GA4GH Passport in the European Genomic Data Infrastructure - GDI

Tommi Nyrönen CSC

For 17th FIM4R meeting 15th February 2023 CERN Geneva

Funded by the European Union

GDI website

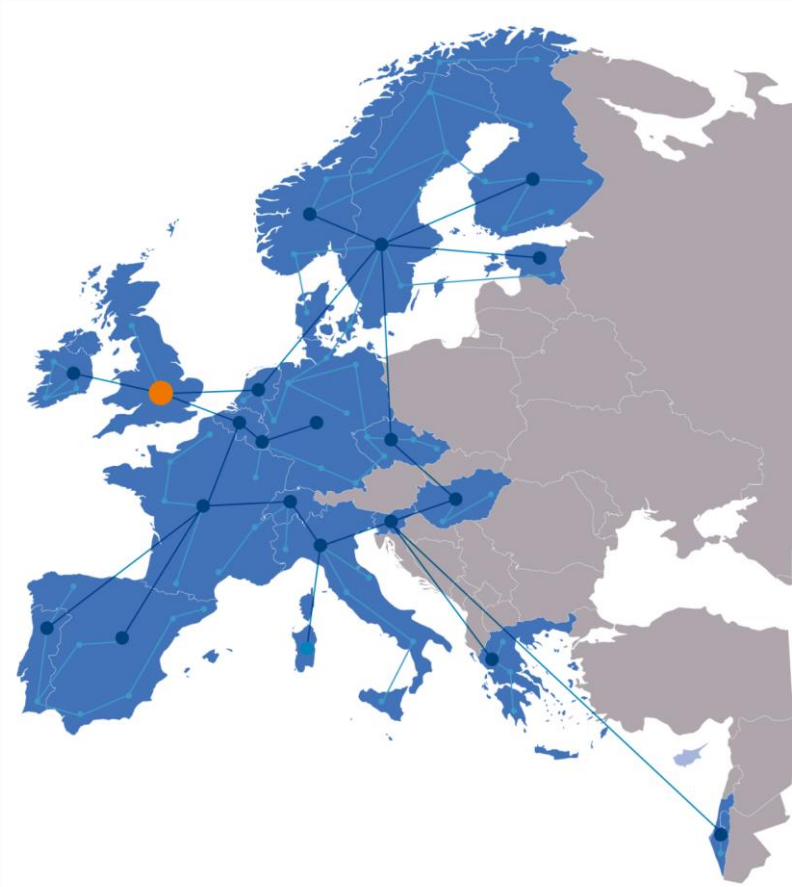
@GDI_EUproject

/company/gdi-euproject

ELIXIR

ELIXIR est. 2014 is an intergovernmental organisation that brings together life science resources such as databases, software tools, training materials, standards and compute resources, from across Europe.

- Find and share data
- Exchange expertise
- Agree on best practices in scientific research
- 22 members (countries)
- Part of FIM4R community since the very beginning



Sensitive data derived from human subjects is a challenge for data management services

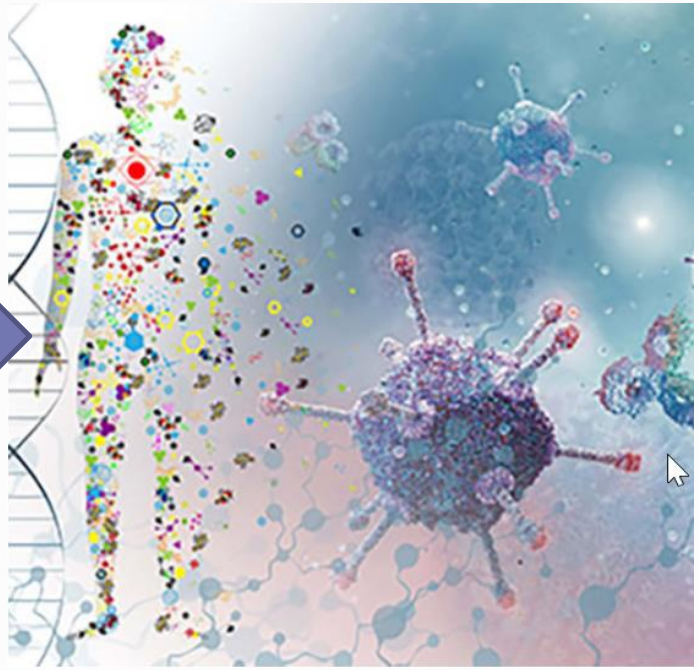
European Commission

What personal data is considered sensitive?

- personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs;
- trade-union membership;
- **genetic data**, biometric data processed solely to identify a human being;
- **health-related data**;
- data concerning a person's sex life or sexual orientation.

Article 4(13), (14) and (15) and Article and Recitals (51) to (56) of the GDPR

Precision medicine **AND** data protection

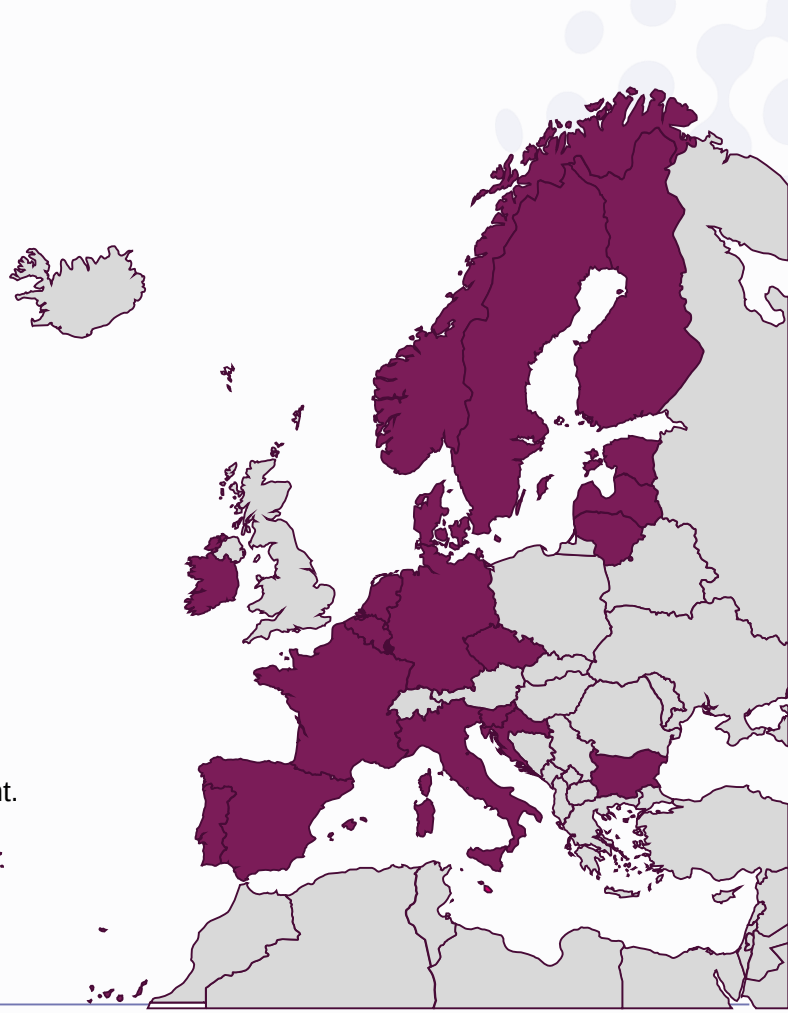


Funded by the European Union

20 EU countries

European Genomic Data Infrastructure

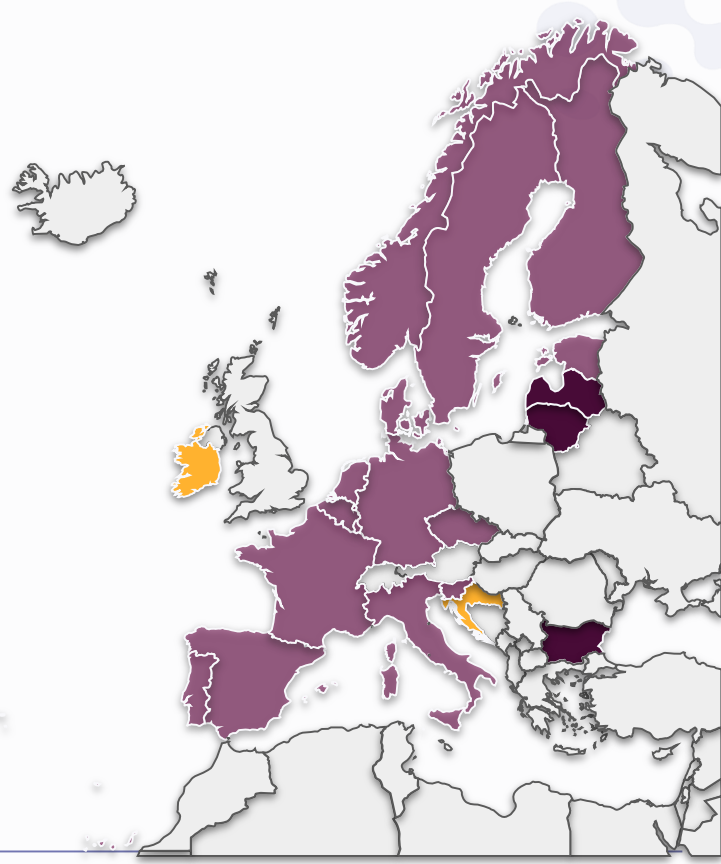
Project launched Nov 2022
Finland and Sweden lead data infrastructure deployment.



Funded by the European Union

Countries commitment to GDI by 2026

- Fully operational and integrated into 1+MG infrastructure: **Belgium, Czechia, Denmark, Estonia, Finland, France, Germany, Italy, Luxembourg, Portugal, Slovenia, Spain, Sweden, The Netherlands, Norway**
- Fully operational national node but not yet integrated in the 1+MG infrastructure: **Bulgaria, Latvia, Lithuania**
- Onboarding: **Croatia, Ireland**



Infrastructure will exist in 2024 with at least 6 countries

Funded by the European Union

What is GDI setting out to do?

Support the EU 1+Million Genomes (1+MG) initiative (Digital Europe policy) ambition to enable secure access to high-quality genomics and the corresponding clinical data across Europe for better research, personalised healthcare and health policy making

Establishing a **federated, sustainable and secure infrastructure based on open community standards** to access genomic and related phenotypic and clinical data across Europe

Building on the Beyond 1 Million Genomes (B1MG) project outputs

Funded by
the European Union

Data Infrastructure functionalities

Funded by
the European Union

1+MG infrastructure & TEHDAS Data Journey

Data Journey		Data Infrastructure
Data Preparation	Pre-processing to agreed standards, annotation with metadata etc	Data Reception
Data Inclusion	Physical transfer of data incl., legal transfer of data to 1+MG to enable visibility in data catalogue	
Data Storage & Management	Including GDPR Compliant processing environment, data versioning, backup etc	Storage and Interfaces
Data Discovery	Discovery of data using GDPR compliant APIs e.g., Beacon	Data Discovery
Data Access	A mechanism(s) by which the data controller can authorise access to select dataset(s)	Data Access Management Tools
Data Use	Data processing for the approved purposes	Data Processing

Funded by the European Union

Data Discovery

- Public visibility and search of genomic data, requiring descriptive non-sensitive metadata like summary level descriptions of data
- Link to Data Access Management where users can apply for detailed data access. Users need to Login

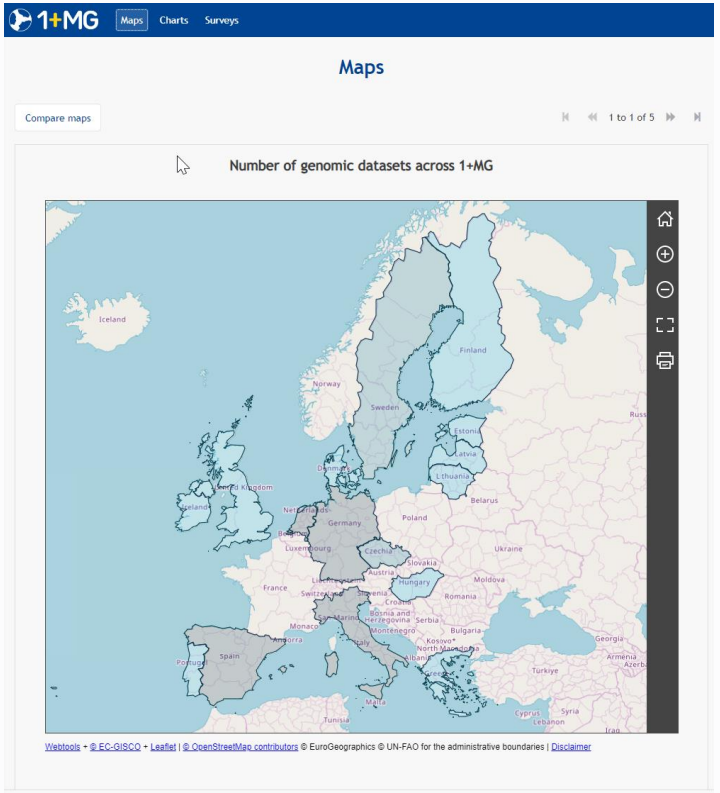
Data discovery | Access management tools | Data processing | Data reception | Storage and interfaces

Funded by the European Union

Data Reception

- Uniform processes such as quality control and standardisation
- Receiving (e.g. upload) or access (e.g. API) to data and metadata
  - Adhering to global standards and principles (e.g. GA4GH, FAIR)
  - Genotypic and phenotypic data
- Reception means logical description of datasets
  - Data becomes actionable on the 1+MG infrastructure even if they are stored nationally or locally

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



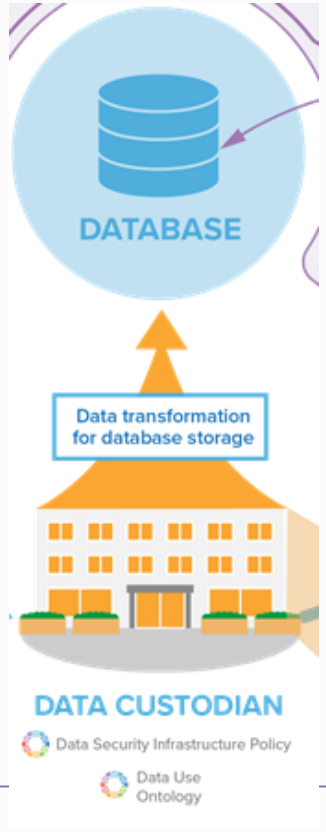
Funded by the European Union

https://dashboard.onemilliongenomes.eu/

Storage & Interfaces

- Organisations store data and offer interfaces (APIs) that form the technically interoperable (standard-based) infrastructure backbone
- Aim is to leverage existing investments in e-infrastructure capacities that can provide data privacy and confidentiality

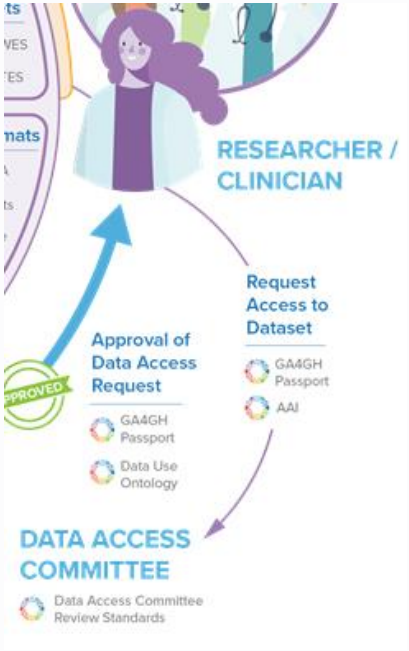
Data discovery | Access management tools | Data processing | Data reception | Storage and interfaces



Funded by the European Union

Data Access Management Tools

- Management of data access in an ELSI compliant way, i.e. facilitation and audit of secure access
  - e.g. central access portal, central access review process linking to data custodians, collaboration/data use agreement guidance
- Tools manage
  - user applications for data use
  - data access authorisations from the data controllers
  - communication of access rights to (distributed) infrastructure services providing data and processing



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

Funded by the European Union

Processing (compute)

- Local, high-performance and cloud computing with security standards appropriate for 1+MG to analyse data by user who has acquired the access rights for the intended data use
- Processing happens in many places on the infrastructure (local or distributed)
- Security specification & audit is likely for appropriate processing systems

Data discovery

Access management tools

Data processing

Data reception

Storage and interfaces

Funded by the European Union

Synthetic datasets

- Synthetic dataset is a genome and phonome data that does not relate to an identifiable natural, useful for driving technology forward in 1+MG
- In Finland produced in collaboration with high-performance computing

Funded by
the European Union

Proof-of-concept

Funded by
the European Union

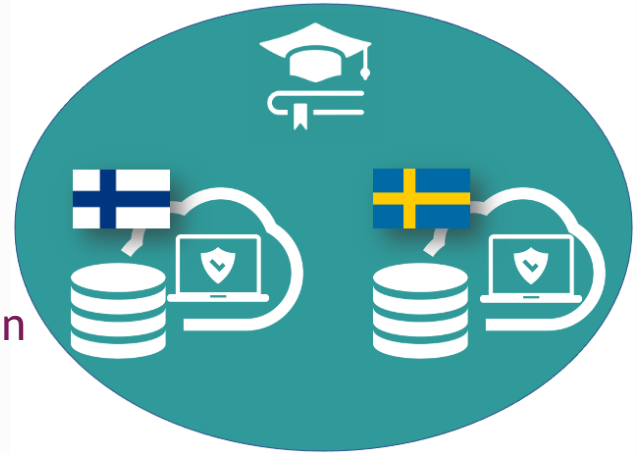
1+MG Proof of Concept - high-level structure

- Each 1+MG signatory country
  - Will provide a Data hub
  - Each country manages their own internal data (e.g. regional hubs)
  - Data hubs provide cross-border data analysis
  - Overall data infrastructure provides 5 main functionalities

Funded by the European Union

Rare Disease Scenario

- Rare disease researcher is investigating congenital myasthenic syndromes
  - Parents and child are all sequenced
  - Child has a de-novo mutation in the RYR1 gene
  - Clinical features include:
    - Neonatal hypotonia & distal athrogryposis
    - Inability to walk & recurrent lower respiratory tract infections
- Proof of Concept data is split between two data hubs in Sweden and Finland
- Driving questions for data infrastructure development



1. Are there any other individuals with the same mutation or allelic variant?
2. If there are, what is their phenotype? What can be derived from them for the prognosis?
3. What is the variant frequency across different populations?
4. Is there a relationship between the gene mutation and disease?

Funded by the European Union

Proof of concept – rare diseases

https://www.youtube.com/watch?v=6MtIJA4xXdU

Funded by
the European Union

General 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 genomic variant of interest
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





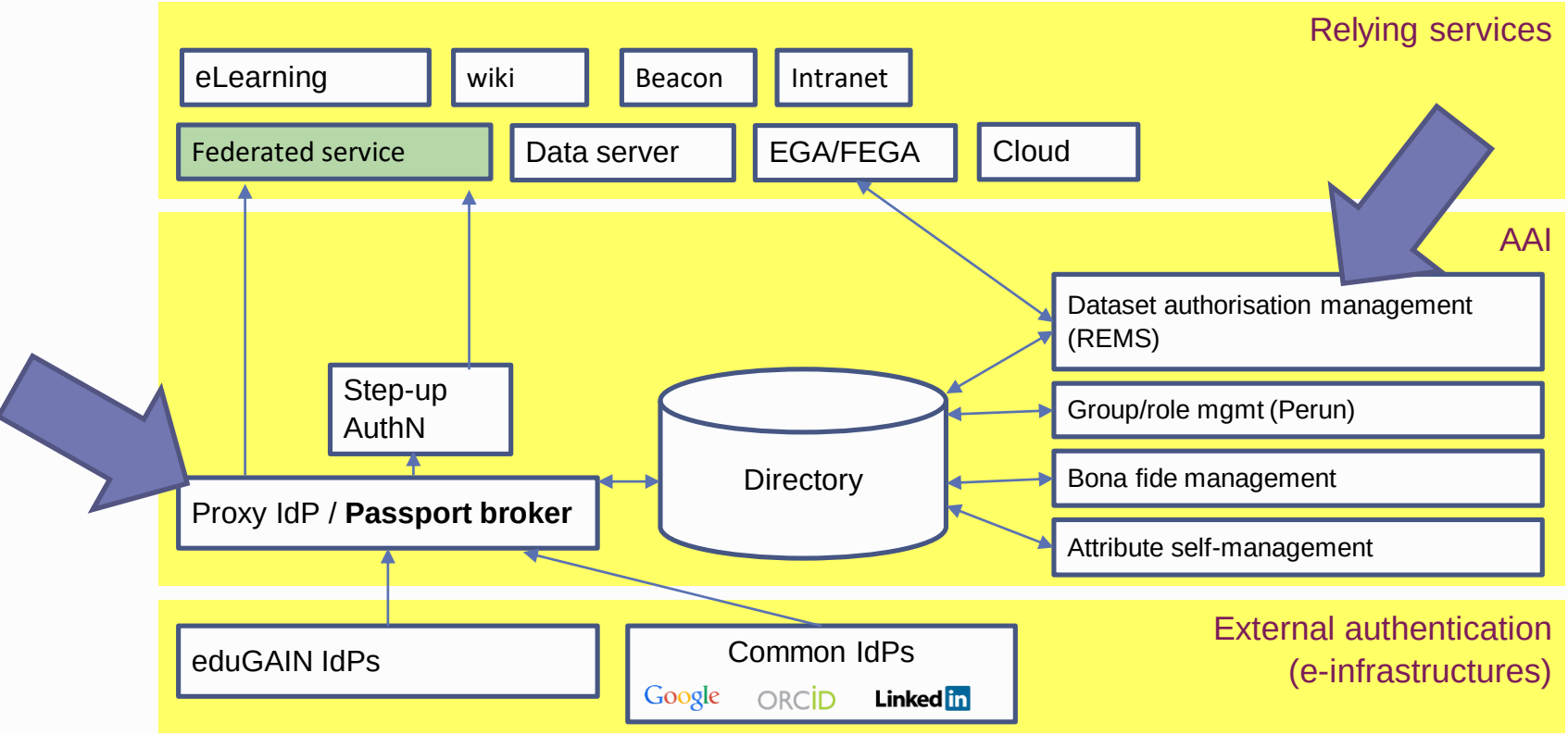

Funded by
the European Union

Under the hood -- Focus:

Federated data access

Funded by
the European Union

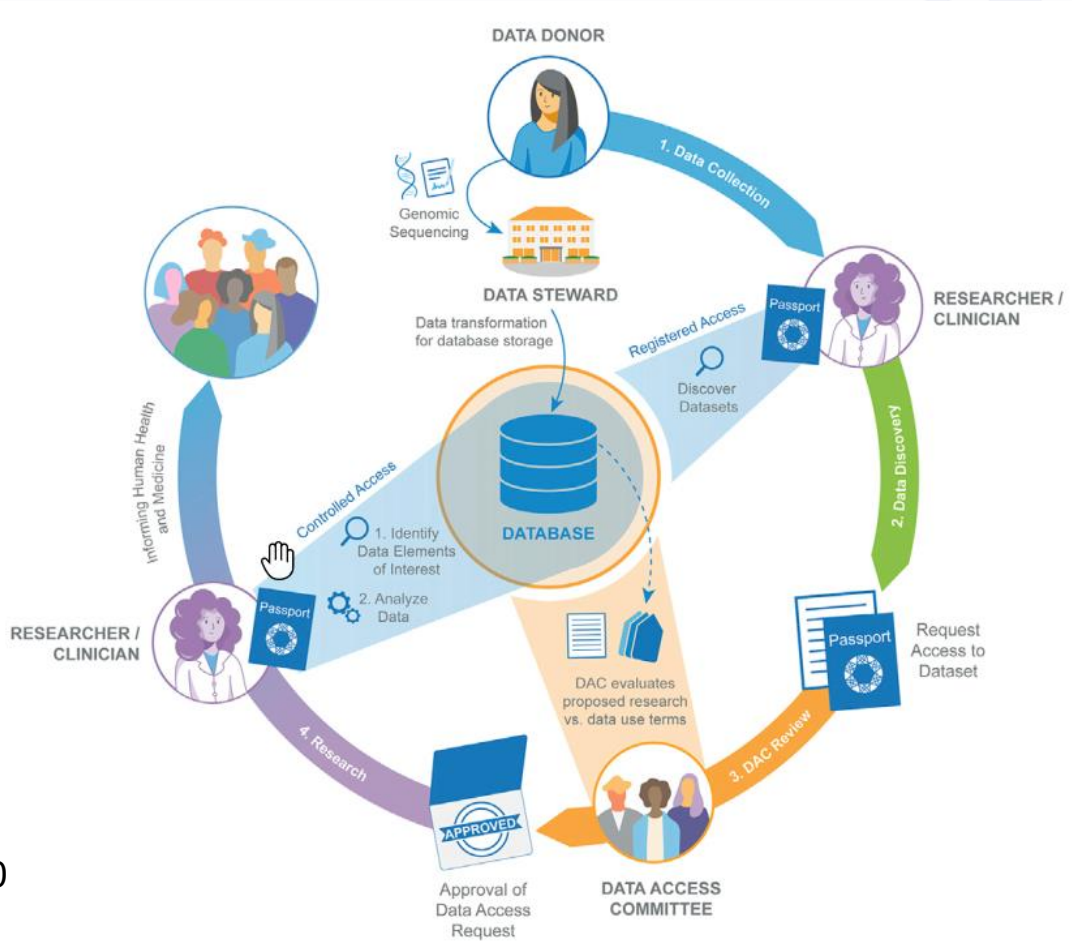
ELIXIR AAI original design – now moved to Life Science Login (2022)



GA4GH Passport

A standard to encode machine-readable data access permissions for individual users. Passports are used as part of a federated data regulatory process to authenticate and authorize data users in managing access to human biomedical datasets.

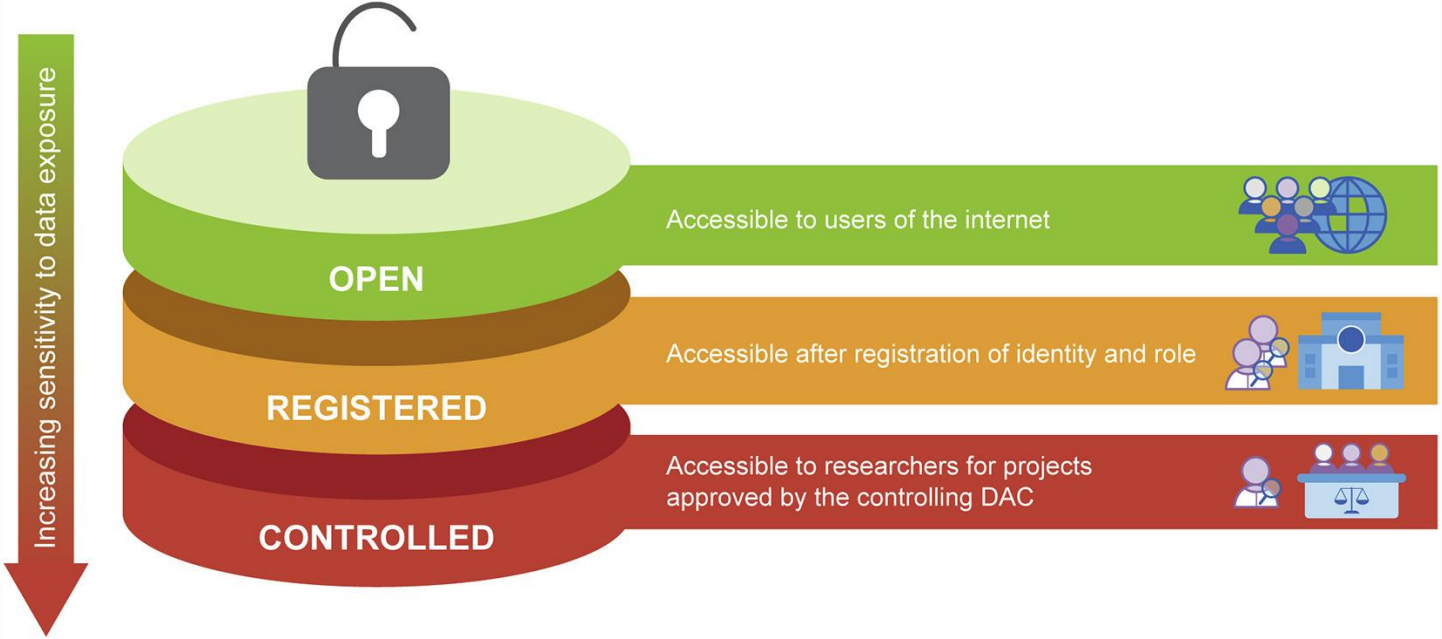
Passport is one of the technical standards used in GDI.



Voisin et al., 2021, Cell Genomics 1, 100030
https://doi.org/10.1016/j.xgen.2021.100030

Funded by the European Union

Tiers of data access

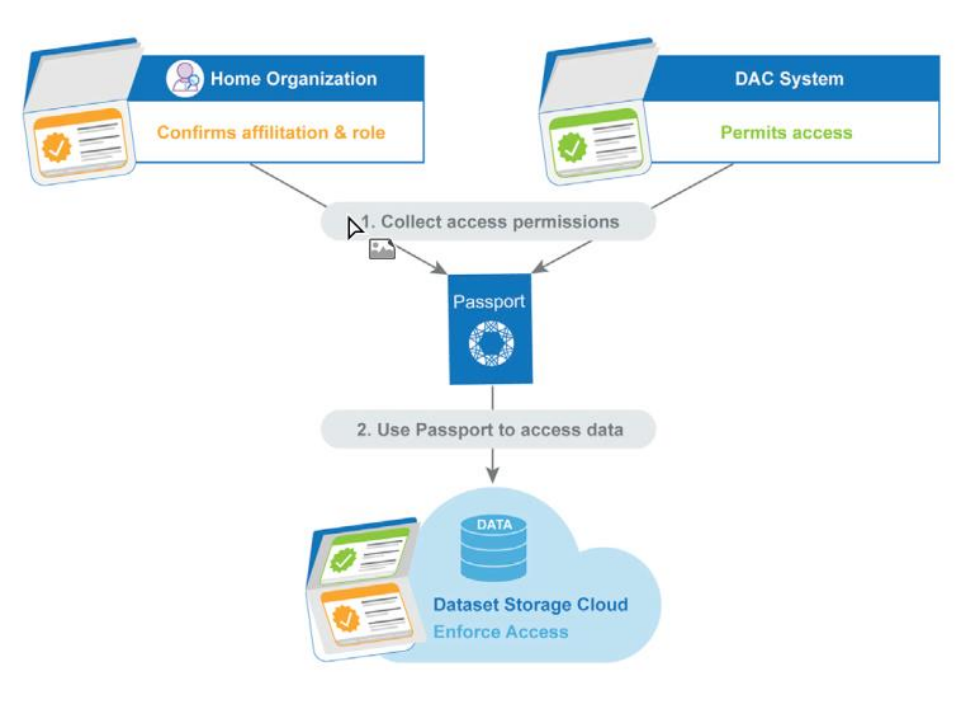


Datasets are commonly shared in databases with either open, registered, or controlled access, depending on the regulatory requirements.

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

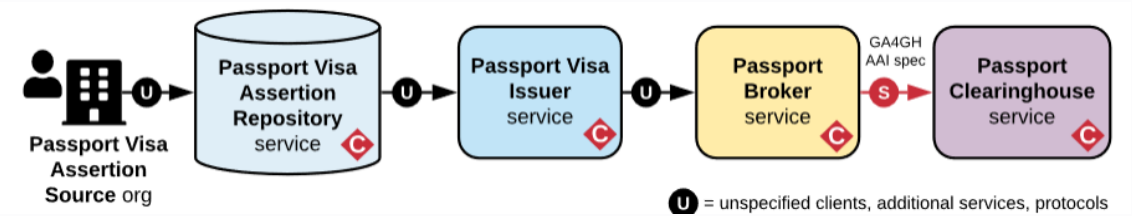
Funded by the European Union

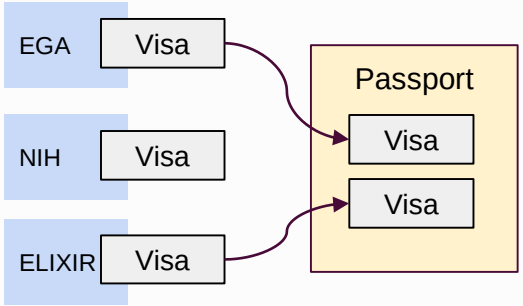
Federated data access with Passport



Funded by
the European Union

GA4GH Passport structure





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. mikael@elixir-europe.org equal to mlinden@csc.fi

Funded by the European Union

Controlled Access (ControlledAccessGrants Visa)

ELIXIR Broker can collect and deliver controlled access visa

- issued by EGA
- issued by REMS

Example JWT (decoded, header and signature stripped)


```
{
  "sub": "EGAW00000019020",
  "iss": "https://ega.ebi.ac.uk:8053/ega-openid-connect-server/",
  "ga4gh_visa_v1": {
    "type": "ControlledAccessGrants",
    "asserted": 1623936445,
    "value": "https://ega-archive.org/datasets/EGAD00001006673",
    "source": "https://ega-archive.org/dacs/EGAC00001000908",
    "by": "dac"
  },
  "exp": 1624274513,
  "iat": 1624270913,
  "jti": "7e223df4-ed66-4426-8b6c-fbb1c1f2de3e"
}
```


Funded by the European Union

Changes in Passport and AAI OIDC Profile 1.2

Global Alliance for Genomics & Health

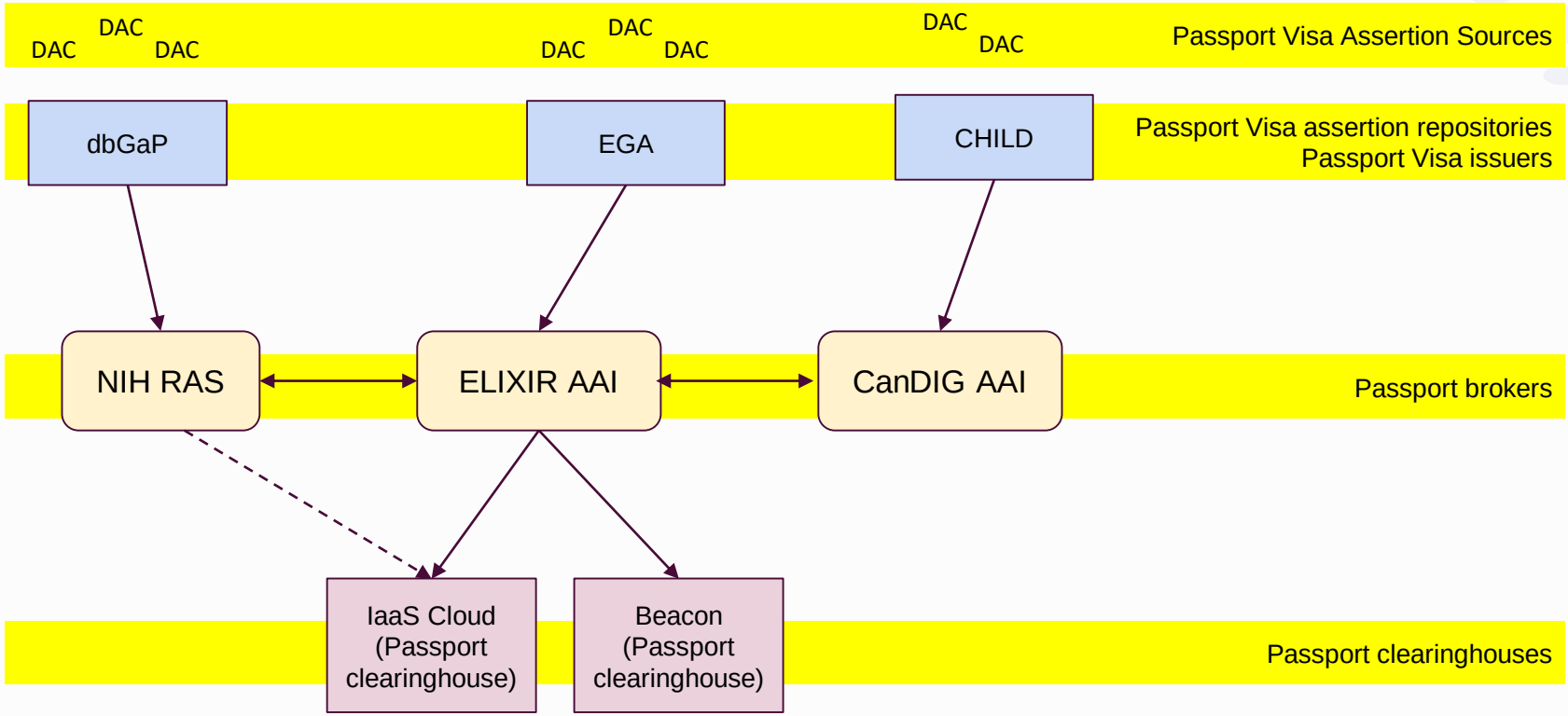
- https://ga4gh.github.io/data-security/changes-1_2
- Minor version, backward compatible with version 1.0
- **Terminology changes**
  - Unified and simplified terminology in both specs
  - Claim → Visa Assertion
  - Embedded Token → Visa
  - Passport Bearer Token → Passport-Scoped Access Token
- Introduced standard **OAuth 2.0 Token Exchange** mechanism from RFC 8693
- **Redefined Passport** as “a signed and verifiable JWT container for holding Visas“
  - Passport in 1.0 was a tuple of an access token and a list of Visas
  - Passport in 1.2 is a token that can be passed among systems
- Added **ES256** signing algorithm based on Elliptic Curve Cryptography (shorter keys than RSA)
- Media types for distinguishing multiple types of JWTs
  - Passports use `vnd.ga4gh.passport+jwt`, Visas `vnd.ga4gh.visa+jwt`

Main change - Passport redefined as a token

Global Alliance for Genomics & Health

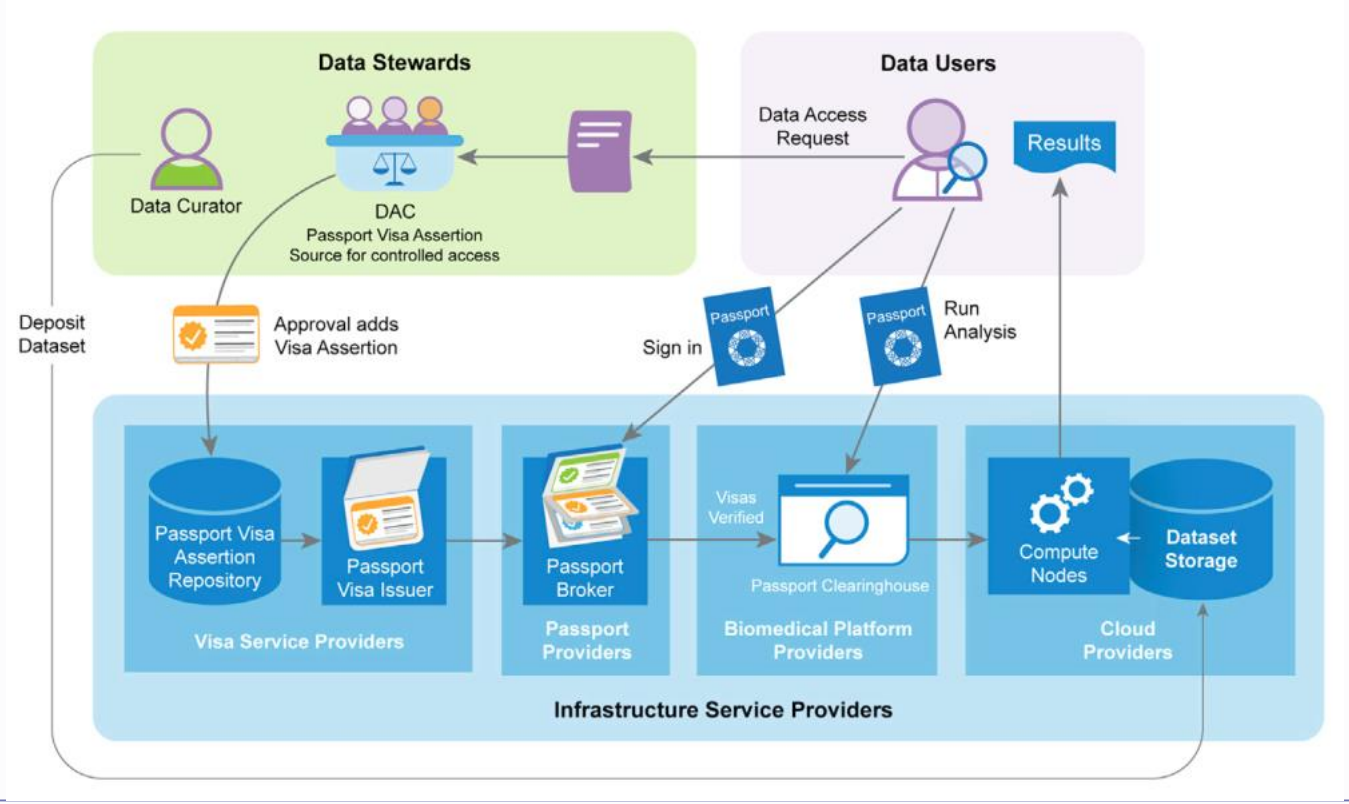
- in version 1.0
  - Passport was defined as ”*GA4GH-compatible access token along with the Passport Claim that is returned from Passport Broker service endpoints using such an access token*“
  - **a tuple of an access token and a list of Visas**
  - obtained from **UserInfo endpoint**
- in version 1.1 (proposed by NIH, not released, skipped number)
  - Passport was a JWT obtained from UserInfo endpoint
- in version 1.2
  - Passport is defined as “*a signed and verifiable JWT container for holding Visas*“
  - **a token** that can be passed among systems
  - obtained from **token endpoint** using OAuth 2.0 Token Exchange mechanism

Example deployment



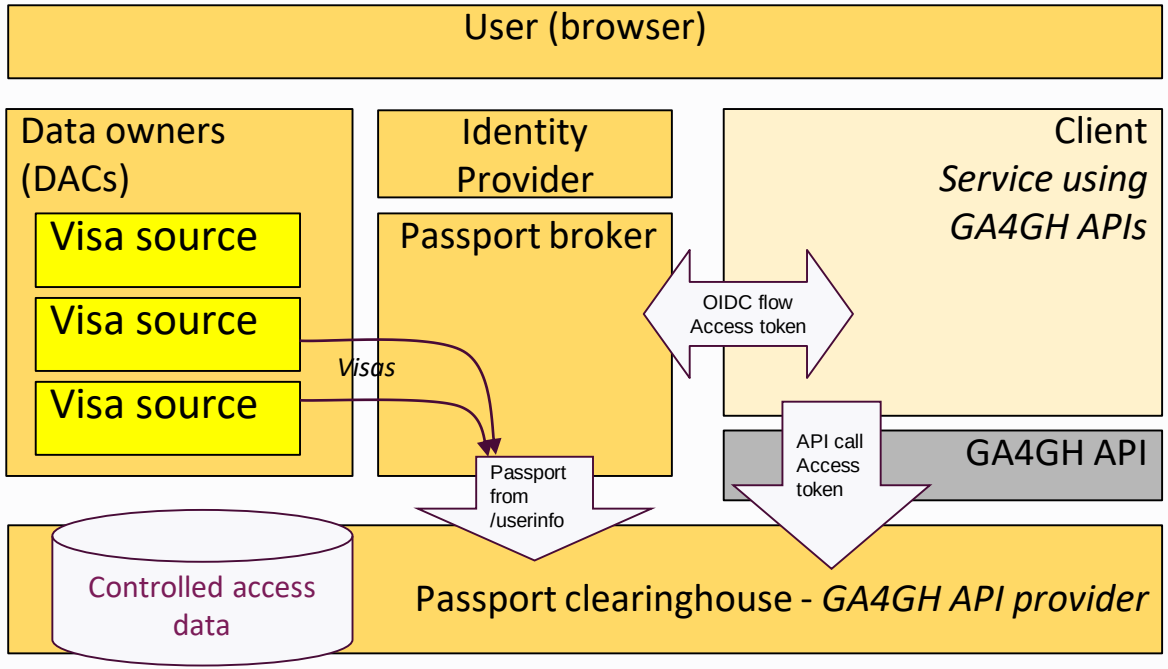
Funded by the European Union

An ecosystem of service providers needs to work together



Funded by
the European Union

Passports/AAI work



1. A client retrieves an access token from a broker
2. The client attaches the access token to the call of a GA4GH API provided by a Passport clearinghouse
3. Passport clearinghouse retrieves and validates the Passport
4. Passport clearinghouse enforces access to data based on the Visas

Funded by the European Union

Conclusions

Funded by
the European Union

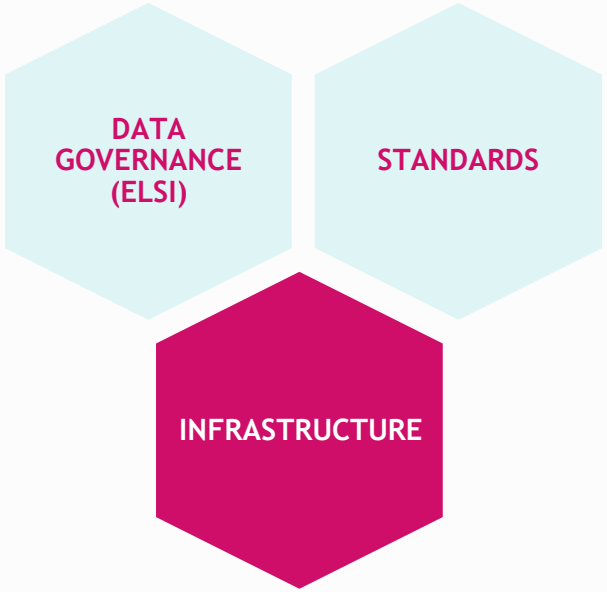
Passport prospects

- **Digital Europe eID** coming online – expected to help resolve identity vetting as part of LS AAI. Thus, more life science AAI technology development emphasis moves to federated authorisation.
- **GA4GH Passport** defines the format and types of **Visas** - JWTs carrying user attributes that define data access metadata coming from data access authoristies, this information is required on e-Infrastructure services
  - 5 standard types of Visas - AffiliationAndRole, AcceptedTermsAndPolicies, ResearcherStatus, LinkedIdentities and ControlledAccessGrants - and custom visas
  - "View my passport": https://echo.aai.elixir-czech.org/
- Currently implemented using OIDC. However! Self-sovereign Identity (SSI) also looks promising because the sources of data access decisions are distributed globally

Funded by
the European Union

GDI 2022-2027

- Advances **technical interoperability** between countries, which is a prerequisite for improved **semantic interoperability**
- Forms a **technical network of 1+MG infrastructure experts** who can train and provide knowledge and technology transfer between existing and aspiring nodes.
- **Co-develops existing services**, including dissemination of new developments and requirements of the infrastructure, across nodes, building technical capacity and resilience. Enabling growth in national expertise. **FIM4R!**
- Work in close collaboration with **use cases** and roll out innovative solutions on federated analysis and learning



Funded by the European Union

Acknowledgements

Funded by
the European Union

GDI Consortium

Funded by
the European Union

GDI Consortium (continued)

Funded by the European Union

CSC Collaboration

Genomikeskus GENOME CENTER

FINDATA TERVEYS & HYVINVOINTI

Global Alliance for Genomics & Health
Collaborate. Innovate. Accelerate.

EOSC-Life

Tryggve

BF Biocenter Finland

thl

CINECA

elixir FINLAND

de.NBI

FINNGEN

BBMRI.fi

Beacon

EMBL-EBI

EUROPEAN GENOME-PHENOME ARCHIVE

FINBB

VTT

SITRA

neic

ACADEMY OF FINLAND

Ministry of Education and Culture

Funded by the European Union

Tommi Nyrönen CSC

Thank you!


facebook.com/CSCfi


twitter.com/CSCfi


youtube.com/CSCfi


linkedin.com/company/csc---it-center-for-science


github.com/CSCfi