

Data Harmonization workshop

Venue of workshop: Virtual

Dates for the workshop: 18th – 19th May 2022

Workshop organisers: CINECA, H3ABioNet

Application opens: 5th April 2022

Application closes: 22nd April 2022

Notification date: 3rd May 2022

Link to application form: <https://bit.ly/cineca-data-harmonization-registration>

Workshop Sponsors: CINECA and H3ABioNet

Contact: [Nicola Mulder](#), [Mamana Mbiyavanga](#)

Workshop webpage: <https://www.cineca-project.eu/workshop/data-harmonization>

Course Overview: The ultimate goal of CINECA's vision of a federated cloud-enabled infrastructure making population-scale genomic and biomolecular data accessible across international borders is to enable large-scale federated data analysis responsibly and securely. This will require integrating and harmonizing diverse, large human cohort data using community standards. Data harmonization within and across cohorts adds value to the data for downstream analysis and interpretation and facilitates cross-cohort meta-analysis.

This workshop aims to discuss ways to address common challenges in cohort data harmonization, work towards practical steps to address them, and share best practices. We welcome any cohort with plans for prospective or retrospective data harmonization, enthusiastic about sharing their experience and learning from others' perspectives in cohort data discovery and analysis.

Topics to be covered:

- Data cleaning and curation
- ELSI considerations in merging data
- Data collection standards, ontology terminology and interoperability standards, metadata models
- Data storage standards
- Data harmonization
- Sharing cohort summary data

Applicants are encouraged to check out the CINECA webinar on 31st March 2022 (<https://www.cineca-project.eu/webinar/bringing-it-all-together>), which highlights some of the relevant standards and applications.

Objectives: After this workshop, participants should be able to:

- Do basic data cleaning
- Understand what data standards & ontologies exist for clinical data
- Map their cohort metadata to a data model
- Understand existing approaches to and algorithms for data harmonization
- Prepare summary data from their cohorts

Intended Audience: Members of cohort projects who are working on data curation and management. Data managers, curators, bioinformaticians, data scientists.

Prerequisites: None, but should be involved in cohort data management or analysis.

Workshop limitations: This workshop will only provide a foundation for continued learning in data harmonization, with some example applications using synthetic datasets. Future bring our own workshops can be arranged for more hands-on work with your cohort data.

Workshop Programme:

Time CEST/CAT	Topic	Speaker
18th May 2022		
14:00	Welcome and introduction, workshop aims	Nicky Mulder
14:15	Data cleaning	Katherine Johnston, Ayton Meintjes
14:45	Machine learning/text mining tools for cleaning data	Isuru Liyanage
15:15	ELSI considerations in merging data	Melanie Goisaufer
15:40-16:00	Break	
16:00	Overview of data collection standards, ontology terminology and interoperability	TBD
16:45	Metadata: GECKO, IHCC	Carles Garcia
End 18:00	Hands-on work to prepare for loading into Atlas	Carles Garcia, Mamana Mbiyavanga
19th May 2022		
14:00	Browse newly uploaded data in Atlas	Carles Garcia
14:20	Other considerations -data storage standards	Alexa Heekes
14:45	Summary from the literature review on data harmonization	Lyndon Zass
15:00	Example 1: DPUK	Sarah Bauermeister
15:20	Example 2: H3Africa CVD	Katherine Johnston
15:45-16:00	Break	
16:00	Example 3: PRIMED	Leslie Lange
16:20	Data harmonization algorithms	Mamana Mbiyavanga
16:45	Cohort summary data -how to generate and share - Phenopackets, MIACC	TBD
17:15-18:00	BYOD and discussion	All