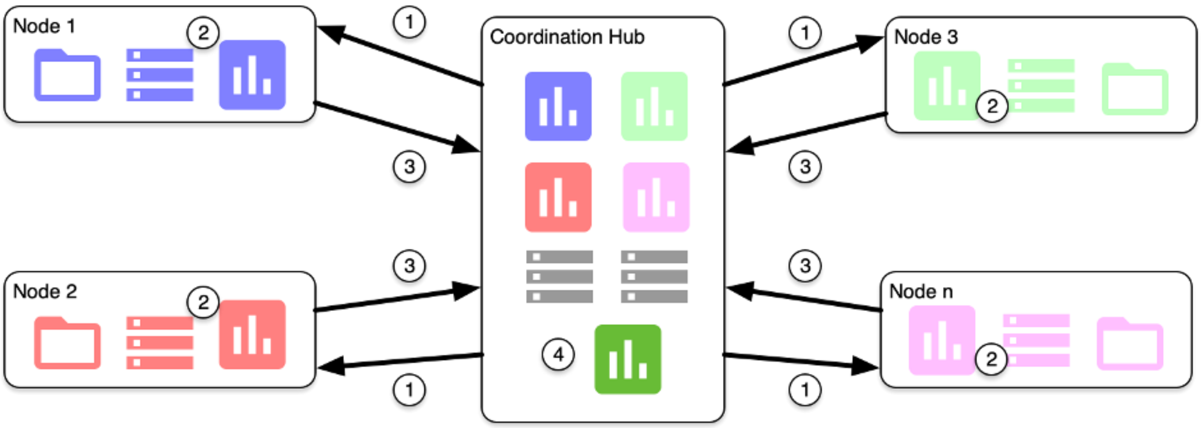


Figure 1 PHIRI Architecture and analysis flow

As in Figure 1, the PHIRI federated architecture consists of a number of country nodes (**PHIRI partners)** acting as Data Hubs, and a **central coordination hub at IACS.** Governance and roles are: (1) The orchestrating hub develops, implements and shares the analytical pipeline and provides

[1] Agyapon-Ntra, K., McSharry, P.E. A global analysis of the effectiveness of policy responses to COVID-19. Sci Rep 13, 5629 (2023). https://doi.org/10.1038/s41598-023-31709-2

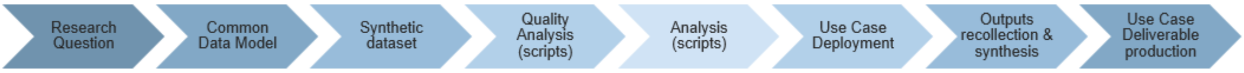
[2] J. Karacic, "Europe, we have a problem! Challenges to health data-sharing in the EU," 2022 18th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob), Thessaloniki, Greece, 2022, pp. 47-50, doi: https://doi.org/10.1109/WiMob55322.2022.9941532.

[3] Vukovic, J., Ivankovic, D., Habl, C. Dimnjakovic J. Enablers and barriers to the secondary use of health data in Europe: general data protection regulation perspective. Arch Public Health 80, 115 (2022). https://doi.org/10.1186/s13690-022-00866-7

[4] González-García, J., Estupiñán-Romero, F., Tellería-Orriols, C. et al. Coping with interoperability in the development of a federated research infrastructure: achievements, challenges and recommendations from the JA-InfAct. Arch Public Health 79, 221 (2021). https://doi.org/10.1186/s13690-021-00731-z

support to the federated research infrastructure for its deployment; (2) Nodes deploy the pipeline local analyses on their premises; (3) intermediate outputs (e.g., aggregated results or models) obtained from the local analyses are sent to the Central Hub, so, no sensitive data is shared across the federation of nodes but only digital objects; (4) the central hub can perform meta-analyses with those intermediate outputs if required.

b. Analysis methodology



Research Question
Common Data Model
Synthetic dataset
Quality Analysis (scripts)
Analysis (scripts)
Use Case Deployment
Outputs recollection & synthesis
Use Case Deliverable production

Figure 2 PHIRI Analysis methodology

During the project, it has been developed a federated analysis methodology to develop population health studies that build on top of the infrastructure. In this methodology, the studies start by defining the research question to be answered. Once it has been agreed the orchestrating hub develops a prototype, a stepwise approach, aiming for full interoperability at any stage of the process; thus, starting with the formalisation of the research query as a common data model for all the nodes, following with the deployment of the analytical pipeline on-premise to run the analyses and, finalising with the collection of the research results and their publication.

The detailed steps of the methodology are depicted in Figure 2, and are as follows:

1. Formalise the Research Question to be answered in the study. This is typically initiated by a researcher in one of the partners in the federation (the use-case leader in the project prototypes), who then asks the rest of the partners to join such a study using their local data, if they agree they become participant nodes.
2. Iteratively build a Common Data Model specification with the contribution of all participants in each research question.
3. Generate a synthetic dataset following the specifications of the agreed common data model, for code preparation and testing.
4. Iteratively develop and test the scripts for the Data Quality Analysis using the synthetic data set, tailored to the quality requirements of the research question (i.e., following a fit-for-use approach to data quality assessment),
5. Iteratively develop and test the scripts for the statistical analysis using the synthetic data set. The study should implement the methodologies and analytical techniques enabling the response to the research question,
6. Containerise the research digital objects using a reproducible solution (e.g., using software containers such as Docker), deploy them in the participant node premises and execute them using the participants' node data. The deployed elements contain the common data model definition, the synthetic dataset (as data example), the data quality analysis and statistical analysis scripts, as well as all possible dependencies these elements may have, to guarantee the correct execution of the analyses and the generation of the outputs.All these digital objects are encapsulated as a single web application named “PHIRI App” with a neat and usable interface that facilitates its execution by the partners.
7. Collect the *local outputs* of each participant, i.e., the outputs each partner generates on their premises with their own data, for the promoter of the question to summarise or meta-analyse.
8. Finally, create the final reports, based on the summary or meta-analysis of the collection of local outputs.

Note that for “feeding” the reproducible environment and executing the data quality and the analysis scripts, the participant nodes should access their own data and transform it to the common data model they agreed upon.

c. Latest developments of the PHIRI FRI

This subsection briefly covers the latest development of the PHIRI FRI implemented within the PHIRI project. Deliverable 7.3 presents extensively the extent of the technical developments.

In summary, the implementation work done during the PHIRI project has foreseen an incremental approach in the communications between nodes. Figure 3 depicts this incremental approach, and as follows:

- Tier 0: the communications between the nodes are manually executed by the partners on each endpoint. In practice, the aggregated results are sent by email from each partner to the coordinator node and the responsible in the coordinator node gathers them manually. This solution was implemented in the first version of the use cases.
- Tier 1: the partner nodes send manually the aggregated results to the coordinator node using a web application hosted in such a node that gathers all the results. This removes the human intervention at the coordination node side. This solution was implemented in the use case demonstrators and it is now being polished using the EGI AAI platform to guarantee the users' authentication.
- Tier 2: using the AAI credentials, the PHIRI App connects directly to the coordination node via an API to submit the aggregated results. This solution was not tested during the PHIRI project.
- Tier 3: using the AAI credentials and distributed algorithms, the partner nodes and the coordinator node can execute complex federated learning algorithms. In the initial steps of the development of the PHIRI RI, different federated learning algorithms were tested internally in the IACS premises[5]. No other tests were performed to connect different partners, as the AAI solutions were not already in place.

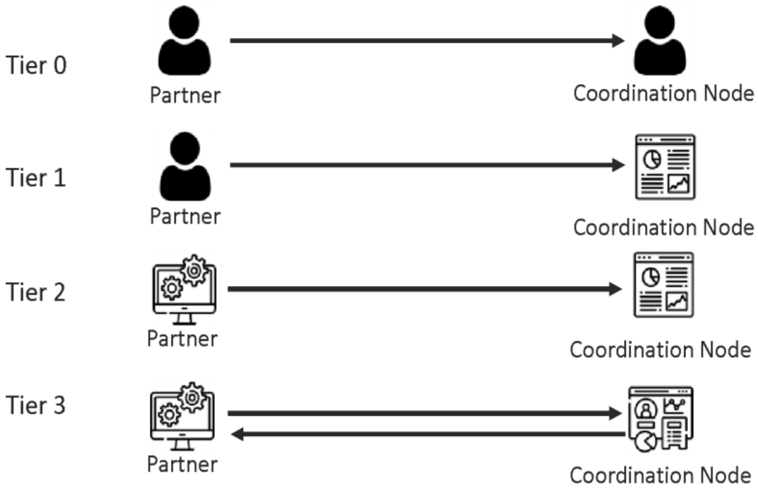


Figure 3 Communications handling in the PHIRI FRI

[5] Luo, C., Islam, M.N., Sheils, N.E. et al. DLMM as a lossless one-shot algorithm for collaborative multi-site distributed linear mixed models. *Nat Commun* **13**, 1678 (2022). https://doi.org/10.1038/s41467-022-29160-4

3. Developer’s and Users’ experience in the deployment of the PHIRI research infrastructure

As part of the activities of the work package, interaction with developers and users was constant and surveys were formalised to capture their experience in the use of the infrastructure in the deployment of the use cases (see work package 6 achievements).

A specific survey was conducted with questions covering the experience with the PHIRI App as well as the experience in a number of topics specific to the use-cases, namely: accessing the required data, transforming the input data to the common data model (ETL), deploying the container and using the web application.

a. General container and PHIRI App-related questions

As can be seen in the following figures (4 to 7), there has been a positive experience running the container solutions, having 70% of partners capable of running it (Figure 4, summing ‘Yes’ and ‘Yes, but currently we host…’ and ‘only if strict security requirements…’). This is a very good figure, considering that the exposure to containerised software was not that high at the beginning of the project, as can be seen in Figure 7, where only 4 out of 20 participants self-reported being advanced or proficient users of Docker. Furthermore, more than 60% of the partners that effectively executed the containerised solution, identifying themselves as intermediate or advanced users, reinforces the view that their participation in this project helped to increase their capacity in this highly demanded technology.

Finally, in the specific question regarding the execution of the PHIRI App, Figure 6, as mentioned before, the web application deployed within the container, just 12% were not capable of running it (a single partner), which represents only 2 out 17 respondents.

All in all, these two questions support the choices made when selecting the technological solutions for the PHIRI FRI.

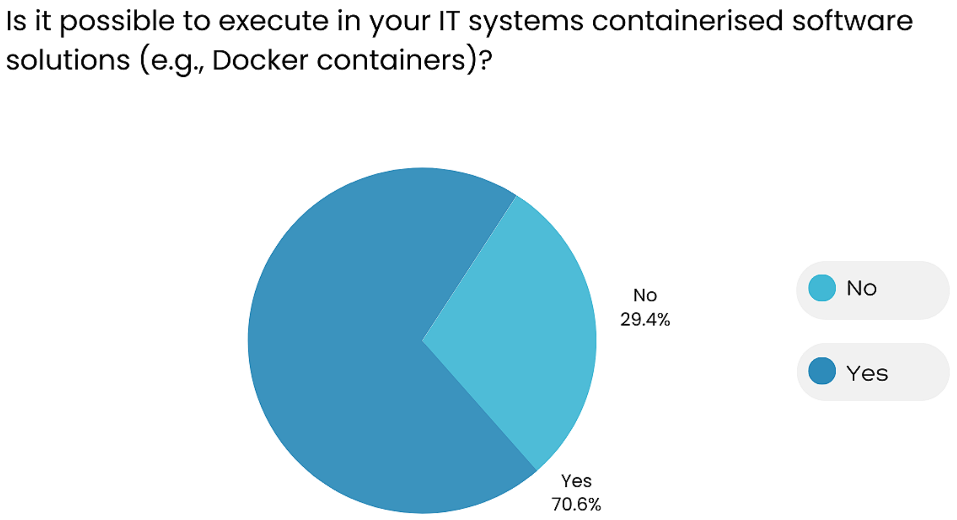
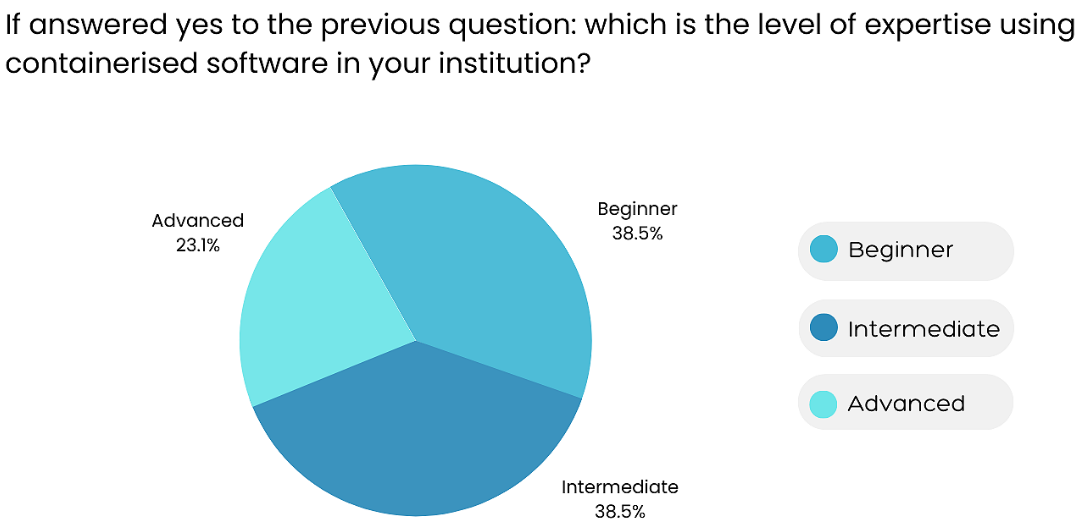
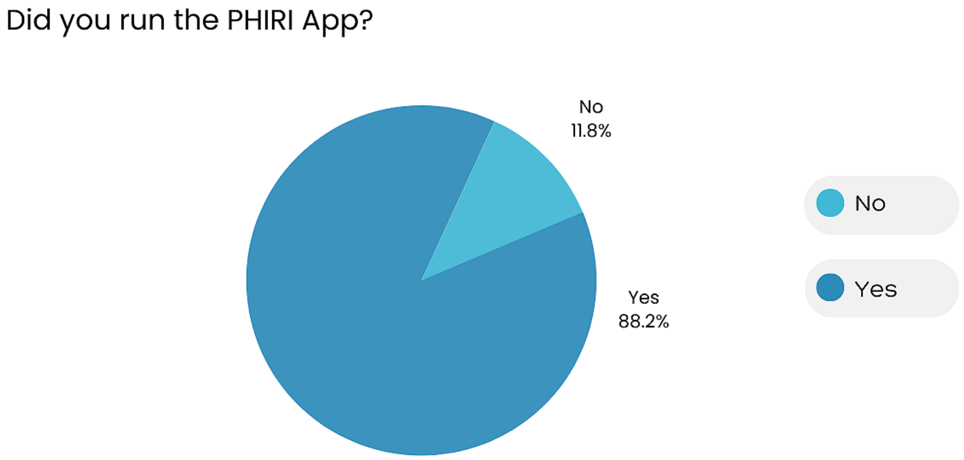


Figure 4 Container executing capabilities.



If answered yes to the previous question: which is the level of expertise using containerised software in your institution?

Figure 5 Containerisation solutions expertise



Did you run the PHIRI App?

Figure 6 PHIRI App execution capabilities

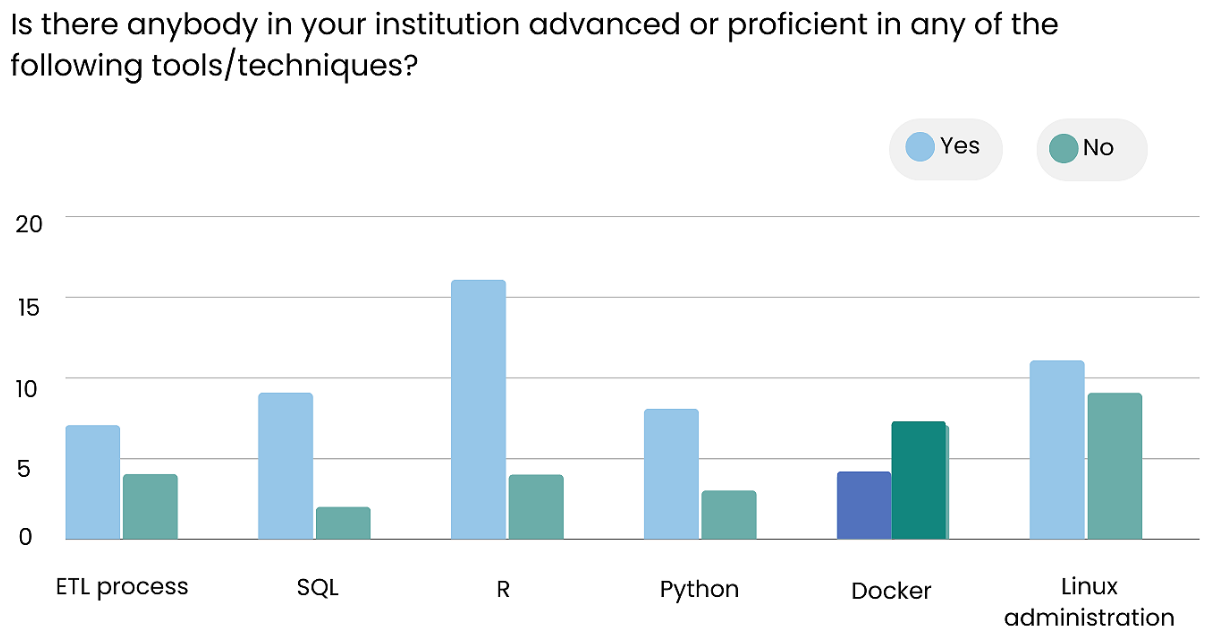


Figure 7 Self-reported IT capabilities at the beginning of the PHIRI project

b. Use-cases specific questions

Regarding the experience running the use-cases, it is also possible to confirm that the elements related to the container deployment and PHIRI App execution received positive reviews. Evaluating the plots in Figures 8 to 11, the responses that capture the partner experiences on each use-case, show somewhat or extremely hard experience when accessing the input data and when transforming the data into the common data model.

These two elements were expected beforehand and reflect classic situations observed in research, not necessarily federated, based on health data sharing in multiple settings; thus: 1) the data governance is not always clear in all setting for all types of data, which cause delays and issues to access to the real data; and, 2) the work to adequate the local data model and its semantics is heavy and complex.

The analysis presented in Section 4 focuses partly on these two main gaps and how they can be addressed from multiple perspectives.

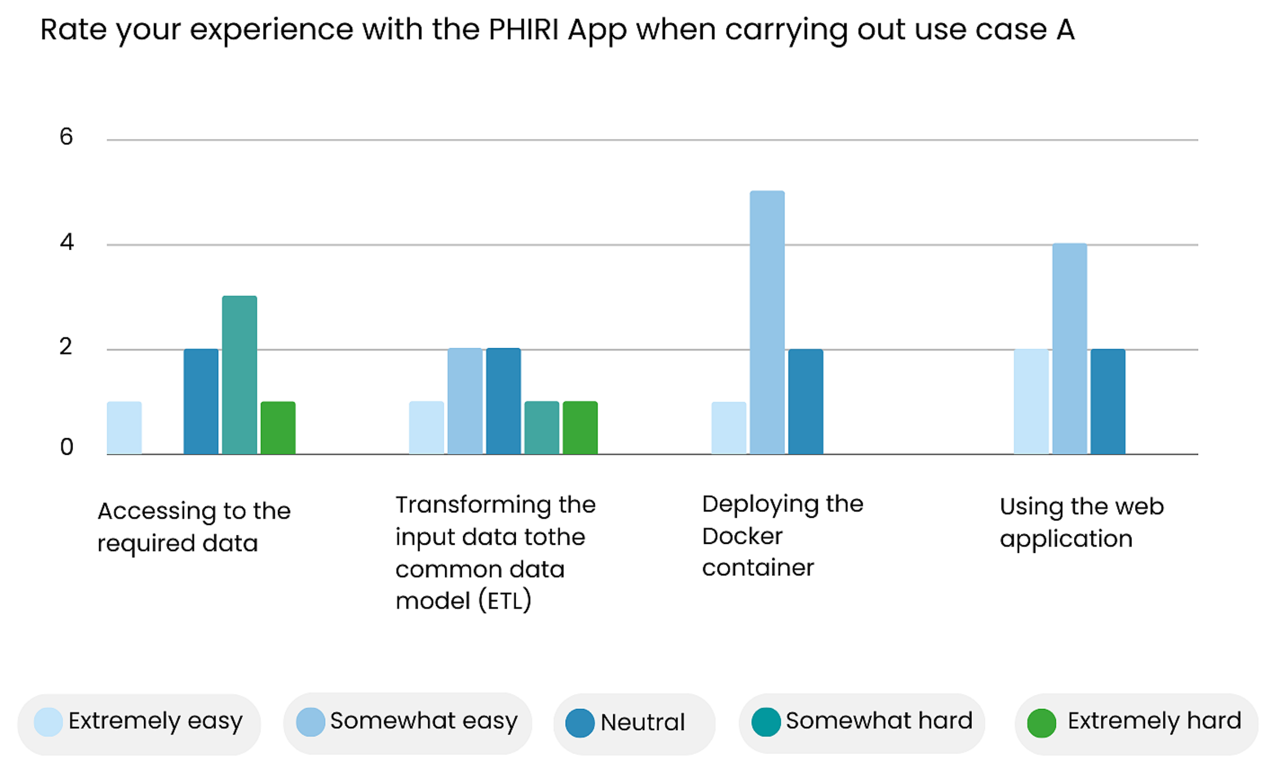


Figure 8 Partners' experience running use-case A

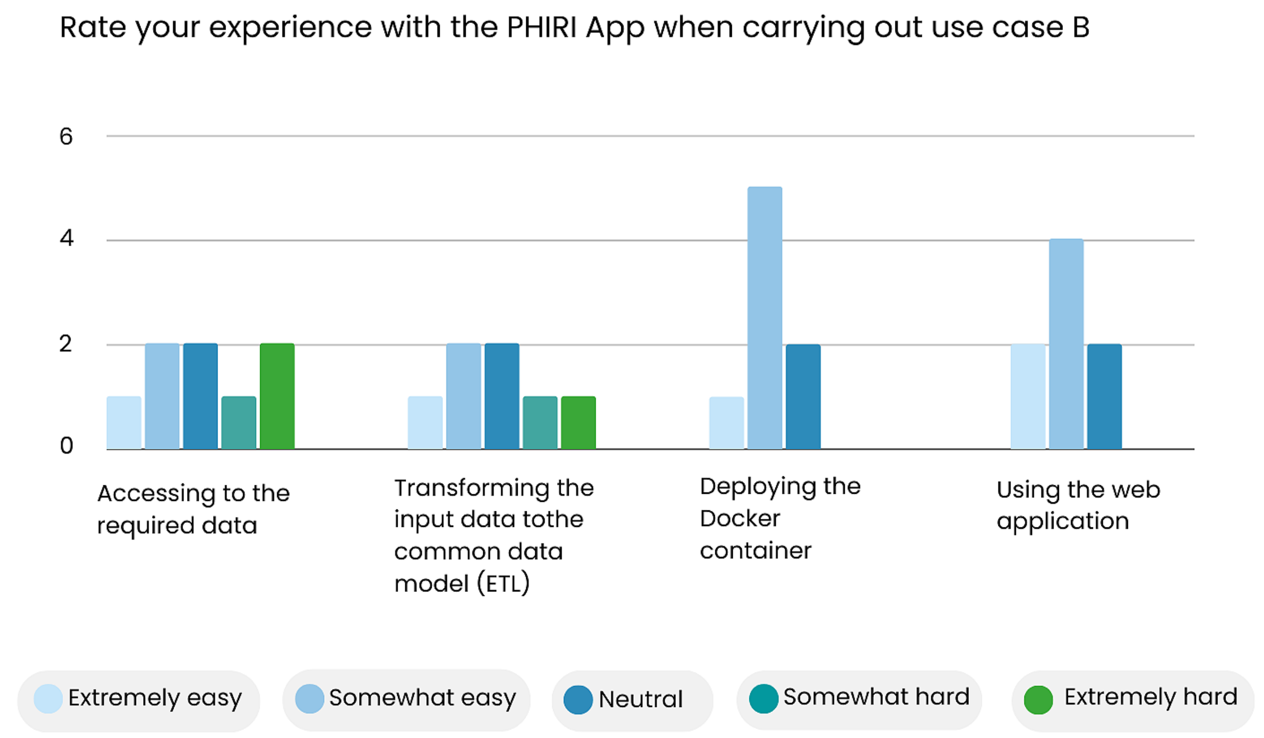


Figure 9 Partners' experience running use-case B

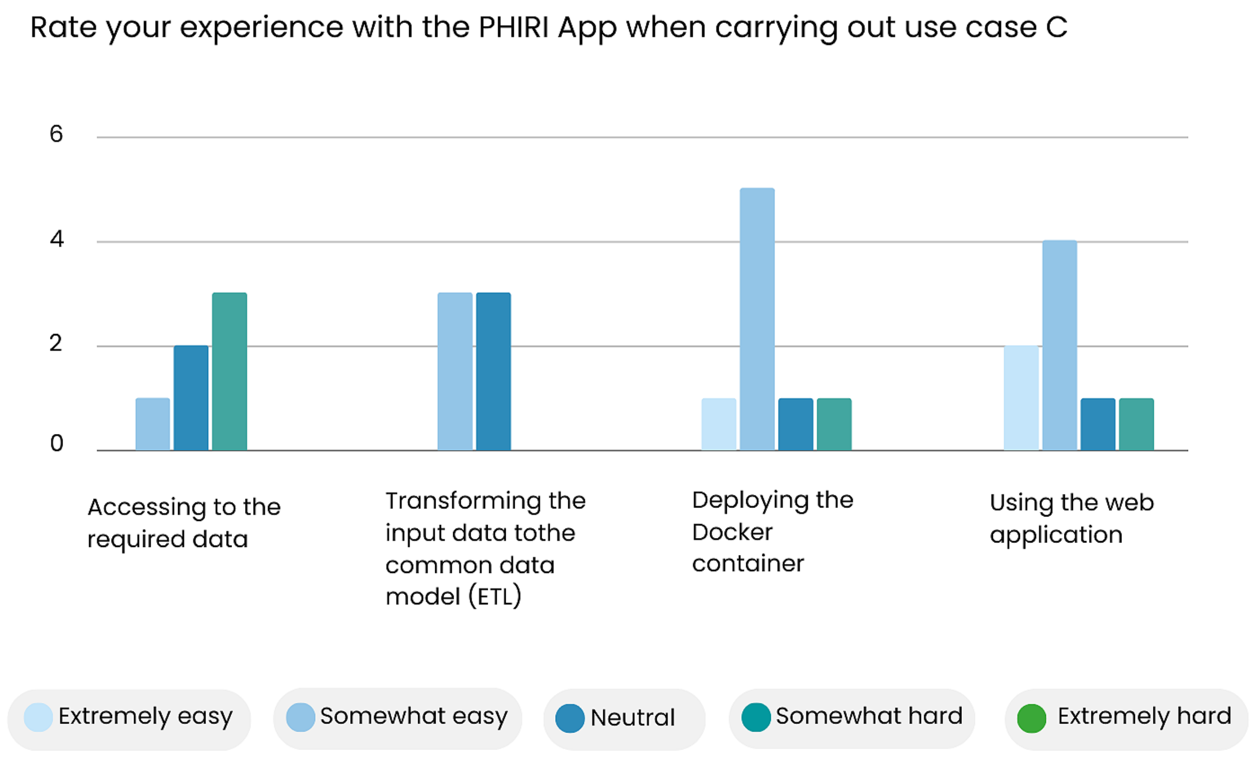


Figure 10 Partners experience running use-case C

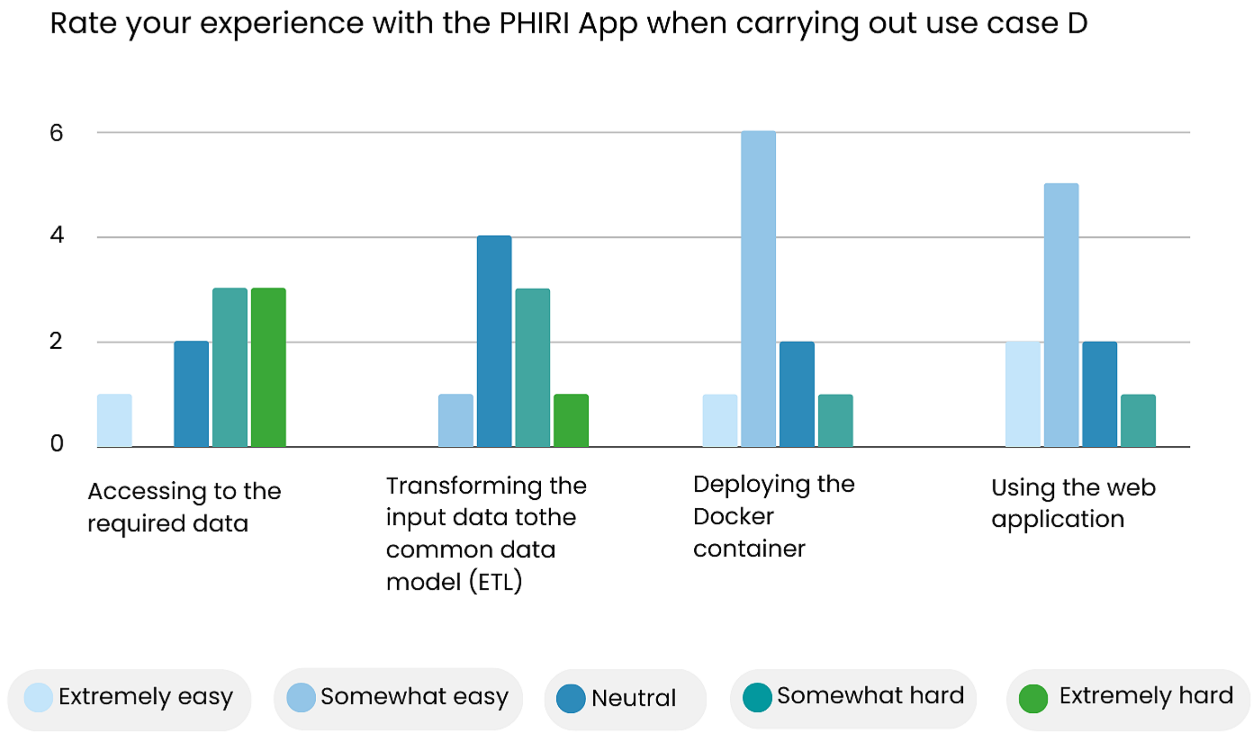


Figure 11 Partners' experience running use-case D

4. Missing gaps and opportunities – What's still lacking?

The main discussions taken in Task 7.5 of the WP7 were related to the analysis of the current prototype of the FRI, its exploitation in the PHIRI methodology, what are the missing gaps when doing this map from the methodology to the architectural solution, and the possible opportunities to the further development of the infrastructure.

The discussions were taken internally, with the participant partners in different internal fora (dedicated tasks workshops or the Developers' Forum) and in a workshop with external stakeholders, which took place on the 26th of June 2023, which represented milestone 7.5.

This section synthesises the discussions and conclusions reached in these different fora.

a. Technical level

i. Interactive initiation of the research question

The current interaction between the central hub and the participant nodes is humanly initiated by coordinating the possible participants in each study. This is done, for example sending an email, setting up a set of meetings to agreeing on a research question. This initial step is expected to always have a high human intervention.

On the other hand, the further steps are subject to be automatised, for example, using an assisting tool to design the common data model, its integration with the health information portal to facilitate the discovery of data, the generation the synthetic data and the creation of the scripts, both for the quality control and for the analysis.

1. *Discussions*

In the workshop discussion, it was highlighted that in other analysis networks, such as EHDEN[6], the initiation of the studies follows the very same methodology. In the EHDEN case, the 'data partners' (the participant nodes), can be contacted through an email distribution list. In this distribution list, the study proposals that have been authorised by the EHDEN governing boards are announced, and data partners can join with their data.

This way of operating could be improved by providing first a platform where the research questions are formalised using the appropriate languages and ontologies. This platform will serve to check which are the open questions in the research infrastructure, exchange between the different research involved and start, for example, pre-designing the common data models.

ii. Automatic deployment and execution of the scripts

Once the scripts have been created (both the quality control and the analysis) they should be deployed to the participating nodes of the study. This corresponds to the sixth step in the methodology (Figure 2). In the current version, this deployment has three stages: 1) include such scripts and the PHIRI App that serves as a front-end in a Docker container; 2) upload the resulting container image to a repository (currently, Zenodo); and, 3) indicate to the participating partners the location of the image and the instructions on how to install and execute it.

[6] https://www.ehden.eu/

1. *Discussions*

This process should be automatised to facilitate the encapsulation of the scripts in a standardised manner. First, it would be necessary to create a development API, to avoid repetition on certain features. This has already been done at the user level with the PHIRI App, which unified the development of all the use cases. The API of the PHIRI App should be opened and documented to support the major scripting languages (R and Python mandatory), it can also use workflow managers for easing the creation of much more sophisticated federated scripts (see next section). This API should serve to facilitate further containerisation using a set of scripts and tools for continuous integration. Ideally, all this process should be supported by a code control system that guarantees the proper development.

Finally, the deployment of the container images should be assisted by a web application that runs on the participant premises, which is able to download the images and present the containers required for a given study, and then run the PHIRI App with the specific analysis codes.

In the workshop, there was an extensive discussion on the use of container solutions, beyond the regular use of Virtual Machines (VMs), as containers facilitate the encapsulation of the environments and dependencies while being more maintainable than VMs. Additionally, it introduced some solutions that help manage the packaging of the requirements in containers for specific architectures[7].

iii. Federated algorithms

In the current incarnation of the PHIRI FRI, the analysis follows a scheme where the analysis scripts are executed on each node with their local data and then the results are sent back manually to the central hub where a meta-analysis is performed. This solution has two limitations: first, there is a human intervention to manually upload the local results; and second, this is a “one-shot” approach, which is far from the current state-of-the-art federated analyses algorithms, that require multiple communication rounds to produce their results. So, for the sake of analysis power, much more complex analysis capabilities (e.g., federated learning) should be provided, as we already explored in the initial steps.

1. *Discussions*

It would be necessary to provide federated learning capabilities to the analysis scripts. As stated, federated analysis algorithms rely on multiple data exchange loops aiming to refine the results, usually in a synchronous manner (i.e., the delays in the communications between nodes should be in the order of milliseconds). These capabilities will rely on using a set of software libraries and runtimes for parallel processing[8]. Parallel processing libraries, such as COMPSs[9] whose developers participated in the workshop, ease the development of the analysis scripts by implementing the coordination protocols to build the analysis algorithms or even part of the algorithms themselves, basically in those elements referring to the data exchange between the participants' nodes.

In all cases, the runtimes for parallel processing should help to guarantee the privacy of individuals whose data is analysed, for example, controlling the data exchanges of the partial results and securing the data transports, as introduced in the following section.

[7] https://github.com/eflows4hpc

[8] Klemm, Michael and Cownie, Jim. *High Performance Parallel Runtimes: Design and Implementation*, Berlin, Boston: De Gruyter Oldenbourg, 2021. https://doi.org/10.1515/9783110632729

[9] https://github.com/bsc-wdc/compss

iv. Secure transport

Secure transport is a non-functional requirement to guarantee secure communications between the nodes and the central hub. Secure communications are intended to avoid main-in-the-middle attacks, where the perpetrator captures the messages sent between two of the partners. If the communications are secure, the messages cannot be read.

1. *Discussions*

As part of the PHIRI project activities, eDelivery was planned to be tested, but for time constraints it wasn’t tested finally. eDelivery is a secure messaging transport protocol developed by the European Commission. It is the standard of choice selected for the HealthData@EU pilot and probably the future implementation of the final HealthData@EU infrastructure of the European Health Data Space (see Section 3). The underlying technology of eDelivery is based on a compendium of security technologies[10] such as Transport Layer Security (TLS)[11].

In the workshop, AI-SPRINT[12] where European researchers are experimenting with highly secure environments for AI, was debated. The architectural proposal for this project includes secure transport and secure analysis. On its side, UK researchers are developing the TRE-FX project[13], to create a network of Trusted Research Environments (TREs), i.e., highly secure analysis systems for analysing data, where communications between each TRE communicate securely with the rest of the TREs in the network.

b. Semantical level

i. Common data models (CDMs)

In the PHIRI methodology, once the research questions have been agreed, there is a process to design a common data model. The outcome of this step is the definition of the common data model necessary for answering such a research question in terms of required variables, data types, semantics of the variables, level of requirement, etc. It will be then the participating nodes that have access to the data the ones responsible for transforming their data to conform with the data model.

Even though it would be always required to transform the data to a format that will be later used by the analysis solution (statistical application or dedicated analysis script), in the PHIRI methodology, the participant partners need to adapt the data to the CDM for each study. This may limit the system's scalability regarding studies a node can participate in parallel.

1. *Discussions*

The main discussions within the PHIRI activities have been around the appropriateness of pushing for a “general purpose” common data model, whose semantics are broader than a single research question, having an *a priori* agreement on how the data has to be organised, and thus easing the way the data is manipulated and analysed. In this way, the data transformation is done only once per participant node, from the data they have access to this CDM and, for each study, a single transformation from the general purpose CDM to the research question-specific CDM can be applied to all the participant nodes.

[10] https://ec.europa.eu/digital-building-blocks/wikis/display/DIGITAL/Documentation+eDelivery

[11] Eric Rescorla. The Transport Layer Security (TLS) Protocol Version 1.3 (RFC 8446). https://www.rfc-editor.org/info/rfc8446

[12] https://ai-sprint-project.eu

[13] https://dareuk.org.uk/driver-project-tre-fx/

This approach has been treated in many fora and has been the solution adopted by the European Medicines Agency for its DARWIN Network (Data Analysis and Real World Interrogation Network)[14]. In the workshop, the managers of DARWIN EU Coordinators Centre presented how Observational Medical Outcomes Partnership[15] (OMOP) CDM will be used in this project and has been already used in the EHDEN network, as commented previously. Currently, the OMOP CDM is gaining much attention, as it is becoming the *de facto* standard for storing clinical data. Even though it may still lack expressivity for some specific clinical elements, the concept of an episode that aggregates multiple contacts is still under development. The governance body that drives its development, the Observational Health Data Sciences and Informatics (OHDSI) initiative[16], and the community of users are very active, and there are a large number of tools that help in minimising the usually huge cost of transformation from the existing data models to OMOP.

All in all, the decision of using specific CDMs per research question or shifting towards general purpose CDMs needs a further evaluation, considering the transformation costs, the literacy and the further exploitation of the standard of choice.

c. Organisational level

i. Sustainable computing capacity provision

As introduced in the first section, the PHIRI FRI includes many pieces that facilitate the federated analysis of health data across the participant partners. These are materialised as a software stack that provides the features required to operate the PHIRI analysis methodology. The different pieces of the infrastructure operate at different levels: there is a single instance of the coordination and meta-analysis stack deployed in the central hub, while there are multiple instances of the local analysis stack deployed in the participant nodes of the federation.

To ensure a smooth operation of the FRI, the software stack needs to be deployed in reliable computing resources, and this capacity provision should be sustainable. During the development of use cases of WP6, each partner sought the system to deploy the local analysis pieces, with the help of the Central Hub team to define the requirements. In turn, the Central Hub defined its own requirements and part of the developments were deployed in the EGI infrastructure as part of the EGI-ACE[17] (Advanced Computing for EOSC) project collaboration.

1. *Discussions*

The main element in the discussion about the computing capacity provision is the need to outsource it. It is not the core business of the participant nodes to deploy and maintain a computational infrastructure, for this reason, it would be useful to contract with a trustful provider of such services. Thanks to the software solutions selected, the deployment of the software stack should be easy, independently of the provider, but the providers need to guarantee certain good practices to ensure the reliability of the overall infrastructure, considering the high sensitivity of the data.

Apart from the EGI-ACE project, where the computing (and services) capacity provision is done by The EGI Foundation, a pan-European federation of public computing facilities was discussed in the workshop. In the workshop, DARWIN representatives pointed out that, for their infrastructure, they

[14] https://www.ema.europa.eu/en/about-us/how-we-work/big-data/data-analysis-real-world-interrogation-network-darwin-eu
[15] https://www.ohdsi.org/data-standardization/
[16] https://ohdsi.org/who-we-are/
[17] https://www.egi.eu/project/egi-ace/

selected the Andrea Cloud provider[18], which is ISO 27001[19] certified, and related to information security management systems, so it can provide computing capacity using the Microsoft Azure Cloud with the highest security requirements.

In any case, the computing capacity provision has to be decided coordinately to define the minimum requirements and dedicated personnel to operate, ideally as part of the participant nodes teams.

ii. Data access provision

It is out of the scope of the PHIRI Federated Research Infrastructure the curation, management, and storage of real-world data, being the PHIRI nodes gateways to access such data. This distinction is crucial to understanding the limits of such an infrastructure and highlighting the critical need to find a common governance framework and interfaces to the data required in the analyses.

In the current FRI, the data discovery and data access mechanisms have relied on the discussion with the partners participating in the use cases and the initial pilot. In this way, the use case leader during the definition of the CDM and the participating nodes had to identify the appropriate data sources and procure the data access protocols (e.g., data requests, ethical approvals, etc.) of use in their settings. These elements have been omitted from the methodology described in Section 1.2 because those are *internal* particularities on how each participant governs data access.

Although limited in its scalability, this way of solving the data access provision has demonstrated its practical utility. Many other projects may follow the same methodological approach and may benefit from the use of the PHIRI's Health Information Portal to discover datasets of interest while relying on the data access request methods of each of them and the computation capacity provision when accessing the different datasets.

1. *Discussions*

It is important to analyse two situations for the further development of the PHIRI FRI regarding data access. First, it is the possibility of accessing synthetic datasets or data lakes, where data that replicates real-world data on its structure and semantics are available for the researchers to start the exploration of the real-world data, before deploying their actual scripts. This synthetic data should be generated by the actual data holders, as they have the local knowledge of the original datasets.

The second situation is the alignment of the PHIRI FRI with the HealthData@EU infrastructure, defined in the European Health Data Space (EHDS) legislative proposal[20], which regulates the first data space in the EU, and establishes clear procedures to access health data for its secondary use, facilitating the discoverability (that may complement the Health Information Portal), the data access request mechanisms (that may subsume the current data access procedures applied in the PHIRI nodes) as well as the computational capacity for the data analysis, in the form of Secure Processing Environments (SPEs). In this context, and considering a successful deployment of such an infrastructure, it would be required to define a clear governance to establish how PHIRI nodes may interact with HealtData@EU actors, namely the coordinator Health Data Access Bodies, to provide seamless access to real-world data as well as define the possible shared use of the computational capacity.

In Section 3, there is a large description of how the interaction between the PHIRI FRI and the HealthData@EU infrastructure in its actual conception has been conceived.

[18] https://www.andrea-cloud.eu
[19] https://www.iso.org/standard/27001
[20] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52022PC0197

d. Legal framework

The current PHIRI Federated Research Infrastructure has been designed to comply with the GDPR, being each node in charge of taking responsibility for compliance, minimising the use of potentially sensitive data (i.e., the contents of the CDMs), avoiding any personal data transmission in the federated analysis schema, and using ad-hoc synthetically generated dataset following the designed CDM to facilitate the development of the quality control and analysis scripts.

i. Discussions

The implications of the European Health Data Space regulations when it enters enforcement are yet to be seen, and implementation acts analysed; the same applies to the AI regulation (AI Act), where complying with specific provisions in the developing of AI models will be required; for example, in hypothesis generation, development of decision support systems for the healthcare sector

e. Summary of the missing gaps and opportunities for further development

Elements	**Gaps in the current version**	**Opportunities for developm't**
Research initiation	No support to formally start a research study in the PHIRI FRI	Create a platform for researchers to launch research studies and check open research studies proposed by other researchers
Deployment/Execution of analysis scripts	Missing automatisation to encapsulate the scripts and deploy the solutions.	Create an open repository of existing containers of scripts, based on open APIs, that can be reused to create new codes and deploy them.
Federated learning algorithms	No purely federated learning algorithms executed in the platform	Adapt existing federated learning algorithms to the population health context and create new ones using open APIs and runtimes
Secure Transport	No protocol for secure data transport selected	Participate in the discussion to select secure transport layer, aligning with other data sharing initiatives
Common Data Models	Scalability issues when working with bespoke data models may limit the PHIR FRI scalability	Help in the development and adoption of general purpose data models
Computing Capacity provision	No sustainable computation capacity provision selected	Adopt EU supported computation capacity providers (e.g., EOSC)
Data access provisions	PHIRI FRI participating nodes are not strictly data access providers	Establish clear governance to facilitate the data access within the infrastructure and to external data sharing initiatives.
Legal framework	Uncertainty on the development of EU Digital Strategy, Data-related Acts (EHDS, AI, DGA, others)	Participate in the discussions to clarify the legal framework and its implications for the PHIRI FRI.

5. PHIRI linkage with the major actors on health data sharing

a. European Health Data Space

The European Health Data Space is a major initiative supported by a legislative proposal published in early May 2022. One of its biggest ambitions is to develop the HealthData@EU infrastructure that will facilitate access to health-related data generated in European health systems, cohort data as well as data from health registries for its further reuse in regulation, policy making, innovation and research, the so-called secondary use of health data. De facto, the HealthData@EU will bring a key number of services on top of which PHIRI could build and extend HealthData@EU capabilities. In this scenario, PHIRI could act as a specialised façade (i.e., a knowledge broker designed as a digital experience platform) serving as an intermediary empowering body for the population health researchers.

The Data Lifecycle of Figure 12 displays the steps required from the data collection to its exploitation in a research project. This is a common view with the one presented in the TEHDAS Joint Action WP6 and WP7[21]. The Data Lifecycle is divided into two parts: 1/ Data preparation consisting of the collection, standardisation, and publication of data for its further exploitation by population health researchers. 2/ A users' journey covering all the steps a researcher should follow on a research project. This includes data discovery, the request to the health data access bodies, its effective use, and the finalisation of the research project. In this way, the data preparation part is common, and the users' journey is repeated for each research project.

It is reasonable to say that the TEHDAS Joint Action outputs[22], where PHIRI project coordination and PHIRI WP7 coordination had a major involvement, have had a substantive impact on this proposal paving the road for PHIRI to connect as a research infrastructure node.

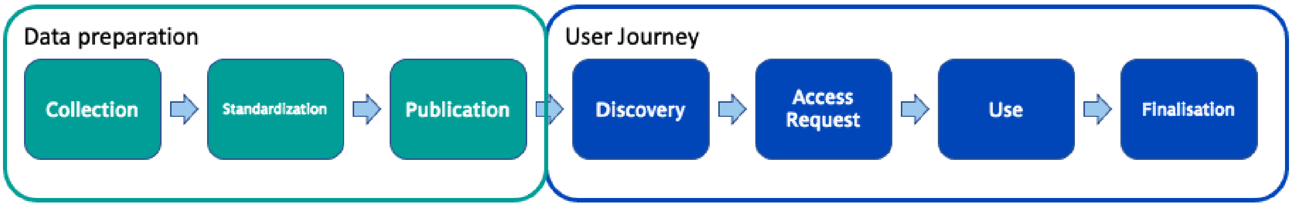


Figure 12 TEHDAS Data Lifecycle

While the EHDS regulation will guarantee the proper organisation of data holders, covering all parts of the data preparation (Figure 12), the PHIRI knowledge hub could provide the users with guidelines and best practices on how to operate the HealthData@EU services. Besides, the PHIRI knowledge hub would include the know-how to deal with the HealthData@EU Secure Processing Environments for data analysis. In this scope, PHIRI wold provide complementary services to HealthData@EU and its users, such as virtual labs and computational environments aimed to develop data analysis algorithms (in the form of scripts or HealthData@EU applications (e.g. Docker images)) that will later be reproducibly run in HealthData@EU premises. Endorsing the principles of open and reproducible science (and related technologies) and, supporting the virtual labs' enhanced user-centric approach,

[21] https://tehdas.eu/results/tehdas-suggests-minimum-technical-services-for-the-european-health-data-space/
[22] www.tehdas.eu

the PHIRI knowledge hub would also provide a capacity-building service that will allow researchers to get the most out of the services and tools provided in HealthData@EU.

Regarding the architectural design, Figure 13 depicts the vision of the interaction between the HealthData@EU architecture proposed in TEHDAS (left-hand side of the picture, in green), the vision for the PHIRI research infrastructure (right-hand side of the picture, in grey), and their interconnections. The HealthData@EU Health Data Access Bodies (HDABs) are expected to provide discovery services, data permit application services and results management services, in some cases, with the help of a Central Platform node, operated by the European Commission. For data use, HealthData@eu will provide Secure Processing Environments (SPEs), dedicated hardware and software facilities which data users will access to analyse the data they have been granted access to use. Such environments will be highly regulated to guarantee the security and privacy of the data deposited in them.

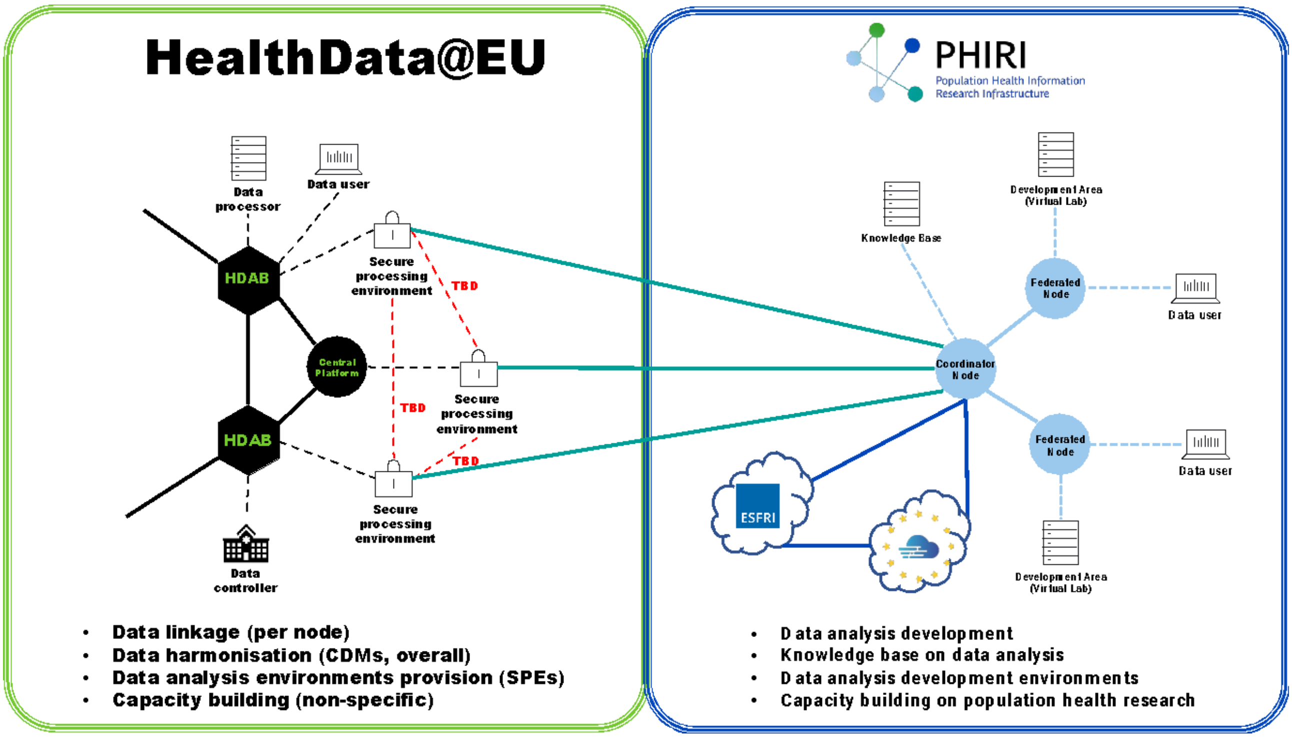

HealthData@EU
Data processor
Data user
HDAB
Secure processing environment
TBD
Central Platform
Secure processing environment
TBD
HDAB
TBD
Secure processing environment
Data controller
Data linkage (per node)
Data harmonisation (CDMs, overall)
Data analysis environments provision (SPEs)
Capacity building (non-specific)
PHIRI
Population Health Information Research Infrastructure
Development Area (Virtual Lab)
Knowledge Base
Federated Node
Data user
Coordinator Node
Federated Node
Data user
ESFRI
Development Area (Virtual Lab)
Data analysis development
Knowledge base on data analysis
Data analysis development environments
Capacity building on population health research

Figure 13 PHIRI and EHDS interactions

As can be seen in Figure 13, the PHIRI knowledge hub will federate a number of country nodes that will provide Development Areas (i.e. virtual labs and computational environments) to data users. These Development Areas will serve as “Playgrounds” to develop analysis algorithms. The necessary training, synthetic or anonymised datasets that mimic the real-world data available in the HealthData@EU will be provided within the virtual labs. The linking point between both architectures is based on the expected ability of the HealthData@EU SPEs to provide machine-to-machine interaction that will be used by PHIRI for deploying the analysis algorithms.

This design will facilitate population health researchers to develop their analytical algorithms. It will simplify the access to a computational workspace (the Development Areas) instead of going through the HealthData@EU SPEs, which will require going through the overall real-world data access. Finally, it will ease the deployment of the developed solutions from the development areas to the HealthData@EU SPEs.

b. European Open Science Cloud

The European Open Science Cloud (EOSC) is a major European initiative aimed at creating a federated infrastructure for open science and research data management. It was conceived as part of the European Commission's Digital Single Market strategy to promote open and collaborative research practices across Europe.

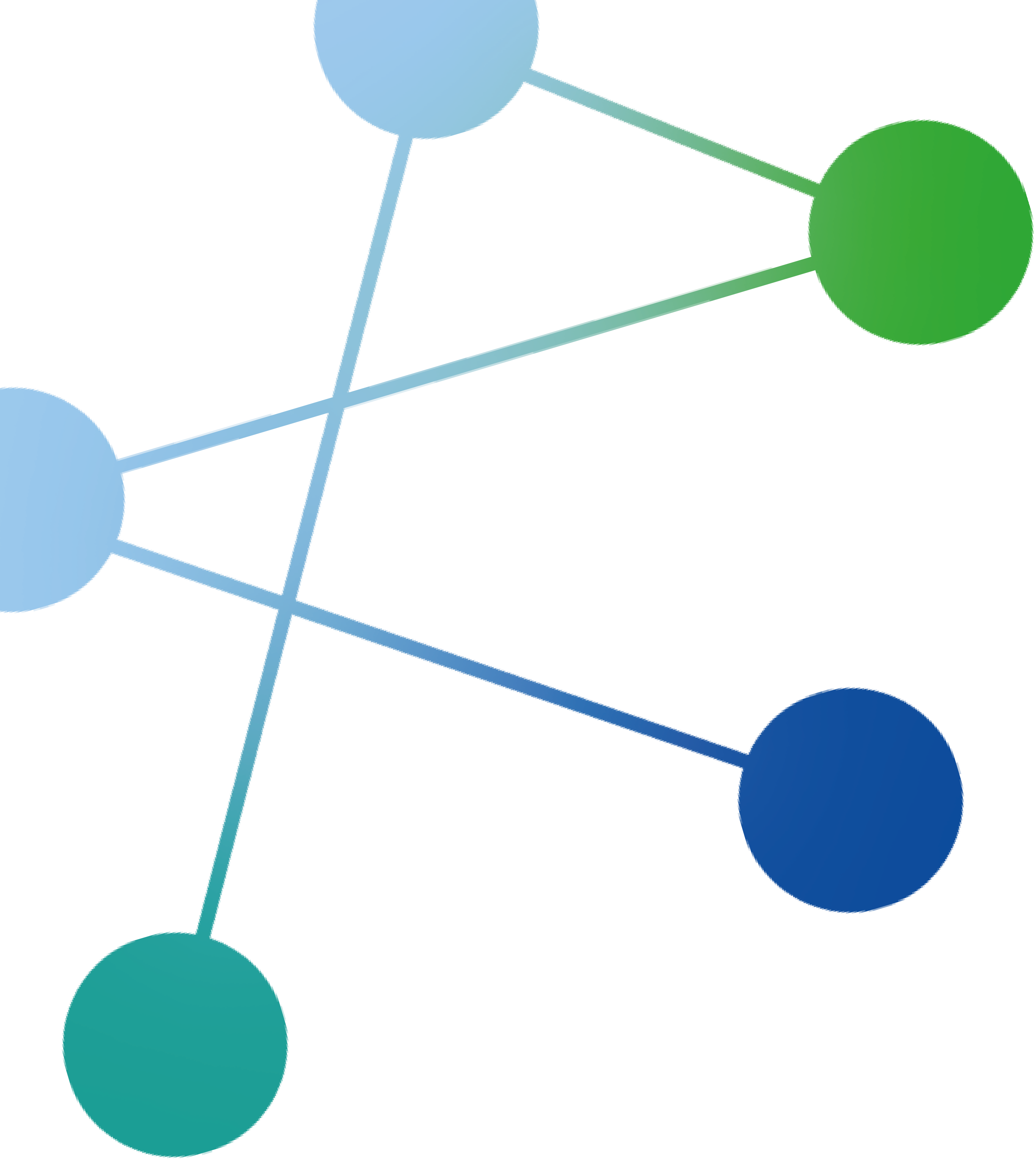
As its current status, the EOSC provides a compendium of open software, training around open science and services that range from distributed storage, solutions to support FAIR data management, and initial steps to provide computational services, and has the ambition of extending its capabilities to facilitate the data exchange within communities and between communities and thus becoming the Open Science Data Space. Further developments of the PHIRI FRI are expected to rely on EOSC services, such as the computation ones mentioned in the computation capacity sustainable provision or the AAI capabilities already explored within the EGI-ACE context.

It is also important that it is expected that PHIRI will help to tailor the “EOSC Health” community. This community is under discussion in the HealthyCloud, a linked project whose aim is to propose the Strategic Agenda to develop a European Health Research and Innovation Cloud, a project where PHIRI has major representation as it is coordinated by IACS and Sciensano, GöG and THL participate as beneficiaries. The aim of the EOSC Health community is to integrate the demands of the different research infrastructures and networks dedicated to health-based research to drive the development and curate those EOSC services required to manage and process highly sensitive data.

6. Disclaimer

Disclaimer excluding Agency and Commission responsibility

The content of this document represents the views of the author only and is his/her sole responsibility. The European Research Executive Agency (REA) and the European Commission are not responsible for any use that may be made of the information it contains.



IACS

Centro de Investigación Biomédica de Aragón (CIBA)
Avda. San Juan Bosco, 13, planta 0
50009 Zaragoza, Spain
Ebernal.iacs@aragon.es
www.phiri.eu
@PHIRI4EU