European Genomic Data Infrastructure

GDI - Bridging genomic research across Europe

TNC 2024

Slávek Licehammer, Dominik František Bučík
(orig: Melissa Konopko, Dylan Spalding, Tommi Nyronen)

Funded by the European Union

GDI website

@GDI_EUproject

/company/gdi-euproject

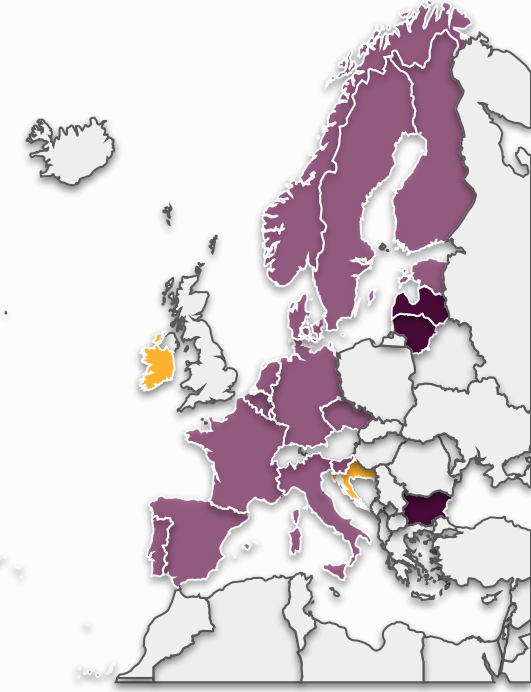
Agenda

- Genomic Data Infrastructure Introduction
- Main GDI user story
- Started kit
- Data Access Management
- GA4GH Passports and Visas
- Standardization

Funded by
the European Union

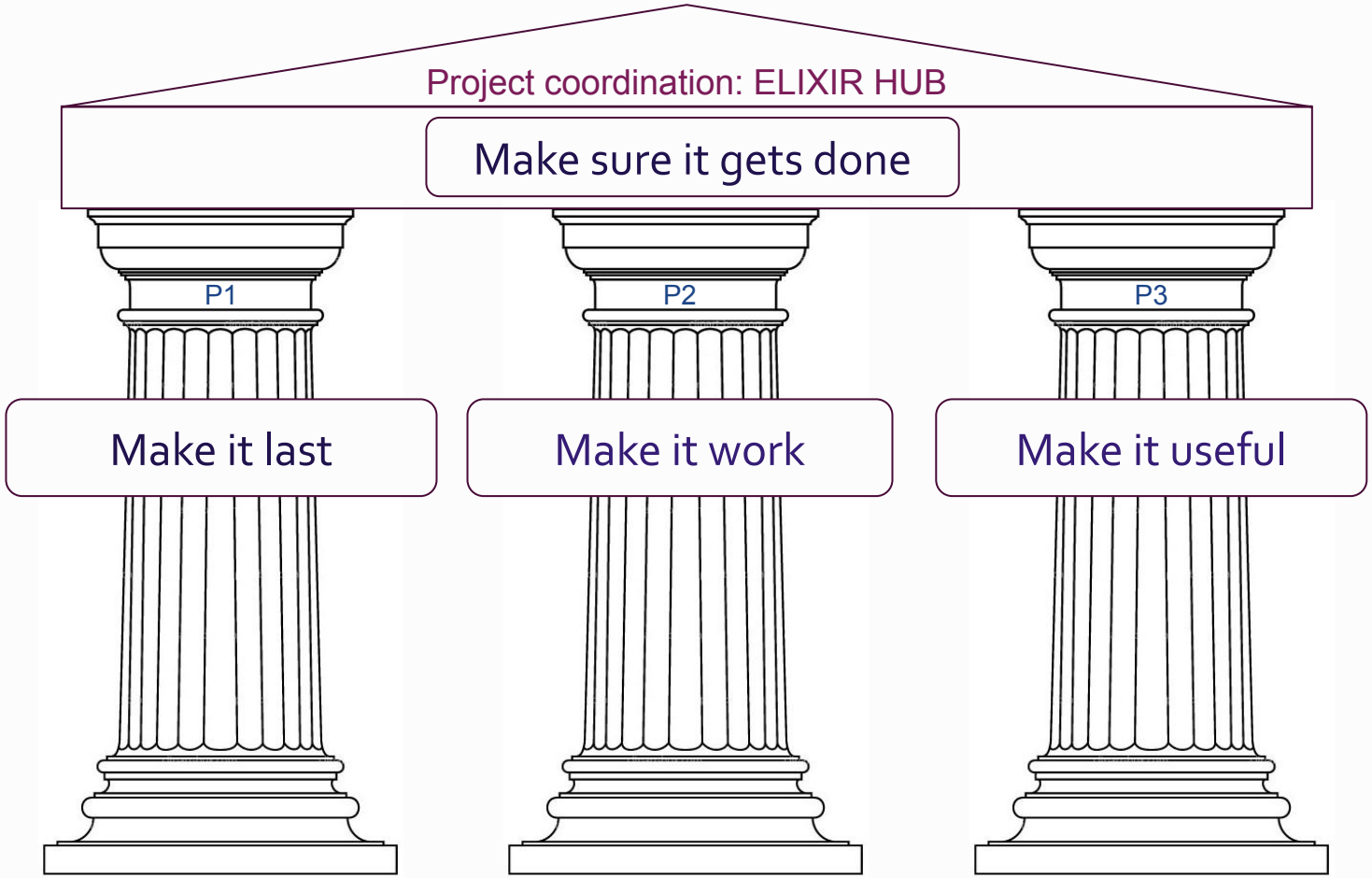
Genomic Data Infrastructure

- Support the EU 1+Million Genomes (1+MG) initiative (Digital Europe policy) ambition
- Establishing a federated, sustainable and secure infrastructure based on open community standards
- 20 1+MG signatory countries in GDI
  - Will provide a node or Data Hub
  - Each country manages their own data (e.g. regional hubs)
  - Data hubs provide cross-border data analysis
- Expected that genomic and phenotypic data clinically derived and from Genome of Europe
- Overall data infrastructure provides 5 main functionalities



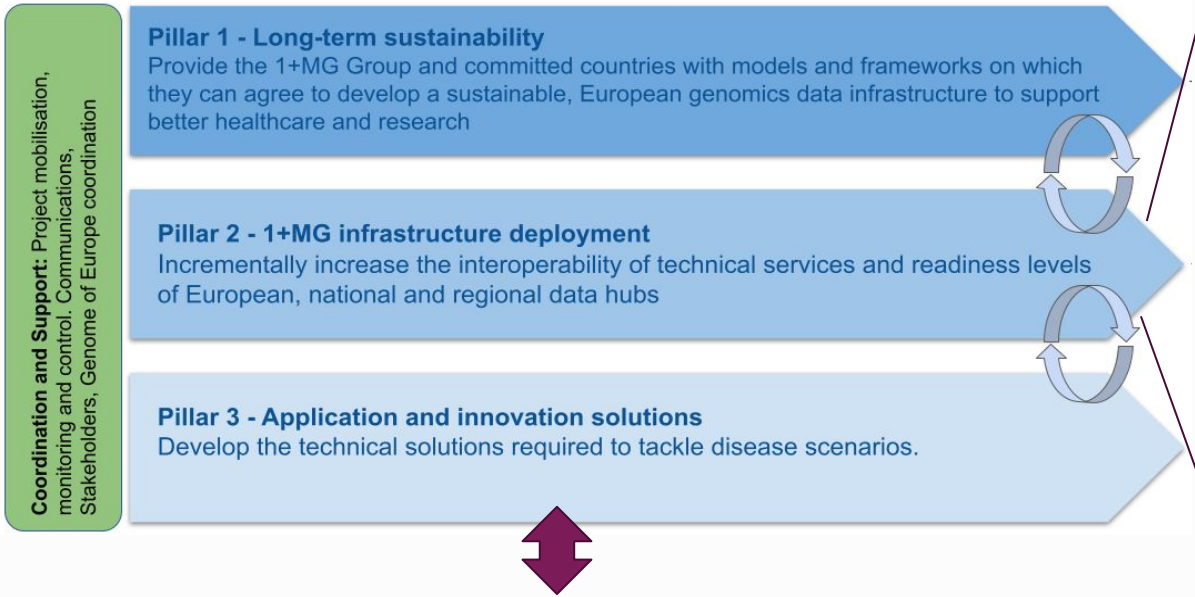
Data discovery | Access management tools | Data processing | Data reception | Storage and interfaces

Funded by the European Union



Funded by
the European Union

Genomic Data Infrastructure



Use Cases: Cancer, Infectious disease, Rare Disease, common complex disease, Genome of Europe

WP3 1+MG Infrastructure
- Onboarding
- Deployment
- Operational

WP4 European level operations
- User portal
- Helpdesk

WP5 Technical Outreach
- Dissemination
- Capacity building

WP6 Data Management
- Data management support
- Realistic synthetic data

Funded by the European Union

Generalised end to end user Story from 1+MG

1. User discovers phenotypes of interest aggregate data in 1+MG data
2. User logs in via LS AAI (registered level), and discovers both a genomic variant and treatment regime and/or phenotype of interest via Beacon[1]
3. User applies for data access to 1+MG data
4. Data Access Committee grants access to a virtual cohort
5. User executes analysis as a controlled access user on this virtual cohort across federated locations





[1] **WG8 – RD:** Do you have any individuals with a mutation in the RYR1 gene and a similar phenotype to congenital myasthenic syndrome?

WG9 – Cancer: Do you have any individuals with a mutation in the PTEN gene, who have BRAF biomarkers and are being treated with vemurafenib?

Funded by the European Union

Starter Kit

- Starter Kit is effectively a proof-of-concept – **not** a production system
  - **Evolved version of the B1MG PoC**
- Demonstrate how a set of applications and components can be linked via standards
  - **Each application or component a product with an assigned product owner**
- Primarily Node level
  - **Deployed at each node**
  - **European level operations in WP4**
- Support knowledge transfer and capacity building
- Outreach to Pillars I and III as well as use cases
- Deployed in waves across nodes
- Synthetic data included (e.g. B1MG rare disease dataset, CINECA UK1, B1MG cancer dataset)
- Each application or component can be replaced depending on local requirements as the node moves towards operational stage

Funded by
the European Union

GDI Starter Kit

Product	Outline	Functionality
Sensitive Data Archive	Securely stores data	
LifeScience AAI	Provides a federated Identity	
REMS	Tool to allow data access applications and decisions	
Beacon	Genetic and phenotypic data discovery	
Beacon Network	Federated network of Beacons	
Synthetic Data	Artificial anonymous data	
htsget	Secure genetic data distribution standard	
Containerised Computation	Computation via containers, e.g docker or singularity	
Federated Computation	Federated workflows, e.g. Nextflow	
FAIR Data Point	Datasets Metadata storage in FAIR format	
User Portal – Data Catalogue	European level catalogue of data within deployed nodes	
User Portal – Access management	European level data application and access management tool	

External to the starter kit

Node Connection

Funded by the European Union

Global Alliance for Genomics & Health

Infrastructure

1+MG
5 FUNCTIONALITIES

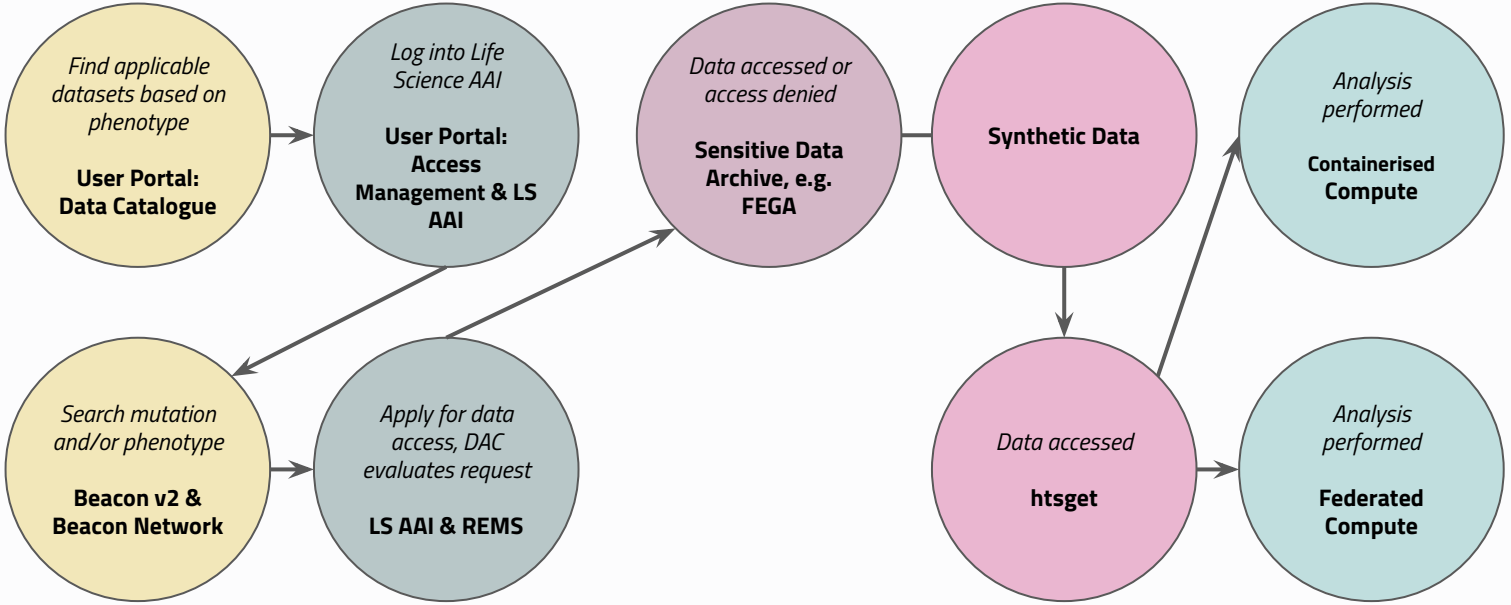
Data Discovery

Data Access Management

Storage & Interfaces

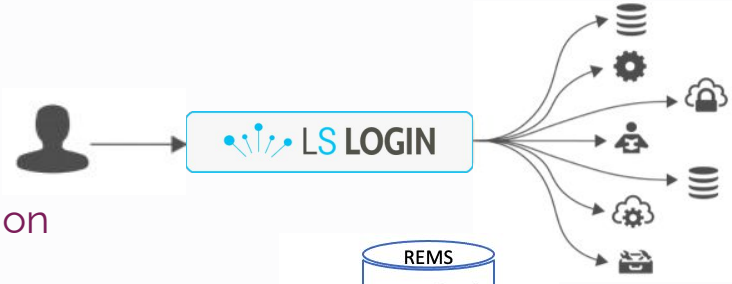
Data Reception

Data Processing



Funded by
the European Union

Data Access Management

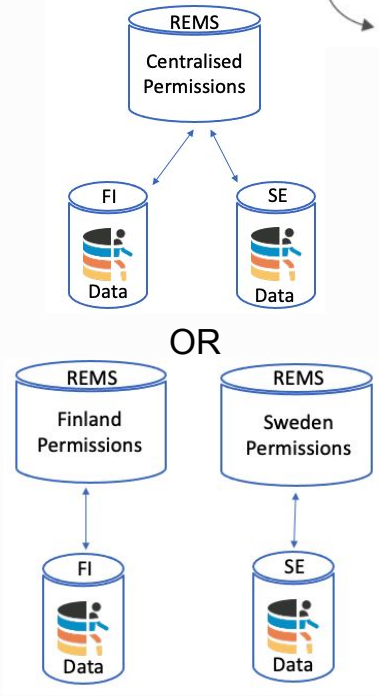


Function:

- **Management of data access** according to Data Protection by Default & Design, i.e. facilitation and audit of secure access
- Tools manage:
  - data access applications for data use
  - data access authorisations from the data controllers
  - communication of access rights
- **AAI is connected to most of the products.**

GDI Starter Kit Elements:

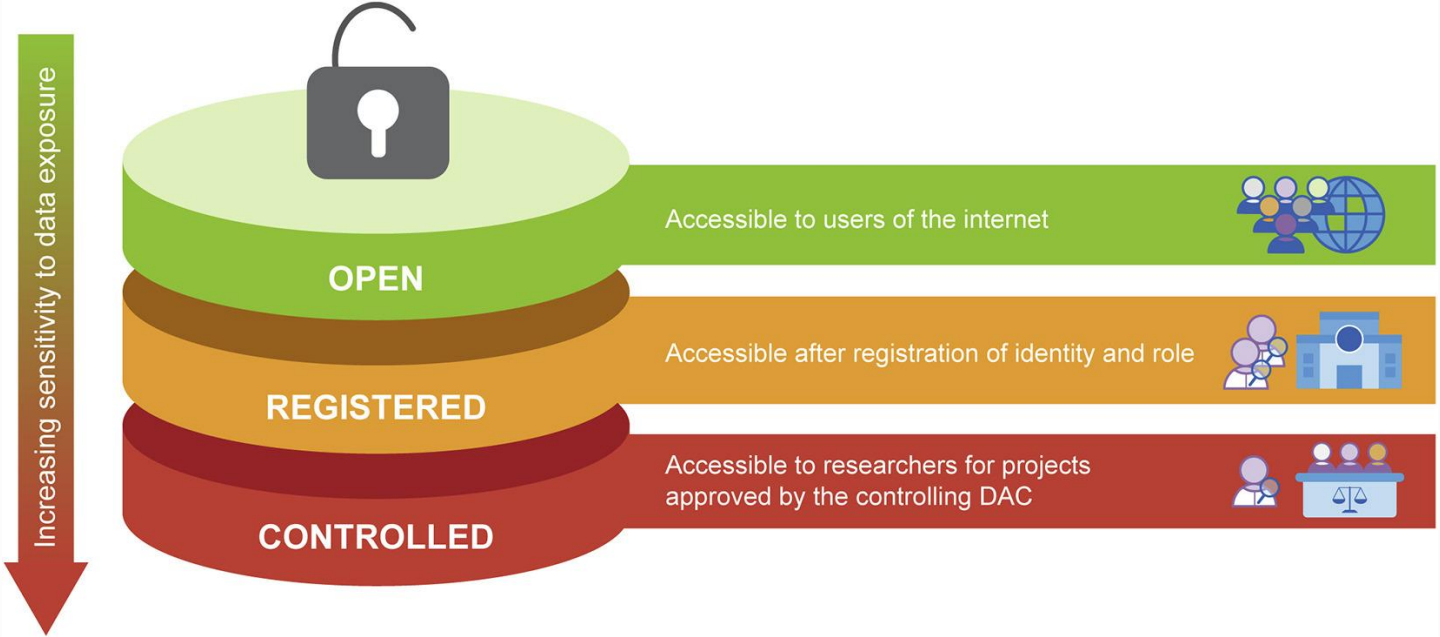
- LifeScience AAI - validate identity and permissions
- REMS - DAC approval
- User Portal: Access management - central monitoring



Data Discovery | Data Access Management | Storage & Interfaces | Data Reception | Data Processing

Funded by the European Union

Tiers of data access

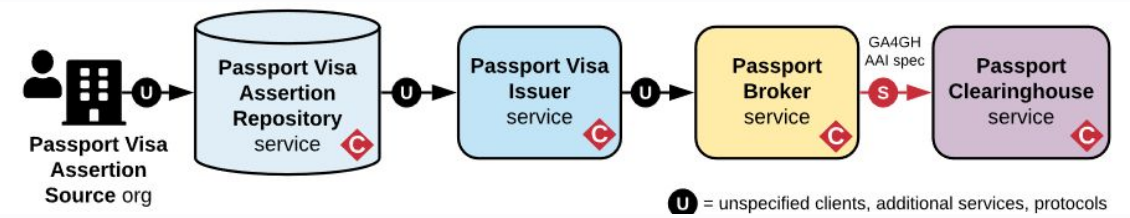


Datasets are shared in these tiers, depending on the regulatory requirements.

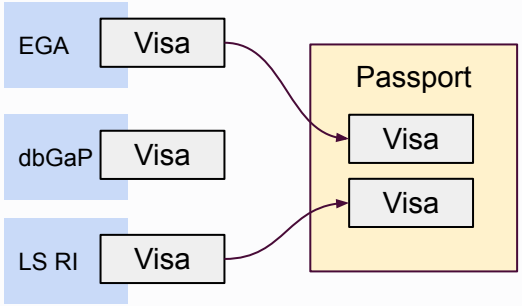
Dyke et al., 2021, *Eur J Hum Genet* **26**, https://doi.org/10.1038/s41431-018-0219-y

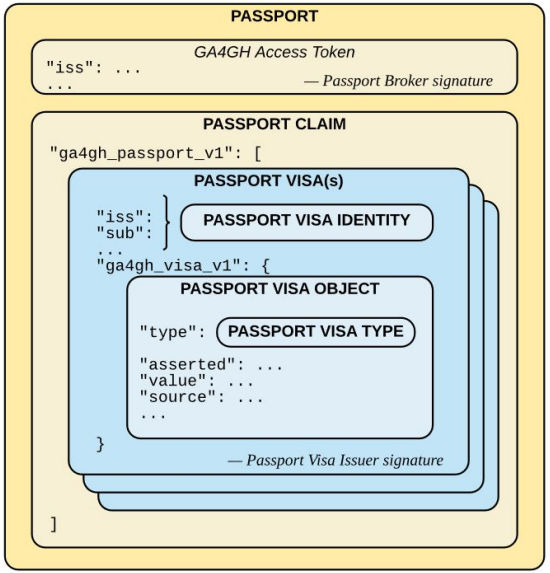
Funded by the European Union

GA4GH Passport in brief



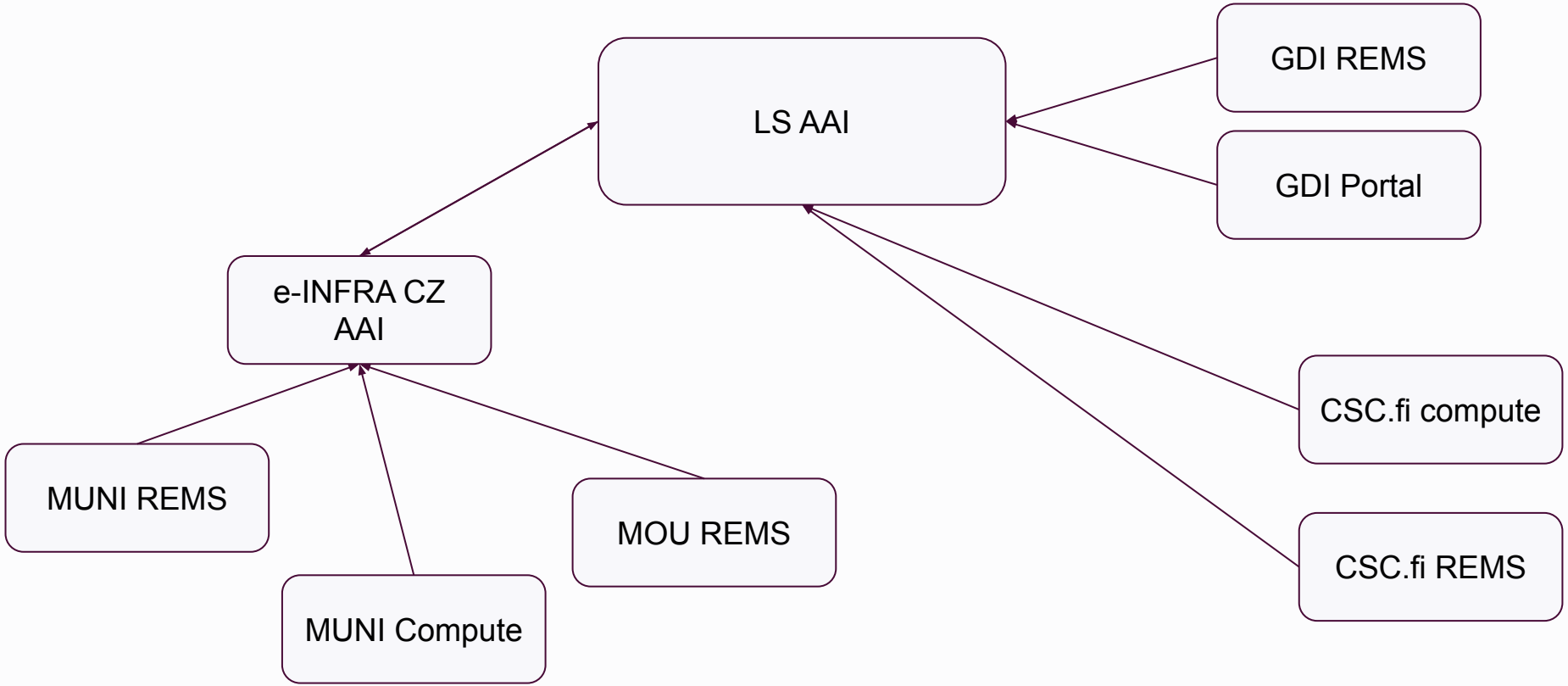
Visa type	Description
AffiliationAndRole	User's role within their institution - e.g. faculty@cam.ac.uk (eduPersonAffiliation)
AcceptedTermsAndPolicies	Acknowledged terms, policies, and conditions - e.g. attestations for registered access
ResearcherStatus	Bona fide researcher status - e.g. for registered access
ControlledAccessGrants	Permission to controlled access datasets - e.g. EGA, dbGaP
LinkedIdentities	Mapping of user identities - e.g. jdoe@elixir-europe.org equal to jdoe@lifescience-ri.eu





Funded by the European Union

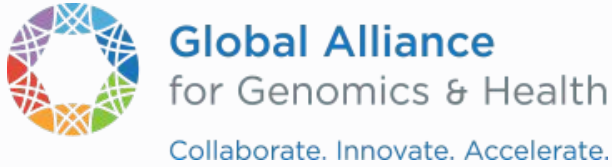
Incorporating existing national infrastructure



Funded by the European Union

Standardization

- AARC Blueprint Architecture
  - Including AARC Guidelines
  - Planned alignment with AARC TREE
- Identity federations, eduGAIN
- GA4GH Visas and Passports
- EOSC AAI
- **Standardization is the key**



Funded by
the European Union

Questions?

Funded by
the European Union