CINECA: Developing a cloud-based federated infrastructure for international human data sharing and analysis

Authors: **Leslie Glass** (EMBL-EBI), Mamana Mbiyavanga (UCT), Thomas Keane (EMBL-EBI), Lauren Fromont (CRG), David Bujold (McGill Univ), Nicky Mulder (UCT), Jonathan Dursi (UHN), Elisa Cirillo (The HYVE), Melanie Courtot (EMBL-EBI), Mikael Linden (CSC), Patrick Ruch (HES-SO), CINECA consortium members

The aim of the CINECA project is to deliver a federated infrastructure for data discovery of human genetic and phenotypic data, facilitating transcontinental human data exchange for research and clinical applications. ***This presents 5 key challenges:***

- ***Challenge 1:*** Federated data discovery - standardised methods and portals for federated search and discovery of relevant human data.
- ***Challenge 2:*** Interoperable authentication and authorisation infrastructure - incorporating standardised researcher IDs which include trusted researcher credentials.
- ***Challenge 3:*** Harmonised cohort level metadata - a common metadata model, alignment with community standards & semantic interoperability is essential to perform analyses.
- ***Challenge 4:*** Federated analysis interoperability for research and healthcare applications - analysis is federated and migrated to the data, using standardised interfaces & tools.
- ***Challenge 5:*** Trans-national harmonised ELSI framework - enabling sharing within an effective ethical, legal and social framework which adheres to national and European regulations, and respects the rights of the participants.

GA4GH standards used in CINECA

- **Clinical & Phenotypic Data Capture Work Stream:** Phenopackets
- **Data Use & Researcher Identities Work Stream:** Data Use Ontology (DUO), Passports
  Data Security Work Stream: Authentication and Authorization Infrastructure (AAI)
- **Cloud Work Stream:** Task Execution Service (TES), Data Repository Service (DRS), Tool Registry Service (TRS), Workflow Execution Service (WES)
- **Large Scale Genomics Work Stream:** Genetic Variation Formats (VCF), Read Data Formats (SAM/BAM/CRAM), htsget API, Reference Sequences (Refget), Genetic Data Encryption (Crypt4GH)
- **Discovery Work Stream**: Beacon API v2 - in development, Data Connect API, Service Registry
- **Regulatory & Ethics**: Consent Policy v2, Framework for Responsible Sharing of Genomic and Health-Related Data

WP1

Beacon API v2 · Service Registry · Data Connect API

- **WP1** is deploying emerging GA4GH product Beacon version 2 (see **Figure WP1-1**), and extended it with:
  - a service registry https://service-registry-demo.ega-archive.org to gather all Beacons
  - extended query services (D1.2. video and video), using Data Connect (https://github.com/ga4gh-discovery/data-connect)
  - a model for cohorts (link to blog) into Beacon logical schema (**see Figure WP1-2**)
- Ultimate goal for WP1 is to gather these discovery services under one single portal to be used by researchers from all over to query about genomic and clinical data

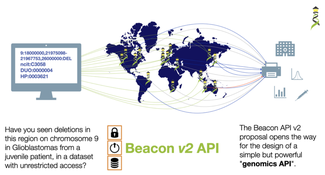


Figure WP1-1. Beacon v2 will allow researchers from all over the world to query databases for specific genomic and clinical information, while ensuring data privacy and security.

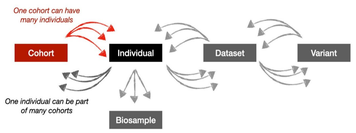


Figure WP1-2. Introduction of cohorts into Beacon v2 logical schema

Cohorts

refget API · VCF CRAM · htsget API

- CINECA has produced a set of cohort-specific **synthetic datasets** based on the phenotypic data from four participating cohorts - UK Biobank, CoLaus, H3Africa & CHILD Cohort Study
  - To increase accessibility to cohort data for standards development
  - Mitigating ethical and legal privacy concerns that arise with cohort data sharing
  - Open access and fully accessible under Creative Commons Licences.
- Descriptions and links can be found on our web page here.

WP7

PIPEDA · POPI Act · GDPR

Consent Policy v2 · Core elements of responsible data sharing

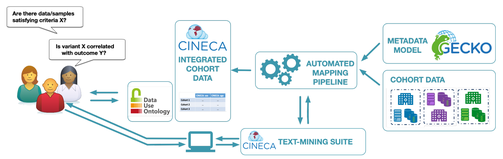
- **WP7** has produced a Catalogue of Canadian, European and African ethical and legal gaps which:
  - focuses on how the CINECA project can be efficiently conducted - especially with respect to data sharing
  - while being legally compliant with relevant laws, regulations, and with established ethical guidelines and practices across three continents
- CINECA ELSI team have delivered multiple webinars and training events, including a Webinar on the Ethical, legal and societal issues in international data sharing.

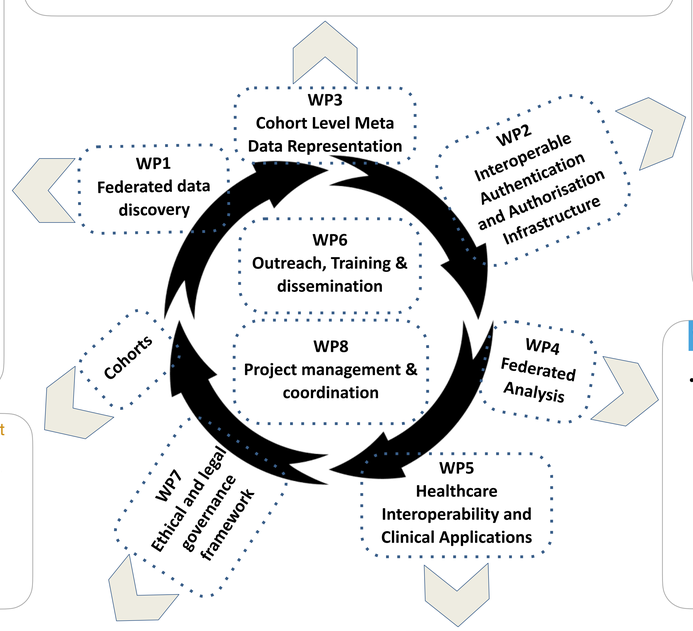
WP3

DUO · Phenopackets · GECKO

WP3 supports:

- cohort data integration and harmonisation for powerful downstream cross-cohort analyses
- Develops a semantic model representing shared attributes across cohorts
- text-mining pipelines for metadata enrichment and standardisation
- contributes to GA4GH Data Use Ontology development to provide machine-readable data use conditions
- WP3 models reused by other international consortia such as the IHCC

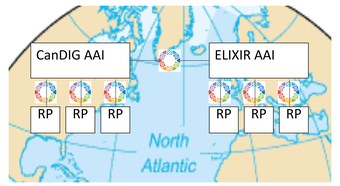




WP2

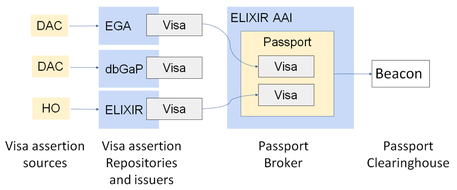
GA4GH AAI · GA4GH Passports

- **WP2** is building on the GA4GH Passports standard to develop a service that delivers researcher's trusted credentials, such as roles and permissions, to access sensitive datasets.
  - Implemented GA4GH Passport support to ELIXIR AAI (blog here, detailed demo deliverable here). An ELIXIR user can log in using their home organisation, enabling them to attach their affiliation to their passport.
  - Also delivered REMS (Resource Entitlement Management System), an electronic tool that DACs can use to review the DARs.



ELIXIR AAI & Canada AAI Integration

- Upgraded ELIXIR AAI and Canada AAI to support GA4GH protocols for CINECA interoperability with other global cohort infrastructures.



GA4GH Passport assembly in ELIXIR AAI

- To assemble a GA4GH Passport, ELIXIR AAI pulls information from several sources:
  - Visa assertion repositories (e.g. EGA, REMS) for Controlled Access Grants visas
  - Its internal sources for the visas for registered access
  - Researcher's home organisation (HO) for roles (federated identity management)
- The issuer's signature (JWT) in the visa is retained in the passport.

WP4

TES · DRS · TRS · WES

- **WP4** aims to implement a technical framework to run different types of federated analysis.
  - To enable the analysis of input information that is split across several controlled access human data cohorts without exporting the raw data
  - Bringing the analysis to the data in a secure cloud-based infrastructure.
  - See video demonstrator of a *common framework for designing portable federated pipelines* here and GitHub repo here.

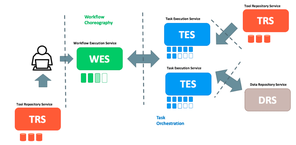
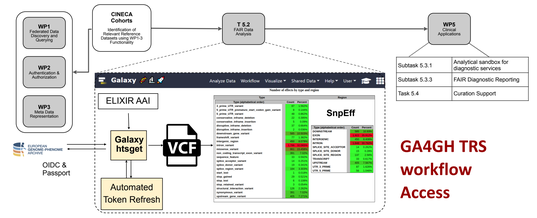


Figure WP4 - GA4GH TES implementation scheme

WP5

GA4GH AAI · GA4GH Passports · htsget API · TES

- **WP5** is building Curation-support Services for Somatic (or Clinical) Variants that will use the services from all of the technical work packages (WP1-4).
- One aspect in development are the Query Expansion Services: Ontology and Data-driven expansion and Variant expansion.



GA4GH TRS workflow Access

CINECA

Common Infrastructure for National Cohorts in Europe, Canada, and Africa

Get in touch:
https://www.cineca-project.eu
info@cineca-project.eu

Consortium Institutions

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825775.
This project has received funding from the Canadian Institute of Health Research under CIHR grant number # 404896