PHIRI
Population Health Information Research Infrastructure

Small-scale prototype of PHIRI federated infrastructure

Deliverable 7.1, 20.12.2021 (updated 13/05/22)

Javier González-Galindo
Francisco Estupiñán-Romero
David Chichell-Ruíz
Juan González-García
Carlos Tellería-Orriols
Ronan Lyons
Enrique Bernal-Delgado

IACS

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

PHIRI, the Population Health Information Research Infrastructure, 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 (Ethical, Legal and Social Issues) and FAIR (Findability, Accessibility, Interoperability, and Reuse) principles. It has the objective: 1) to develop and nourish the Health Information Portal for COVID-19 with FAIR catalogues on health and health care 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, the Work Package (WP) 7 develops the technological substrate for the development of a federated research infrastructure that allows mobilizing sensitive data to respond multiple research queries in multiple sites, while preserving GDPR principles.

The task 7.1 demonstrates the suitability of this federated research approach in the production of accurate and timely research outputs for a rapid policy response on COVID-19; in particular, it establishes the governance mechanism and implements the technological solutions to respond to four uses cases carried out in multiple sites (countries):

1. Impact of the 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)

In this Deliverable 7.1 we describe the building blocks of the development of the small-medium scale PHIRI federated research infrastructure; thus, the development of the common data model por each of the use cases, the implementation of the scripts for data quality assesment and the script for analyses, all of them contained within technological solution, Docker Image, that is a secure environment for the implementation of the Federated Research Infrastructure (FRI) across nodes. This version of Deliverable 7.1 (13/05/22) already contains features of Mid-size prototype of PHIRI Federated Infrastructure (Deliverable 7.2) that is being developed in parallel.

PHIRI Federated Research Infrastruture (FRI)

Created by Javier González-Galindo, Francisco Estupiñán-Romero, David Chichell-Ruíz, Natalia Martinez-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 Mª Carmen Rodriguez 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)** and a **central orchestrating hub at IACS**. The exchange of information follows the steps described below (Figure 1).

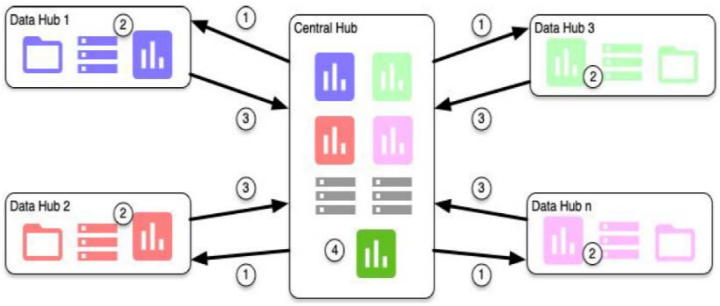


Figure 1: Visual representation of the exchange of information across the PHIRI hubs

1) The orchestrating hub develops, implements and shares the analytical pipeline and provides support to the federated research infrastructure for its deployment. 2) Nodes acting as data hubs deploy the pipeline on premise; 3) so, no sensitive data is shared across the federation of nodes but digital objects (e.g, aggregated outputs or models) 4) The orchestrating node can perform meta-analyses with those outputs if required.

PROTOTYPING THE ORCHESTRATION OF A USE CASE

The orchestrating hub has developed a prototype, a stepwise approach, aiming full interoperability at any stage of the process (figure 2); thus, starting with the formalization of the research query as a common data model for all the nodes, following with the deployment

Figure 2: stepwise approach from the formalization of a research query to its final output

of the analytical pipeline on-premise to run the analyses and, finalizing with the collection of the research results and their publication

MATERIALIZATION OF THE PROTOTYPE

The PHIRI Federated Research Infrastructure (FRI) is supported by a containerized reproducible solution for data analysis to be deployed on-premise by each participant partner. This solution (a Docker image) has been published in ZENODO (https://doi.org/10.5281/zenodo.5729310**)** and constitutes a small-medium scale prototype that includes all the pipeline components required to carry out the research queries foreseen in WP6 and the demonstration pilot (Version 2.2.1). The building blocks of the PHIRI infrastructure are:

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

The common data models are the consequence of the formalization of the research queries included in these use cases in a way that is semantically interoperable for all the participating nodes. The research queries are:

a) Use case A - Vulnerable populations, inequalities and risk factors with direct or indirect impact on health care outcomes during the COVID-19 pandemic: To clarify whether health care utilization 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 has been any increase in surgical and/or co-adjuvant (i.e. radiotherapy, chemotherapy, hormonotherapy and immunotherapy) treatments delay because of the COVID19 crisis in eligible women diagnosed of breast cancer using individual level health record, 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 has been any increase in healthcare utilisation of mental health treatments as a consequence of the COVID19 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 a number of 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 a number of data sources -COVID-19 monitoring and surveillance registries, hospital data, and administrative data.

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; (v2.0.0)

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. For all use cases data quality assessment reports are customized 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 (ie, the data quality assessment provided recommends to follow), the analytical scripts are run to produce the actual research results. Analytcal algorithms can be found here:

WP6 – Use Case A scripts (R Markdown) https://doi.org/10.5281/zenodo.6359850; (v1.1.0)

WP6 – Use Case B scripts (R Markdown) https://doi.org/10.5281/zenodo.6359893; (v4.0.0)

WP6 – Use Case C scripts (R Markdown) https://doi.org/10.5281/zenodo.6380733; (v2.0.0)

WP6 – Use Case D scripts (R Markdown) https://doi.org/10.5281/zenodo.6359904; (v1.1.0)

WP7 – Pilot scripts (R Markdown)

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

Annex 1 Technical specifications

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

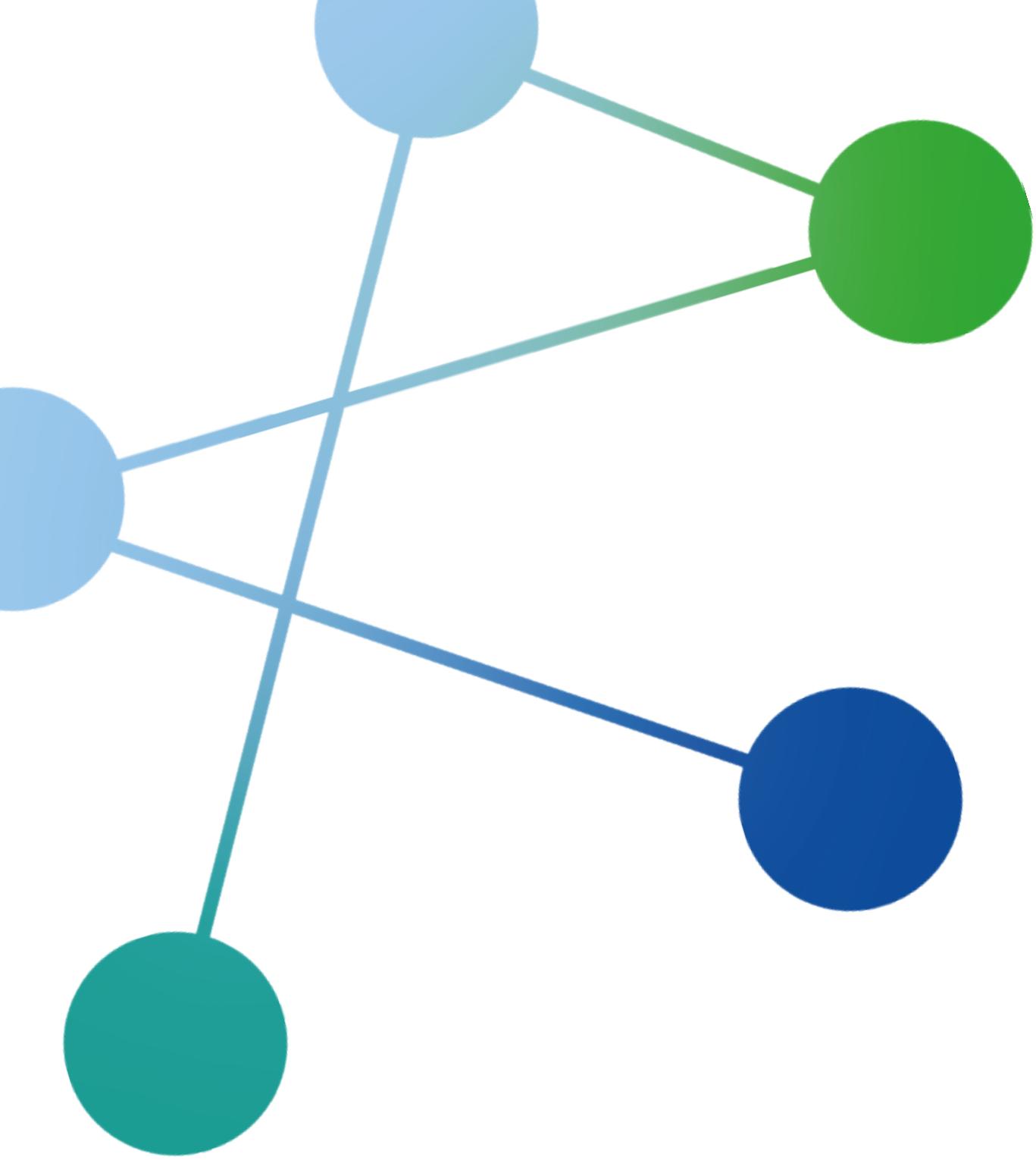
Specific PHIRI FDI Docker deployment requirements

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

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