ORCHESTRA

ORCHESTRA will develop a data repository constituting a large population cohort based on cohorts from multiple countries. The result will support studies to improve public health and vaccine strategies for tackling COVID-19.

In a newly funded project called ORCHESTRA ("Connecting European Cohorts to Increase Common and Effective Response to SARS-CoV-2 Pandemic") the High-Performance Computing Center Stuttgart (HLRS), together with 27 centers for public health and high-performance computing from 15 countries in Europe, Africa, South America, and Asia, will help develop a data infrastructure for collecting and analyzing COVID-19 patient data from across Europe and other parts of the world.

"Although HLRS itself doesn't conduct biological or epidemiological research, the kinds of computing resources and expertise that we provide are going to be very useful for ending the COVID pandemic," said HLRS Director Prof. Michael Resch. "We are delighted to be working with such a diverse, multidisciplinary team to support scientists and clinicians in this important and potentially high-impact effort."

The three-year project is being led by Prof. Evelina Tacconelli at the University of Verona, Italy. The project budget of nearly €20 million is funded by the European Union's Horizon 2020 research and innovation program under the ERAvsCORONA ACTION PLAN. HLRS will focus on building the computing infrastructure and data management framework for collecting, storing, integrating, and sharing critical data related to the pandemic.

"For a data engineer, this work will be exciting both as a technical challenge and because it offers an opportunity to contribute useful expertise in the fight against COVID-19." says Dr. Björn Schembera, a researcher in data management who will lead HLRS's participation in ORCHESTRA within WP7 (data management). "Bringing together cohort analysis and advanced data technology will give epidemiologists new insights into the pandemic"

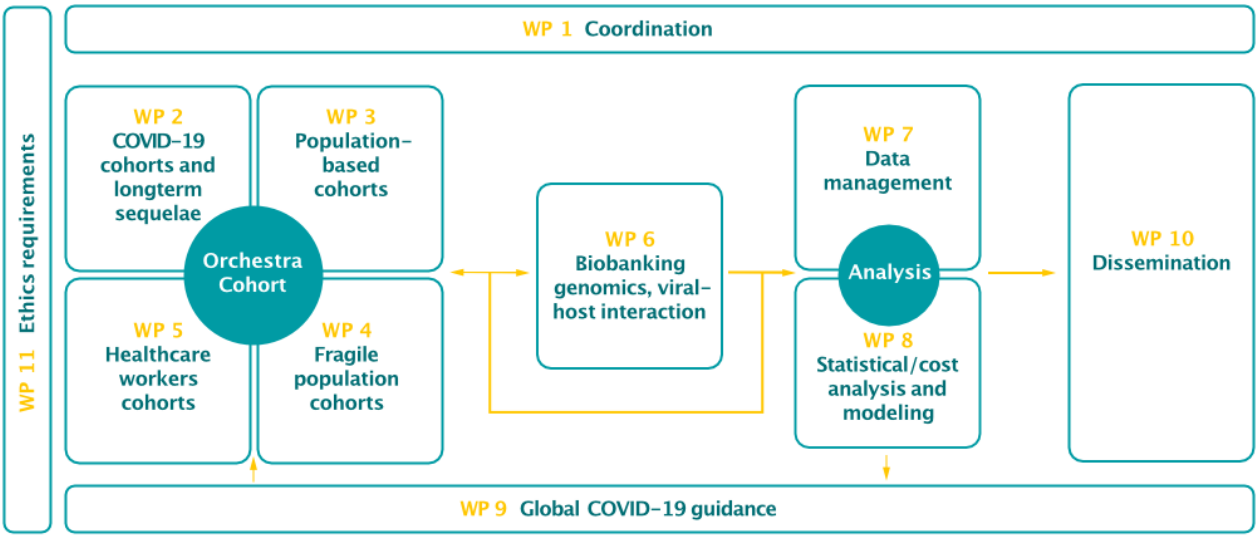

Figure 1ORCHESTRA project structure and cohorts (https://orchestra-cohort.eu/work-packages/)

A resource for COVID-19 research

The large patient cohort of ORCHESTRA will be built from various cohorts (COVID-19 positive and long-term sequelae (WP2), population-based (WP3), fragile population (WP4), healthcare workers (WP5)) and biobanking genomics (WP6), as depicted in figure 1. The resources will enable retrospective analyses of risk factors for disease acquisition and progression, as well as prospective follow-up aimed at exploring long-term consequences of the virus. Such knowledge will be valuable for preparing for and managing potential future waves of COVID-19 spread, or other kinds of future pandemics.

Data infrastructure needed to improve cohort analysis

Working together with scientists at the high-performance computing centers CINECA (Italy) and CINES (France), HLRS will contribute to the development of the data infrastructure needed to support the ORCHESTRA project objectives, as depicted in figure 2.

Each of the IT partners involved in the project will establish a national hub that will be responsible for collecting data from national data providers as well as storing harmonized de-identified data while assuring compliance with the data protection regulations. A cloud-based ORCHESTRA portal will then offer an online interface for sharing, accessing and linking together data managed by the national hubs. The portal will also include machine learning and data analytics tools and methods, which are tailor made to answer research questions posed by the ORCHESTRA project. HLRS will help to build a federated research architecture for cohort analysis based on three layers: national data providers, national hubs, and the centralized ORCHESTRA portal. The center will oversee the implementation of national hubs among the project partners and act as the national hub for cohort data in Germany. A national hub can be envisioned as a system for cohort data and metadata storage complemented with tools to enable federated learning on these data. Federated learning is an approach to jointly analyze data when it can't be moved to other locations due to legal regulations or security restrictions.

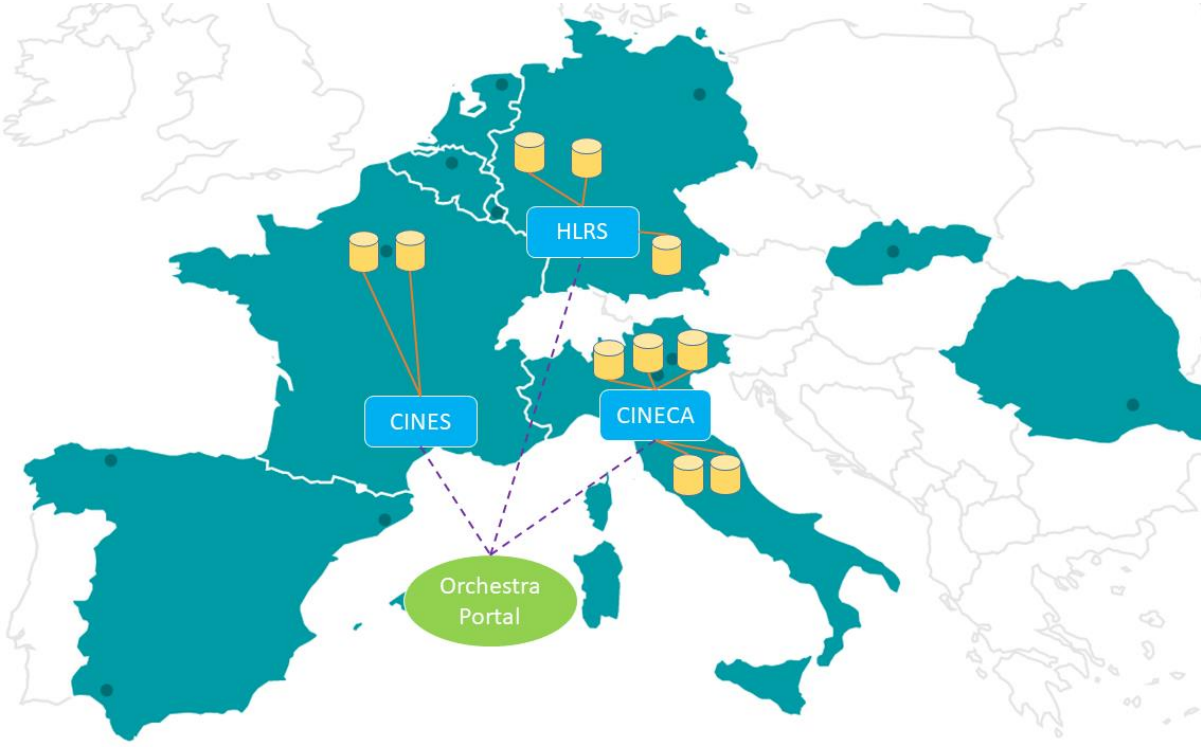

Figure 2: ORCHESTRA data infrastructure with CINES, CINECA and HLRS national hubs

ORCHESTRA in summary

- Pan-european European project to build and analyze a large patient cohort for the conduct of prospective and retrospective studies in order to improve the prevention and treatment of COVID-19 and to be better prepared for future pandemics.
- Data infrastructures and analytics assuring the compliance with the data protection legislation are crucial to achieve the project goal.
- HLRS is leading the design and implementation of the national data infrastructure (national hubs).
- A national hub offers data storage and management for cohort data on the national level (e.g. all German cohorts).
- Each national hub offers a part of the federated learning/analysis components in order to jointly analyze data across borders without violating national legal regulations.

Contact

Dr. Björn Schembera / schembera@hlrs.de / https://orchestra-cohort.eu/

Disclaimer: This text is based on the press release from the ORCHESTRA project and has been written by Chris Williams, Björn Schembera and Miroslav Puskaric.