www.egi.eu
@EGI_eInfra

EGI-ACE Community Workshop

Population Health Information Research Infrastructure (PHIRI)

Juan González-Garcia / IACS-ES
Miriam Saso / Sciensano-BE
16th-17th February 2021

The work of the EGI Foundation
is partly funded by the European Commission under H2020 Framework Programme

Agenda:https://indico.egi.eu/event/5360/

Outline

- Background about the scientific community
- Ambition and challenges
- High-level architecture
- Technical requirements
- Capacity requirements
- Integration support
- Timeline
- Training for external users

Background about the scientific community

- Population Health Research
  - Public Health / Epidemiology
- Network from Joint Action InfAct
  - Representatives of Public Health Institutes from 41 partners of 30 countries
    - 27 National Institutes of Public Health / Research / Disease Control
    - 7 Universities
    - 7 Ministries of Health

INFACT
Joint Action on Health Information

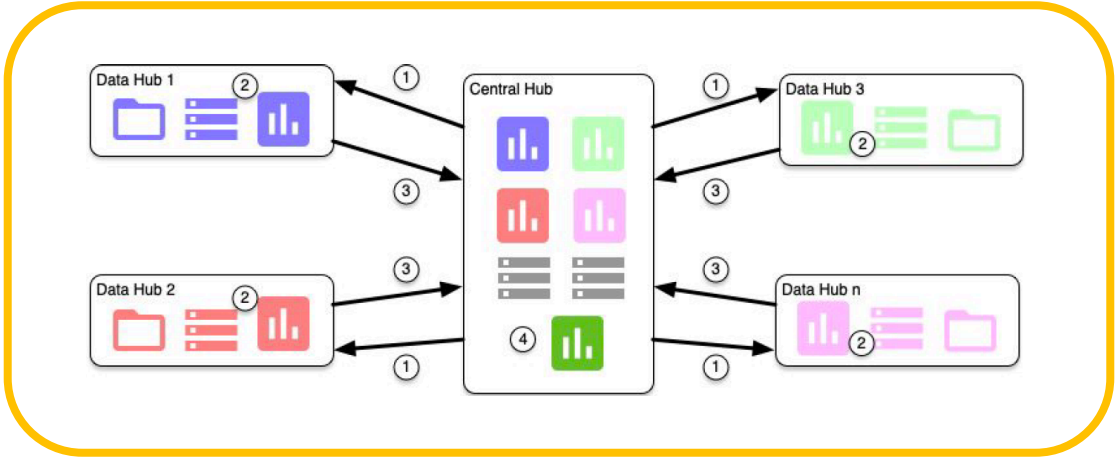
PHIRI
Population Health Information Research Infrastructure

Ambition and challenge(s)

- Build and validate a federated research infrastructure on rapid cycle analysis
  - Demonstrated through COVID19 uses cases (4+1)
  - Valid for future pandemics and (in general) observational studies
  - Stablish a solid governance structure
  - Serve as prototype of Distributed Infrastructure on Population Health (DIPoH)
  - Align with European Health Data Space (EHDS) & others (HealthyCloud, etc.)
- Setup a network of IT developers capable of sustaining and upgrading the FRI
- Setup Health Information Portal on population health
  - Metadata catalogues on population health data sources, studies, guidelines, projects and trainings

Ambition and challenge(s)

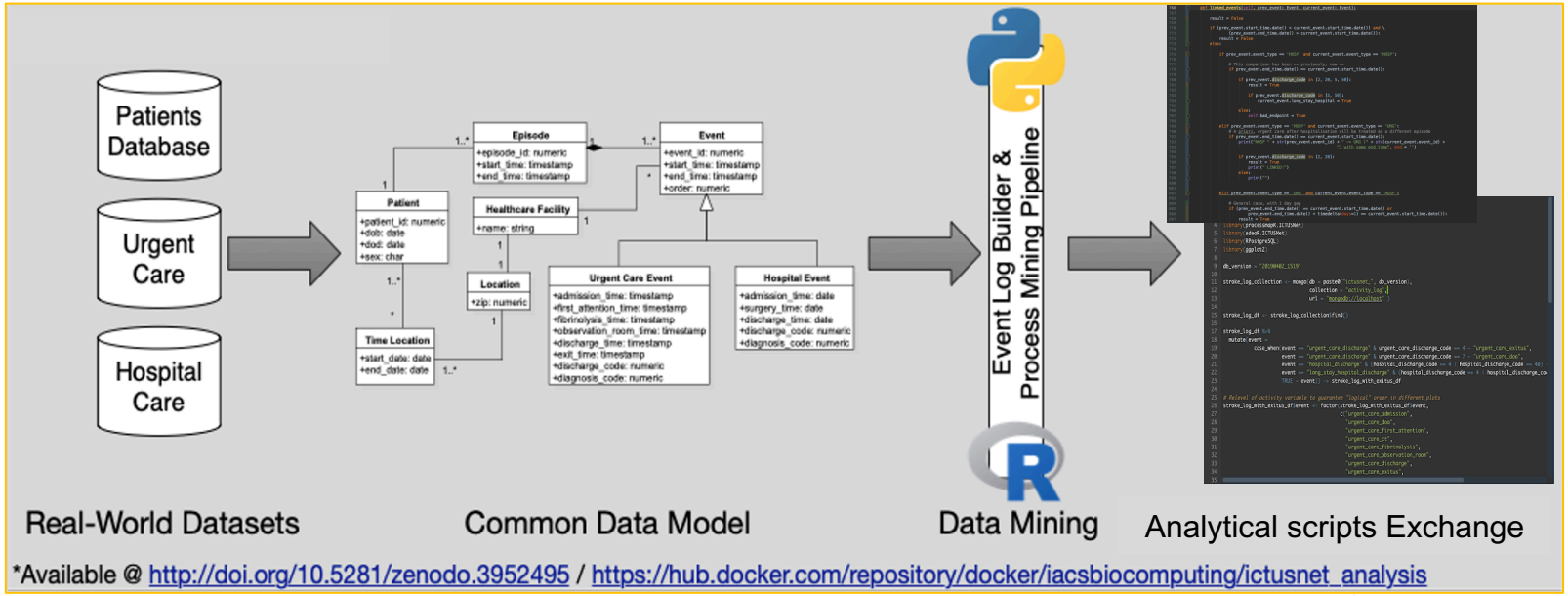
High-level architecture of the PHIRI



High-level architecture of the PHIRI

Routine linkable data / Individual level (anonymised or pseudonymised)

Research Questions



*Available @ http://doi.org/10.5281/zenodo.3952495 / https://hub.docker.com/repository/docker/iacsbiocomputing/ictusnet_analysis

Technical requirements

- Online Storage
- Data transfer
- Cloud compute
- Galaxy frontend?

- **NOTE**: actual requirements should be aligned with GDPR and national regulations

Capacity requirements

- Current dimensions (approx.)
  - GBs to 10s GB data sets per partner (structured data)
  - Local analytical computations of basic ML techniques (regressions, process mining)
  - Low networking usage
- Scale up
  - 10s to 100s GB data sets per partner (adding imaging, others)
  - Pure distributed algorithms (distributed regressions, federated learning)
  - Mid (to high?) network usage

Integration support

- ETL from local datasets to CDMs
- Data motion to computing nodes when required
- Federated learning orchestration

- **NOTE**: actual requirements should be aligned with GDPR and national regulations

Timeline

- Pilot use case to be delivered by Nov 21
  - Status: Surveying data availability
- General use cases to be delivered by Apr 22
  - Status: Defining data models
- Research infrastructure solution to be delivered by Oct 22
  - Production-grade solution (AuthN/AuthZ, stable deployment, stable interfaces)
- Further upgrades by Oct 23
  - RI tuning
  - Establish Common Data Models
  - DIPOH aligned with EHDS

Training for external users

- Capacity building and Developers working group
  - Supporting the specific IT implementation activities
  - Coordinating the aforementioned activities and use cases guidance
- Meetings timeline TBD

Thank you!