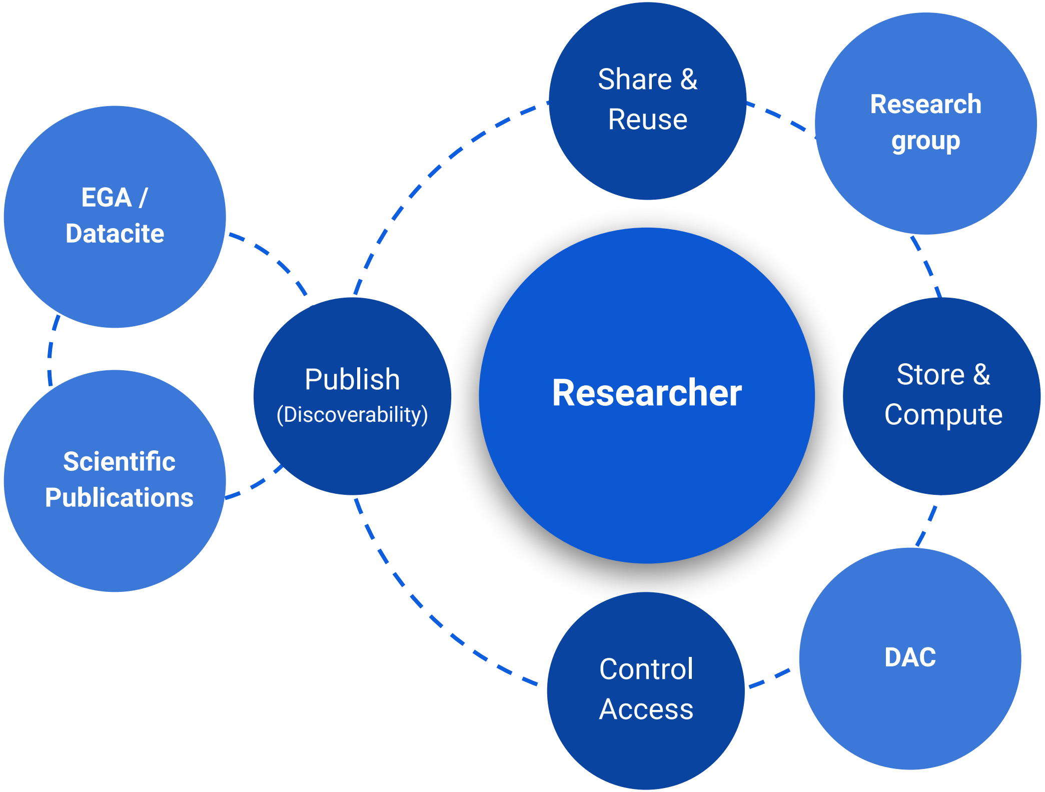
Developing Technologies And Standards for Enabling Sensitive Data Archiving, Sharing and Reuse

Stefan Negru

CSC – IT CENTER FOR SCIENCE

Understanding the landscape



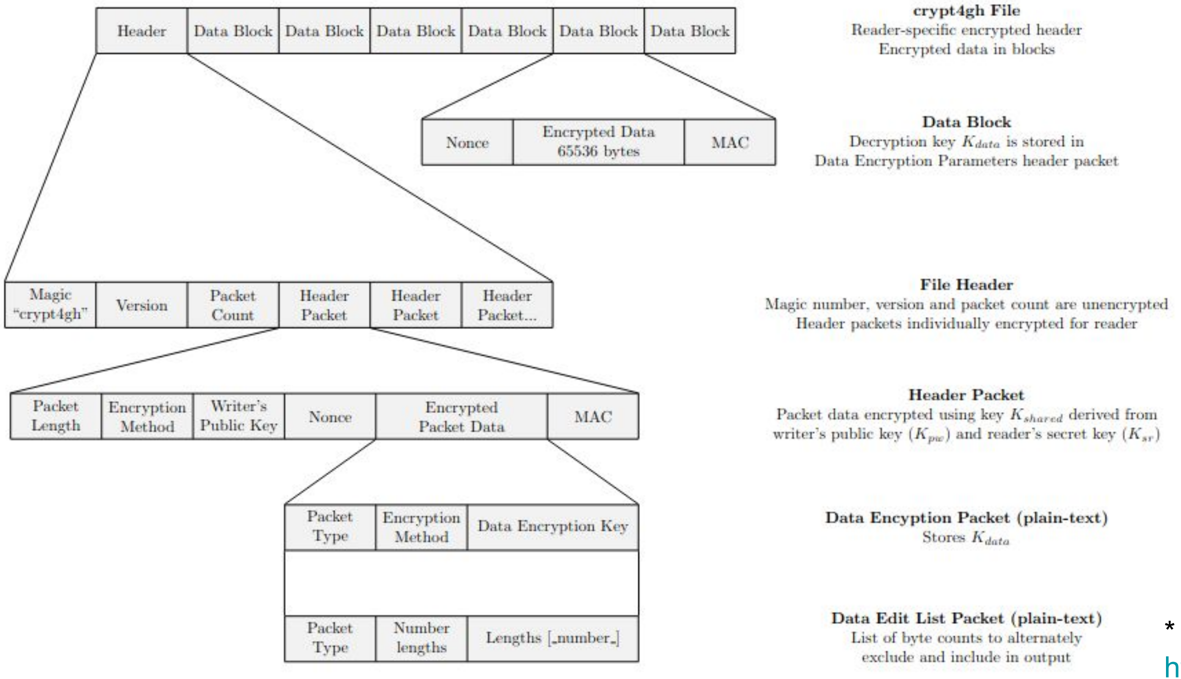
Some principles to consider …

- DARE (Data at Rest Encryption)
    - Facilitating data re-encryption
- FAIR (Findable Accessible Interoperable and Reusable)
    - Discoverable by other researchers
    - Reusable by other researchers
- Controlled Access to data - there is a DAC that controls access
    - Specify and encode permissions for the data, what can be used for and till when

A few APIs/Standards

API/Standards Name	API/Standard Purpose	URL to Spec
htsget	A protocol for secure, efficient and reliable access to sequencing read and variation data	http://samtools.github.io/hts-specs/htsget.html
DUO	Allows users to semantically tag genomic datasets with usage restrictions, allowing them to become automatically discoverable based on a health, clinical, or biomedical researcher's authorization level or intended use.	https://github.com/EBISPOT/DUO
Refget	Refget enables access to reference sequences using an identifier derived from the sequence itself.	http://samtools.github.io/hts-specs/refget.html
Crypt4GH	A file container specification enabling direct byte-level compatible random access to encrypted genetic data stored in community standards such as SAM/BAM/CRAM/VCF/BCF.	http://samtools.github.io/hts-specs/crypt4gh.pdf
Beacon	Discover genomic variants, individuals, and individuals	https://github.com/ga4gh-beacon/beacon-v2/
Phenopackets	It merges the existing GA4GH metadata-schemas work with a more focused model from the phenopacket-reference-implementation.	https://phenopacket-schema.readthedocs.io/en/latest/
GA4GH Passports	GA4GH Passport specification aims to support data access policies within current and evolving data access governance systems.	https://github.com/ga4gh-duri/ga4gh-duri.github.io/blob/master/researcher_ids/ga4gh_passport_v1.md
WES API	Workflow Execution Service API describes a standard programmatic way to run and manage workflows. Having this standard API supported by multiple execution engines will let people run the same workflow using various execution platforms running on various clouds/environments.	https://github.com/ga4gh/workflow-execution-service-schemas

Encryption - Crypt4GH *



crypt4gh File
Reader-specific encrypted header
Encrypted data in blocks

Data Block
Decryption key K_{data} is stored in
Data Encryption Parameters header packet

File Header
Magic number, version and packet count are unencrypted
Header packets individually encrypted for reader

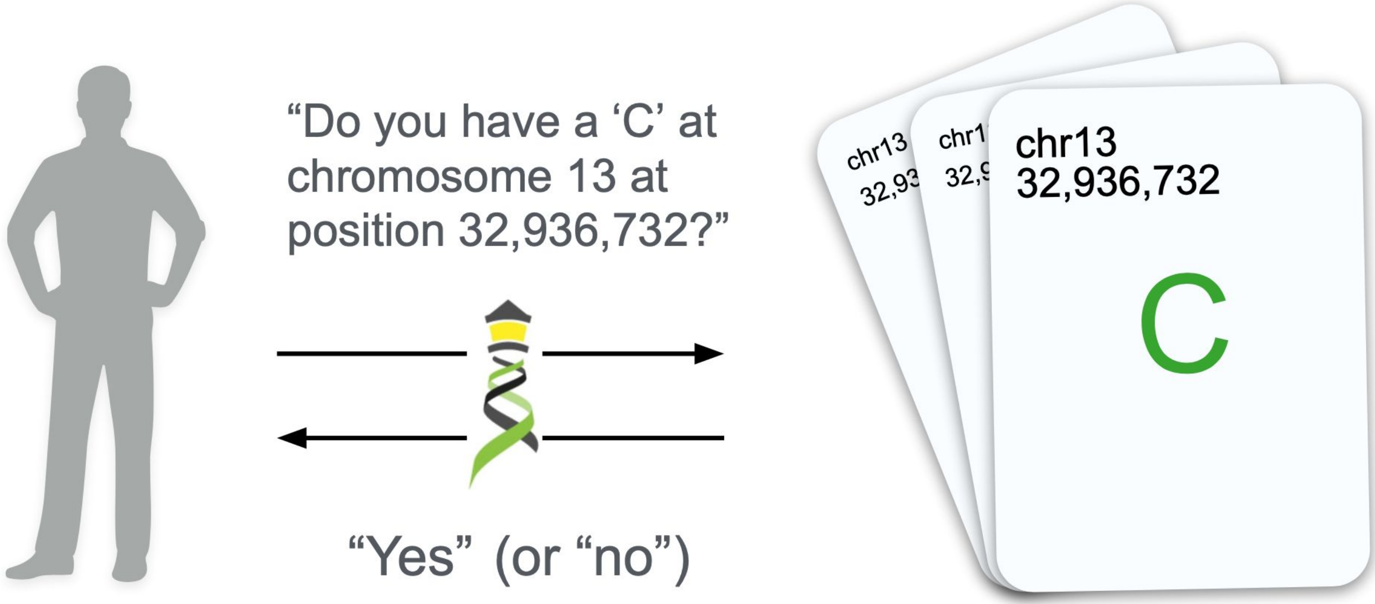
Header Packet
Packet data encrypted using key K_{shared} derived from
writer's public key (K_{pw}) and reader's secret key (K_{sr})

Data Encyption Packet (plain-text)
Stores K_{data}

Data Edit List Packet (plain-text)
List of byte counts to alternately
exclude and include in output

* specification & figure provided by
http://samtools.github.io/hts-specs/crypt4gh.pdf

Discovery - Beacon *



“Do you have a ‘C’ at chromosome 13 at position 32,936,732?”

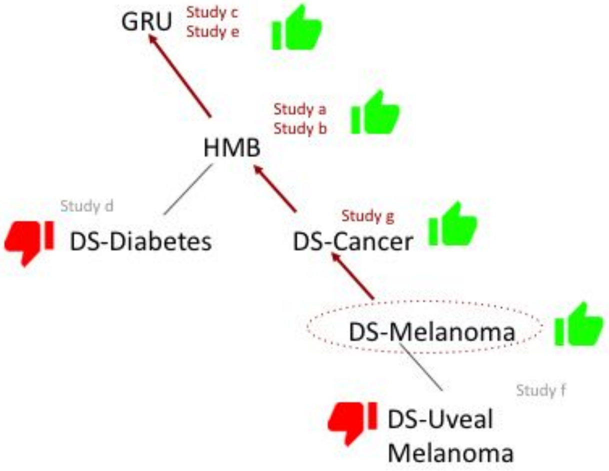
“Yes” (or “no”)

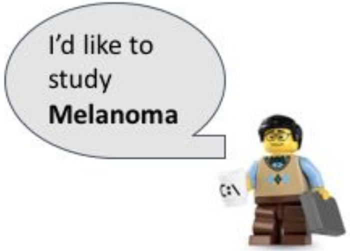
* Figure provided by:http://docs.genomebeacons.org/what-is-beacon-v2/

Controlled Access - Data Use Ontology

Data Repository

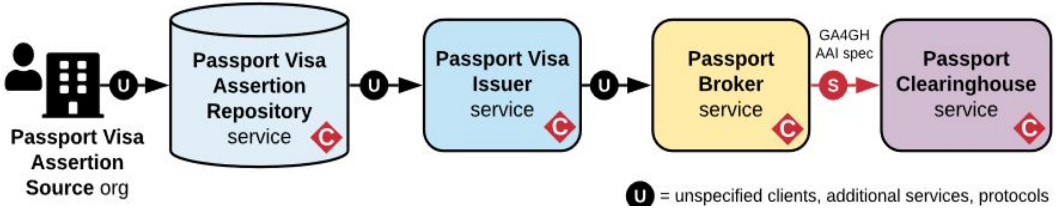
Dataset	DU restriction
Study a	HMB
Study b	HMB
Study c	GRU
Study d	DS- Diabetes
Study e	GRU
Study f	DS- Uveal Melanoma
Study g	Cancer

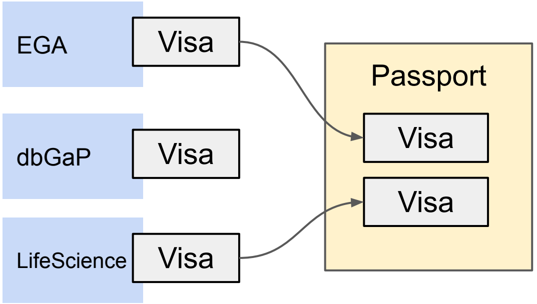




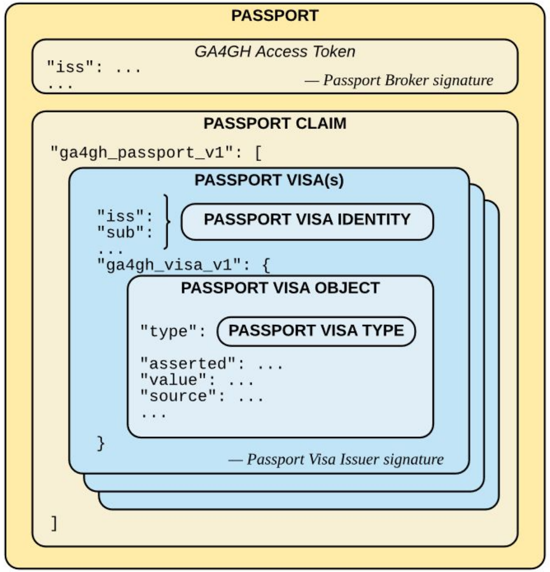
* Figure provided by: https://github.com/EBISPOT/DUO

Controlled Access - GA4GH Passport in brief

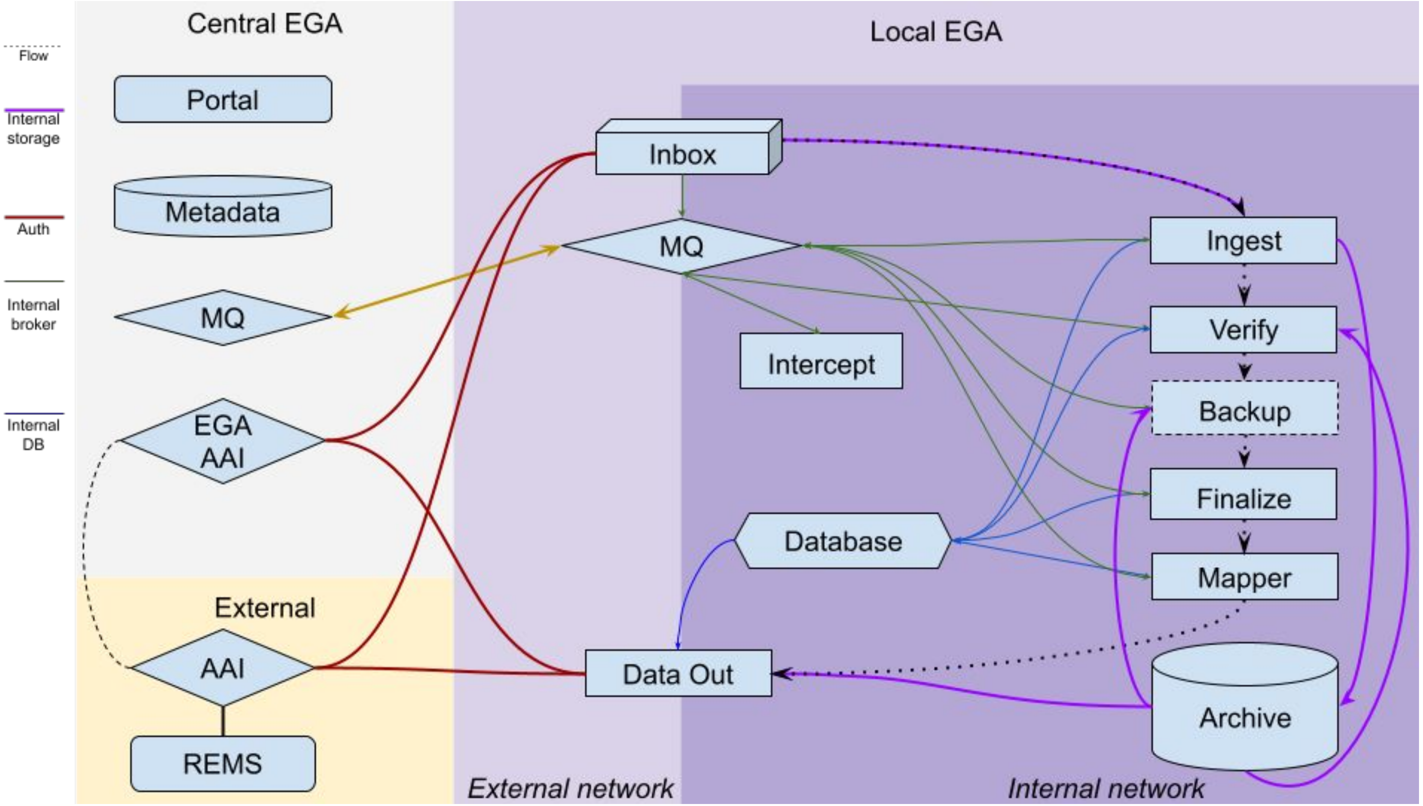

Passport Visa Assertion Source org
U
Passport Visa Assertion Repository service
C
U
Passport Visa Issuer service
C
U
Passport Broker service
C
GA4GH AAI spec
S
Passport Clearinghouse service
C
U = unspecified clients, additional services, protocols



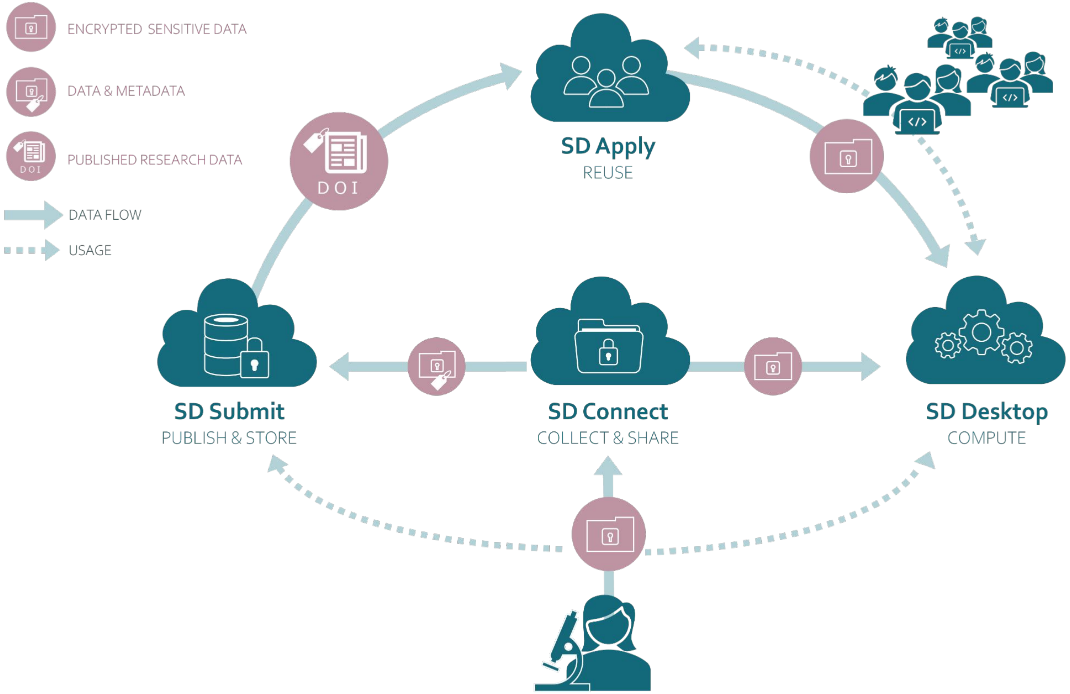
Visa type	Description
AffiliationAndRole	User’s role within their institution - e.g. faculty@cam.ac.uk (eduPersonAffiliation)
AcceptedTermsAndPolicies	Acknowledged terms, policies, and conditions - e.g. attestations for registered access
ResearcherStatus	Bona fide researcher status - e.g. for registered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



Sensitive Data Archive - NeIC Heilsa solution



CSC - Sensitive Data Services for Research *



* more information at: https://research.csc.fi/sensitive-data-services-for-research

Additional Resources

- Crypt4GH libraries
    - Python https://github.com/EGA-archive/crypt4gh
    - Go https://github.com/neicnordic/crypt4gh
    - Rust https://github.com/EGA-archive/crypt4gh-rust
    - C https://github.com/silverdaz/crypt4gh
    - Java https://github.com/uio-bmi/crypt4gh
    - Samtools https://github.com/samtools/htslib-crypt4gh
- NeIC Nordic:
    - https://github.com/topics/neic-sda
- Beacon
    - https://github.com/CSCfi/beacon-python
    - https://github.com/CSCfi/beacon-network
    - https://github.com/EGA-archive/beacon2-ri-api

Thank you!