



PHIRI

Population Health Information
Research Infrastructure

Full operable PHIRI federated infrastructure

Deliverable 7.3, 23.10.2023



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

Table of Contents

Executive summary	2
PHIRI Federated Research Infrastructure (FRI)	3
Annex 1 Technical Specifications	10
Requirements for installing Docker	11
Annex 2 Technical specifications of the demonstrator (Use Case B)	13
Disclaimer	14

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.

PHIRI WP7 has developed the technological substrate for implementing a federated research infrastructure that allows mobilising 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 accurate and timely 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 contained within a Docker Image. The Docker Image is a technological solution that is a secure environment for implementing 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 when a partner completes the local analysis and shares their aggregated outputs.

An additional subtask, 7.2, was to explore using the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) to represent the information requirements specified in the use cases into a single harmonised common data model.

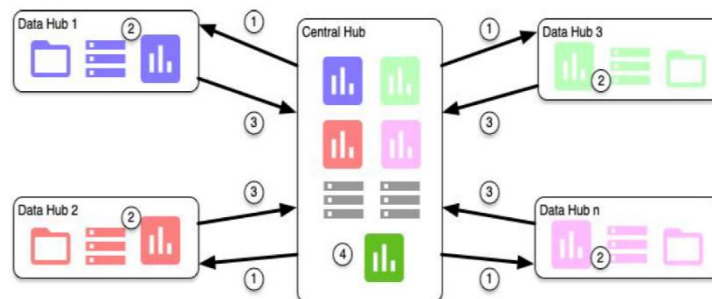
In this Deliverable 7.3 (29/09/23), we provide the final developments made for a fully operable FRI. Specifically, the implementation of the technological stack to enable users' authentication using the EGI AAI platform, thus guaranteeing secure access to the use of the user-to-machine interface referred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

PHIRI Federated Research Infrastructure (FRI)

Created by Javier González-Galindo, Francisco Estupiñán-Romero, David Chichell-Ruíz, Natalia Martínez-Lizaga, Juan González-García, Carlos Tellería-Orriols, Enrique Bernal-Delgado from IACS and Ronan Lyons from SU with contributions from Sarah Aldridge, Simon Thompson from SU and Andrea Schmidt from GOG (Use case A); Pascal Derycke, Nienke Schutte from Sciensano (Use case B); Marianne Philibert and Jennifer Zeitlin from Inserm (Use case C); Cesar Garriga Fuentes and M^a Carmen Rodríguez Blazquez from ISCIII and Mika Gissler from THL (Use case D); Martin Thissen from RKI (WP6 leader)

THE ARCHITECTURE

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



PROTOTYPING THE ORCHESTRATION OF A USE CASE

The orchestrating hub has developed 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.



This stepwise approach includes

1. Formalising the Research Question,
2. Iteratively building a Common Data Model specification with the contribution of all participants in each research question,
3. Generating a synthetic dataset following the specifications of the agreed common data model,
4. Iteratively developing and testing the scripts (code) implementing 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 developing and testing the scripts (code) of 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. Deploying the use case (or research question) using a reproducible deployable solution (i.e. Docker) on-premise in each participant by distributing the common data model, the synthetic dataset (as data example), and the data quality analysis and statistical analysis script in a reproducible way, including the environment to run the analyses and all dependencies required to manage data and to produce the expected outputs reliably,
7. Collecting the local (on-premise) outputs produced by the local analysis of each participant to summarise them or meta-analyse them by the promoter of the research question (or use case leader),
8. Finally, producing the deliverable is based on the summary or meta-analysis of the collection of local outputs.

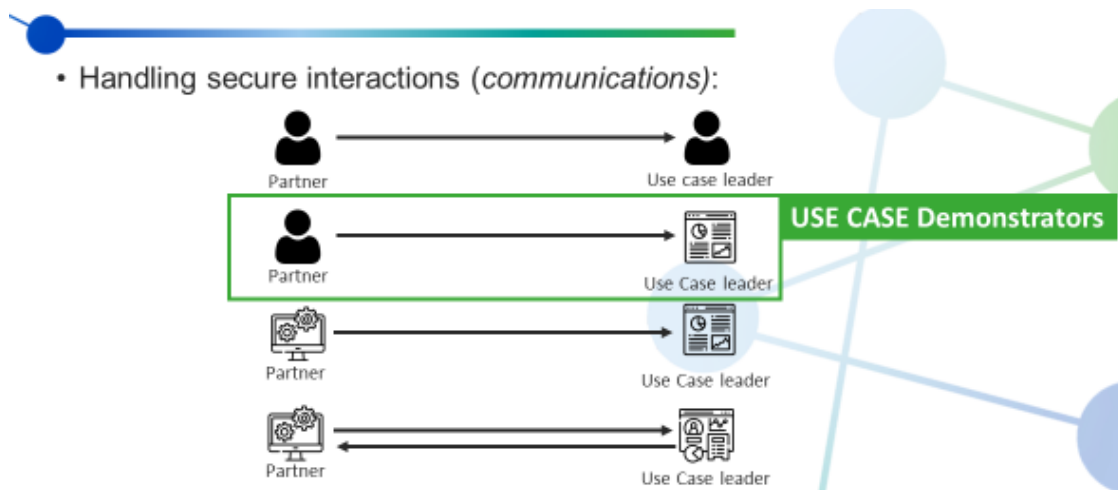
USE CASE OUTPUTS RECOLLECTION AND SYNTHESIS FOR DELIVERABLE PRODUCTION

The last step of the PHIRI federated pipeline is to collect, synthesise and compare the outputs of the local analysis run on-premise by each partner to produce insights derived from the international comparison analysis (or meta-analysis of the outputs). Therefore, outputs of the on-premises analyses are expected to be shared with the central hub to enable their synthesis and further analysis. These outputs are immediately and directly comparable across participating partners in a use case due to previous agreement in the input common data model, common scripting for the analysis and systematic data processing using a common reproducible pipeline. All use cases integrated into the PHIRI app are expected to produce the same outputs independently of their specificities. Those are a) an interactive report on the local analysis, providing immediate feedback for the participant partner for face validation and insight on the scope of the use case research question; b) an interactive data quality report (exploratory data analysis) providing information about the quality of the input data that might be required to interpret the results better; and, c) an aggregated output set that compiles the data required for further comparative analysis with other participating partners.

The communication of those aggregated datasets from the participating partner to the orchestration hub (*marked as (3) in Figure 1*) was originally done via direct email to each use case coordinator, attaching a compressed file –the file could be encrypted depending on the requirements of the participant partner. This organisational setting was considered a pragmatical approach to communicate the local outputs, originally supervised by the partner producing them and not requiring further security measures due to being aggregated following the disclosure policy (*i.e., k-anonymity level*) agreed by partners participating in each use case. However, this human-to-human interaction does not facilitate the execution of the outputs synthesis or the meta-analysis of the aggregated data (*marked as (4) in Figure 1*) as it entails managing multiple asynchronous communications between participants and the orchestrating hub. Although practical for the implementation and production of the results expected from the use cases within WP6, this setting was just conceived as a first step

in developing and implementing a fully distributed research infrastructure within the PHIRI federation both at the organisational and the technical levels.

The step-wise approach towards the PHIRI federated research infrastructure entails the development and implementation of human-to-machine communication of the aggregated local outputs as a second step (tier 1 - i.e., via automated reporting based on aggregated data update); a machine-to-machine unique direction communication (tier 2 – i.e., via an application programming interface (API) on the orchestration hub that could be leveraged by the PHIRI app after completing the local analysis to POST aggregated outputs in the orchestration server); and finally a machine-to-machine bidirectional secure communication (tier 3 – i.e., via establishing a secure communication channel between the orchestration hub and the participating data hubs enabling bidirectional messaging with script execution privileges within containerised environments) (see Figure below).



- Handling secure interactions (*communications*):
 - TIER 0 – HUMAN to HUMAN interactions
 - Example: *Each participant partner execute the analyses and sends an e-mail with a) the data quality report and b) the aggregate data attached to the use case leader for each use case in which they are participating*
 - TIER 1 – HUMAN to MACHINE interaction (with user interface)
 - Example: *Each participant partner execute the analyses, **logs into a website (i.e. health information portal) with user authentication and upload the data quality report and the aggregated data** to enable further meta-analysis or comparison by use case leaders*
 - TIER 2 – MACHINE to MACHINE one-way (automating retrieval of the outputs)
 - Example: *Each participant partner execute the analysis and press “Send outputs” to submit the data quality report and the aggregated outputs to a common repository enabling further meta-analysis or comparison by use case leaders*
 - TIER 3 – MACHINE to MACHINE two-way (distributing algorithms – federated learning)
 - Example: *Each participant partner configures an environment where the required data is available for analyses, and manages authorization for the deployment and execution of analytical algorithms on their data and the authorization for sharing their outputs*

Figure 4 Step-by-step approach to handling secure interactions (i.e., communication) within PHIRI federated research infrastructure

MATERIALIZATION OF THE PROTOTYPE

The PHIRI Federated Research Infrastructure (FRI) is supported by a containerised reproducible solution for data analysis to be deployed on-premises by each participant partner. This solution (a Docker image) has been published in ZENODO (<https://doi.org/10.5281/zenodo.5729310>). It constitutes a fully operative prototype that includes all the pipeline components required to carry out the research queries foreseen in WP6 and the demonstration pilot (Version 2.3.0). The building blocks of the PHIRI infrastructure are:

1) **5 common data models**: one for the demonstration pilot and four referred to the WP6 use cases.

The common data models result from the formalisation of the research queries included in these use cases in a semantically interoperable way for all the participant nodes. The research queries are the following:

- a) Use case A Vulnerable populations, inequalities and risk factors with direct or indirect impact on healthcare outcomes during the COVID-19 pandemic: To clarify whether healthcare utilisation patterns in vulnerable populations vary between settings and over time and is linked to the COVID-19 epidemiological situation, using individual-level health record, administrative and research data combined with ecological/group level contextual data
- b) Use case B COVID-19 related delayed care in breast cancer patients “Was there any delay in the treatment of breast cancer patients because of the COVID-19 stringency measures?”: Demonstrate whether there has been any increase in surgical and/or co-adjuvant (i.e. radiotherapy, chemotherapy, hormonotherapy and immunotherapy) treatment delay because of the COVID-19 crisis in eligible women diagnosed with breast cancer using individual-level health records, administrative and research data combined with ecological/group level contextual data.
- c) Use case C Effects of the COVID-19 pandemic on maternal and newborn health: Investigating the pandemic’s direct (infection by SARS-CoV-2) and indirect effects on perinatal health using routine population birth data and assess whether effects differ by socioeconomic context.
- d) Use case D COVID-19 related changes in population mental health: Measure changes in population mental health associated with the COVID-19 pandemic and, in particular, demonstrate whether there has been an increase in healthcare utilisation of mental health treatments as a consequence of the COVID-19 crisis in eligible patients diagnosed of depression and/or anxiety using individual-level health record, administrative and research data combined with ecological/group level contextual data.
- e) Demonstration pilot for rapid-cycle federated analysis. Piloting the federated analyses of the evolution of several indicators for the monitoring and surveillance of the COVID-19 pandemic (7-day reproductive number, 7- and 14-day incidence rates, regular and ICU admissions, bed occupancy rates) and 7- and 14-day predictions, using individual data from various data sources -COVID-19 monitoring and surveillance registries, hospital data, and administrative data.

The latest version of the common data models can be found here:

WP6 - Use Case A Common Data Model <https://doi.org/10.5281/zenodo.5148013>; (v3.0.0)

WP6 - Use Case B Common Data Model <https://doi.org/10.5281/zenodo.5148021>; (v2.0.0)

WP6 - Use Case C Common Data Model <https://doi.org/10.5281/zenodo.5148031>; (v2.0.1)

WP6 - Use Case D Common Data Model <https://doi.org/10.5281/zenodo.5148039>; (v2.0.0)

WP7 – Pilot Common Data Model <https://doi.org/10.5281/zenodo.5148466>; (v1.0.0)

2) **Synthetic datasets** for the variables included in the data model – these synthetic data set are used to fine-tune the data quality assessment scripts and the analytical algorithms before deployment on-premise

3) **Data quality assessment scripts** Once the data required in each use case, has been obtained and transformed to the common data model in each of the participating nodes, the data quality assessment script will help to understand the quality of the dataset that the researcher is going to use – in particular, completeness of each of the variables, anomalous distributions, the existence of numerous outliers, etc. Data quality assessment reports are customised for all use cases using pandas-profiling functionalities from the 'panda' library in Python v.3.10.

4) **Algorithms for data analysis** scripts for the PHIRI use cases A to D. Once the dataset has been accepted by the researchers (i.e., the data quality assessment provided recommends to follow), the analytical scripts are run to produce the actual research results. Updated versions of the analytical algorithms can be found here:

WP6 – Use Case A scripts (R Markdown) <https://doi.org/10.5281/zenodo.6359850>; (v1.1.2)

WP6 – Use Case B scripts (R Markdown) <https://doi.org/10.5281/zenodo.6359892>; (v4.0.1)

WP6 – Use Case C scripts (R Markdown) <https://doi.org/10.5281/zenodo.6380733>; (v2.0.1)

WP6 – Use Case D scripts (R Markdown) <https://doi.org/10.5281/zenodo.6359904>; (v1.1.1)

WP7 – Pilot scripts (R Markdown) <https://doi.org/10.5281/zenodo.7092521> (v1.0.0)

(Nota bene: Rapid-cycle analysis is enabled through continuous data updates by participant partners deploying this prototype)

5) **Mappings** from Use case A, B and D common data models to [OMOP v5.3 CDM](#) have been included as part of the update of the PHIRI app to mid-size prototype [check at <https://doi.org/10.5281/zenodo.5729310>]. However, they have not been integrated within the analytical pipeline as complete mapping at the variable level. The original idea of representing the use cases' information requirements into OMOP CDM was not possible due to semantic constraints in OMOP to represent certain variables required in the use cases within the scope of population health science.

Thus, a mapping exercise was completed between the CDM for use cases A, B and D to OMOP CDM, both at variable and value levels. Results from this mapping exercise are commented on within this Deliverable, and mappings from use case A, B and D common data models to OMOP CDM v5.3 variables and values are included in the update of the mid-size prototype of PHIRI Federated Research Infrastructure (Deliverable 7.2) in Zenodo. A complete mapping between the use cases data models and OMOP CDM was not possible at the variable level due to constraints regarding variable definitions (*i.e., semantics*) and data types (*i.e., syntaxis*) and the overall rationale of the use cases design (*i.e., population health research*) versus OMOP design guiding principles (*i.e., pharmacoepidemiological research*). However, mapping was possible at the value level mainly using OMOP-valid standard codes following a comprehensive approach. These incompatibilities between PHIRI and OMOP CDMs do not preclude the possibility of recreating –totally or partially- the cohorts defined to answer the research questions posed by PHIRI use cases using OHDSI-OMOP cohort definition based on OMOP value codes. It instead showcases the existence of two different approaches to semantic and syntactic interoperability to enable secondary use of healthcare, both aimed to facilitate reproducible pipelines and replicable outputs, with a) PHIRI approach, driven by the materialisation of a specific research question in a bespoke data model, enabling broader participation (*only depending on data availability and accessibility*), data minimisation (*enabled at both data source selection and entity, variable selection*), comparative analysis (*i.e., via synthesis of local outputs, or meta-analysis*), and enabling linkage with compatible area-level data (*i.e., using*

standard area level categories such as NUTS); or b) OHDSI approach, driven by a standard design, only accessible upon the extraction, semantic mapping and transformation of all data sources available to a standard data model (OMOP CDM), tying research question specification to the semantic constraints of the standard model specification.

6) **Demonstrator of the human-to-machine interface** enabling automatic update of the interactive use case B international comparison report. Technical specifications of the demonstrator are provided as part of this document (*Annex 2*). The containerised application supporting the use case B demonstrator is deployed in IACS servers managed by the Biocomputing Unit (Aragon, Spain). The demonstrator (Use Case B) can be accessed via links in the Health Information Portal within the menu tab for 'Federated Demonstrators' (<https://www.healthinformationportal.eu/services/federated-demonstrators>) or directly at a) <https://phiri.iacs.es/upload> for the 'Upload' interface -to upload new aggregated outputs from the local analysis produced by a partner participating in use case B, b) at <https://phiri.iacs.es/report> for the 'Interactive report' on international comparison analysis of participants in use case B, and c) at <https://phiri.iacs.es/> for an interactive mockup of the web application interface enable upon PHIRI app local deployment in a partners' systems.

7) **Implementation of an identity provider** - EGI AAI platform – to enable secure Authentication, Authorization and Identification access to participate in international comparison analysis (metaanalysis) within the PHIRI Federated Infrastructure. Secure access to the human-to-machine interface enabling automatic updates of the interactive international comparison reports requires the implementation of a robust AAI platform. PHIRI collaborates with the EGI-ACE project as an early adopter of the EGI services and solutions for the project; in particular, the EGI offers the possibility of creating EGI communities and registering applications using EGI-ACE services. One of those services is the EGI AAI platform, providing robust and secure identification services mediated by the OpenID standard, enabling registered users in EGI to securely identify and access authorised third-party applications or services via authentication. WP7 partners integrated and configured a Keycloak modular solution enabling the use of the EGI AAI platform to authenticate users of the PHIRI human-to-machine interface (*pending push to production*). The interface consists of a Nginx server providing the 'upload' and the 'report' interfaces within a Docker container, including a Lua Linux OS distribution with R and Python engines and scripts rendering the Quarto reports for the international comparison on aggregated data from use cases' participants. The Keycloak solution is an open-source solution, implemented in Java, implementing the federated identity protocols SAML2 and OpenID Connect (OIDC)/OAuth2 compatible with the EGI AAI platform. Keycloak implementation's main advantages enable a) a single session sign-in for multiple applications under the same platform, b) identity management of user accounts enabling user authorisation profiles, c) modular access management faceted by application functionality and user profile, d) high-performance and better user experience in the signing in and account management, e) authentication delegation integrating EGI AAI platform, f) Docker compatibility, and g) unified authentication methods facilitating maintenance and customisation of the integration with EGI AAI platform or other AAI platforms in the future.

The integration of secure AAI sign-in offered by EGI implies that PHIRI participants must be registered at EGI-ACE using their institutional accounts to be able to contribute with their aggregated outputs to the international comparative analyses (meta-analyses) resulting from their participation in any of the different use cases. Authorised and authenticated access to the PHIRI user-to-machine interface, for instance, secure the partners' participation in the use case by enabling user profiles to ensure only participants of a certain use case can contribute to the final report updating the latest data available; accessing the latest developments for the analyses and providing a secure environment to which upload the aggregated data outputs from their local analyses before being processed and compared to other countries. Creating the PHIRI community within EGI and integrating the EGI AAI platform to secure access is crucial to deploying a fully operable PHIRI Federated Infrastructure.

Secure access to the PHIRI environment (demo) improves communications and data security by adding an interoperable AAI layer without deteriorating user experience or reducing operational

status. In particular, the EGI AAI solution increases security, facilitates comprehensive management of users' identity, user profiles and account management, enables integration with other third-party applications or new modular development for PHIRI and improves scalability to different use cases and more advanced analyses and reporting solutions, while providing support to current industry identification standards. Finally, this implementation enables logging and auditing all user activity within the PHIRI infrastructure, improving user accountability while enhancing users' experience.

Finally, implementing secure access to the PHIRI environment (demo) is a keystone in developing machine-to-machine interfaces, enabling the implementation of federated learning algorithms running distributed algorithms across the PHIRI participants' network for use cases that may require it.

Annex 1 Technical specifications

The current technical stack of the PHIRI FRI includes (x2) Docker containers, using:

- (x1) Server-side (Back-end)
 - Framework: Nest (NestJS) -> TypeScript (<https://nestjs.com/>)
 - Execution environment: Node.js® >= v14.0.0 (<https://nodejs.org/es/>)
 - Database management system: SQLite (<https://www.sqlite.org/index.html>)
 - Data analysis environment: R version 4.0.4 (<https://www.r-project.org/>)
 - Data wrangling and management: Python (>= Python 3.8.12 (<https://www.python.org/>))
- (x1) Client-side (Front-end Web application)
 - Web server/proxy: Nginx (<https://www.nginx.com/>)
 - Framework: Vue.js -> HTML, CSS, and JavaScript/TypeScript (<https://vuejs.org/>)

Requirements for installing Docker

The PHIRI FRI includes deployment instructions as a PDF document ("deploy_phiri_app.pdf").

General requirements for installing Docker (support):

- Server: 1 server (or VM)
- CPU: 1 CPU minimum (or 2 CPU for VM configuration)
- GPU: No dedicated GPU required
- RAM Memory: 4 GB minimum (8 GB recommended)
- OS: Unix based (Linux or other) with Docker support

Specific PHIRI FDI Docker deployment requirements

- RAM Memory: 8 GB minimum (16 GB recommended)

Annex 2 Technical specifications of the demonstrator (Use Case B)

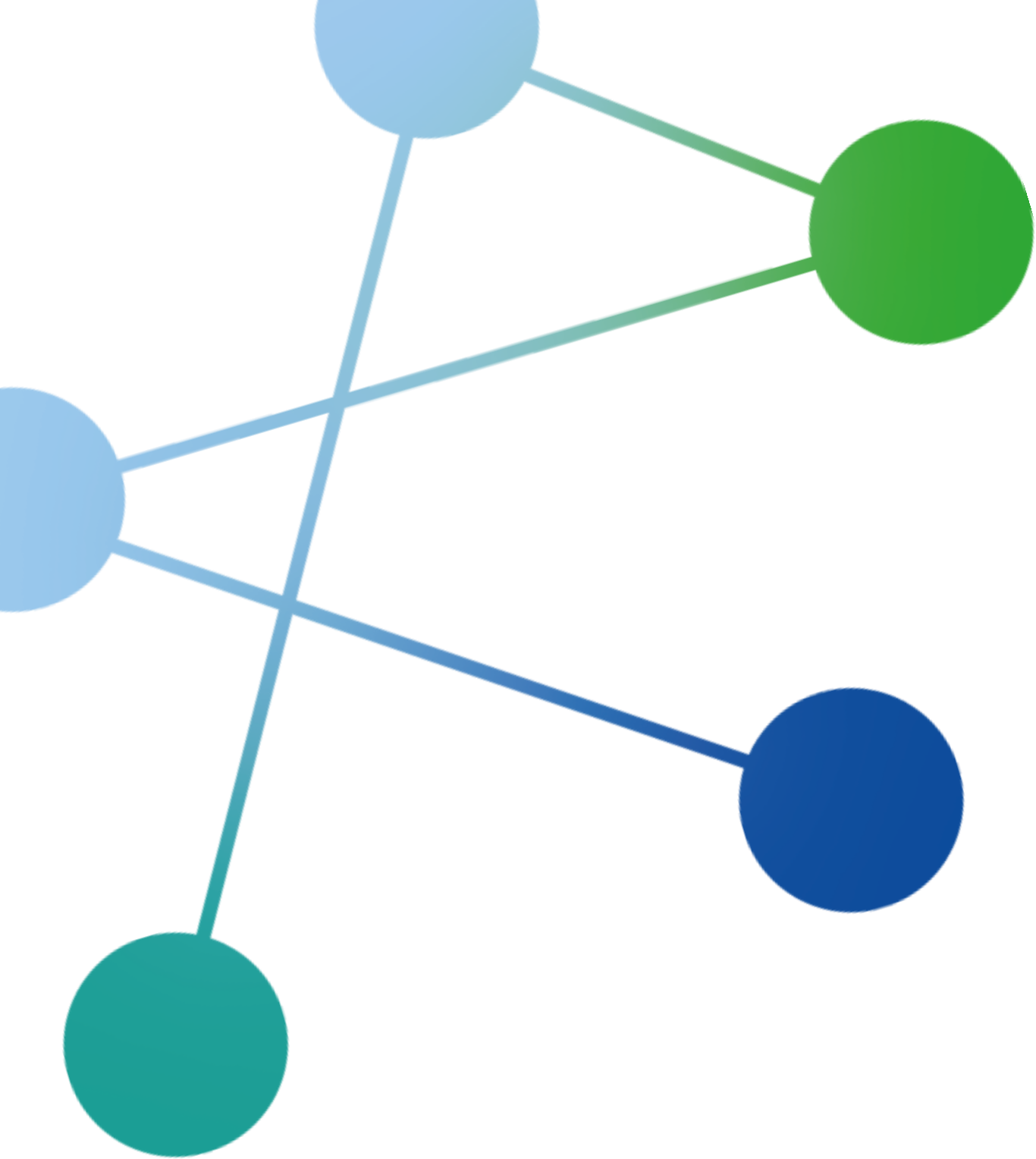
Current technical stack of the Demonstrator includes (x1) Docker container, using:

- (x1) Server-side (Back-end)
 - Framework: Wildfly v.16 (<https://www.wildfly.org/>)
 - Execution environment: Java, Java Runtime Environment (JRE), Java Persistence API (JPA) (latest versions - <https://www.oracle.com/java/>)
 - Database management system: PostgreSQL (<https://www.postgresql.org/>)
 - Data analysis environment: R version 4.1.0 (<https://www.r-project.org/>)
 - Interactive reports: Quarto version 4.2.0 (<https://quarto.org/>)
 - Data wrangling and management: Python (\geq Python 3.8.12 (<https://www.python.org/>))
 - Web server/proxy: Nginx (<https://www.nginx.com/>)
 - Keycloak (<https://www.keycloak.org/>) + OpenResty (<https://openresty.org/en/>)
 - EGI AAI platform: EGI Check-in proxy service (<https://www.egi.eu/service/check-in/>)

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