ELIXIR – Sensitive Data

Approaches to Access and Analysis of Sensitive Data

WUN DAPPER Workshop

Bristol 23 August 2016

www.elixir-europe.org

Alexander Senf, PhD

- EMBL-EBI (ELIXIR Node) 5+ Years
- EGA Team
- Secure Data Access / Integration of access to EGA Data into ELIXIR
  - ELIXIR EXCELERATE Work Packages 4 & 9

(*) EGA: European Genome-phenome Archive

Outline

- **ELIXIR**
  - Structure, Goals, Funding
  - Implementation
- EGA
  - Details, Plans, Local EGA
- ELIXIR Sensitive (Human) Data
  - Individual Projects
- Summary

ELIXIR - Origins

Press Release

EMBL-EBI

Securing the future of Europe's biological data resources

New EC-funded effort to create a sustainable infrastructure for biological data launched

Hinxton, 28 May 2008 – In a contract signed today, the European Commission has awarded €4.5 million to a pan-European consortium to decide upon the best way to unite Europe's biological data resources into a sustainable, integrative bioinformatics network for the life sciences. The European Life-science Infrastructure for Biological Information (ELIXIR) project is led by information storage, access and integration supported by a secure funding mechanism. "Under the Framework Programme for Research, the European Commission provides support to the preparatory phase for the construction of new research infrastructures," commented Mr Robert-Jan Smits, Director of Directorate B (European Research Area: Research

http://www.embl.de/aboutus/communication_outreach/media_relations/
2008/080528_hinxton/press28may08.pdf

ELIXIR - Distributed infrastructure for life science data

- ELIXIR: consolidates Europe's national centres, services, and core bioinformatics resources into a single, coordinated infrastructure.
    - Simplifies access to quality-controlled data and services
    - Ensures long-term sustainability of Europe's core resources
    - Facilitates open access and open science in life science research
- Implemented through National ELIXIR Nodes, coordinated by ELIXIR Hub
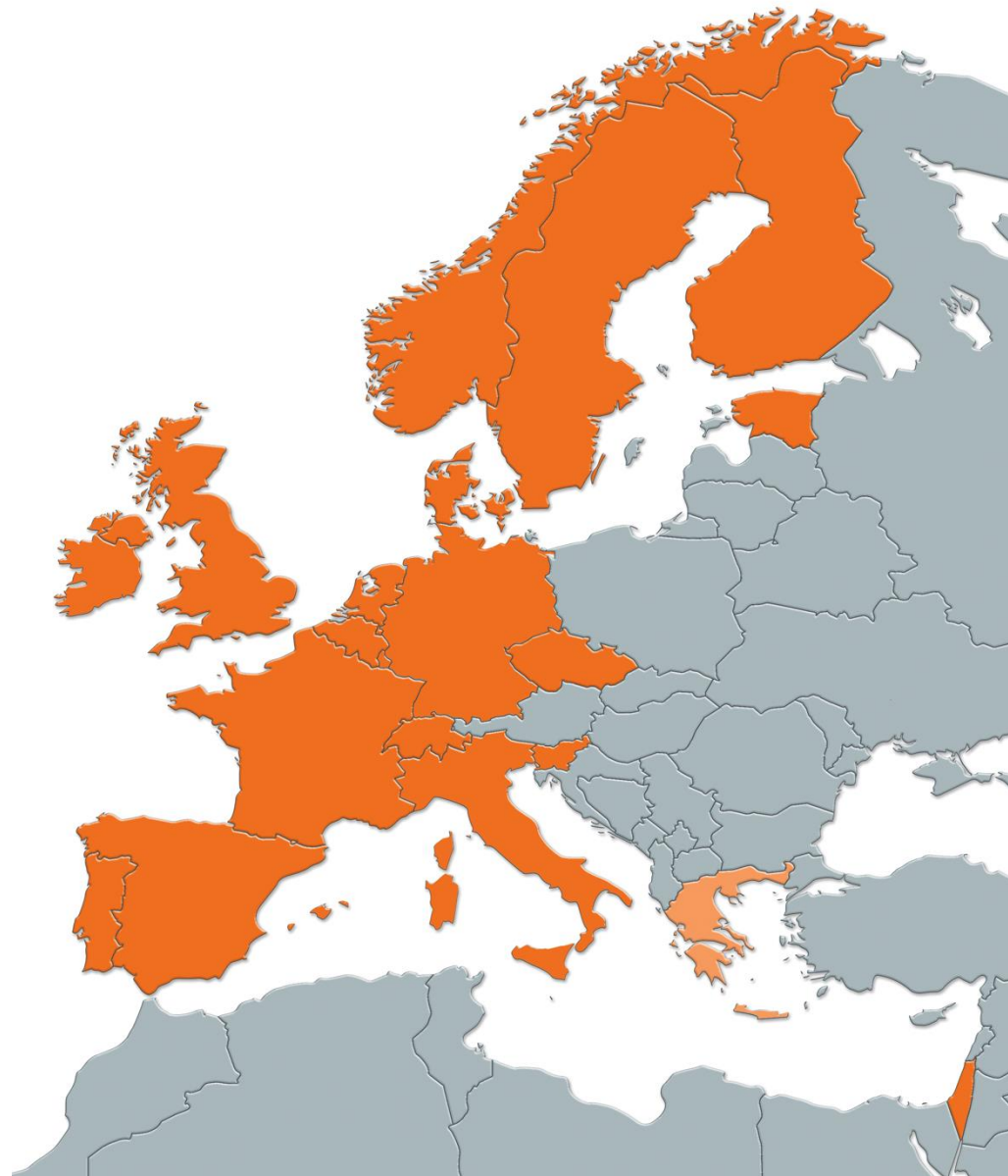   - Nodes pursue their own priorities and interests

www.elixir-europe.org

ELIXIR Membership

ELIXIR Members

Belgium	Czech Republic	Denmark	EMBL
Estonia	Finland	France	Germany
Ireland	Italy	Israel	Luxembourg
Netherlands	Norway	Portugal	Slovenia
Spain	Sweden	Switzerland	United Kingdom

ELIXIR Observers

Greece

Latest member – Germany! (joined in August 2016)

ELIXIR Nodes

- Each National Node pursues their own Plans and Priorities

- Coordinated and Integrated through the ELIXIR Hub

ELIXIR Services

Data deposition:
ENA, EGA, PDBe, EuropePMC, …

Data management:
Genome annotation
Data management plans

Added value data:
UniProt, Ensembl, OrphaNet, …

Data Interoperability:
Standards,Identifiers,
Ontologies

Compute:
Secure data transfer, cloud
computing, AAI

Bioinformatics tools:
Bio.tools

Industry:
Innovation and SME programme
Bespoke collaborations

Training:
TeSS, Data Carpentry,
eLearning

(*) AAI = Authentication and Authorisation Infrastructure

Technical Coordination Priorities

- Don't Create Solutions which already exist

- Make use of / connect existing Resources
  - EUDAT, EGI, GridFTP, EMBL-EBI
  - Commission Missing / Glue services

- Guided by Scientific Pilot Use Cases
  - Ensure services provide what is needed for the Science

- Pursue funding to help with implementation
  - Horiozon 2020 4-year grant, started December 2015
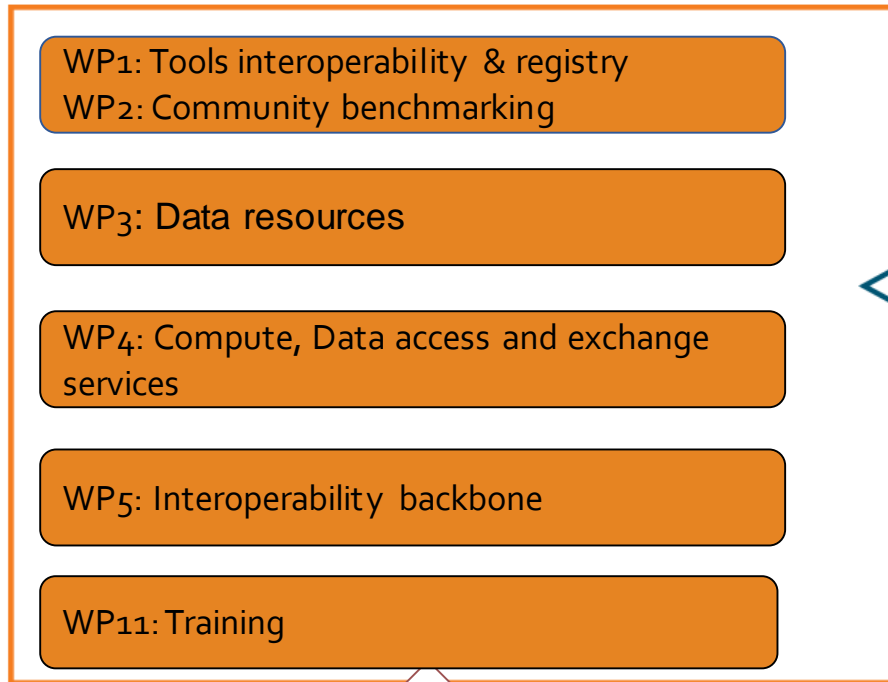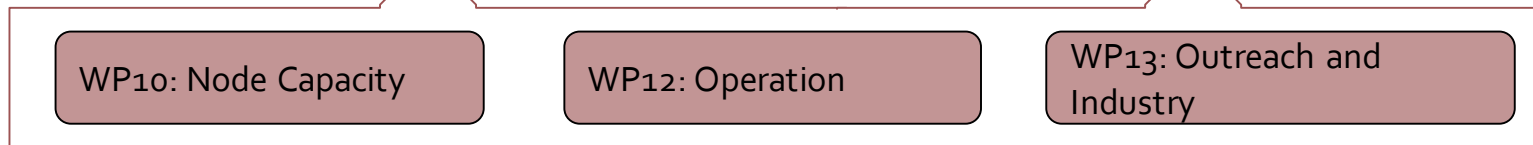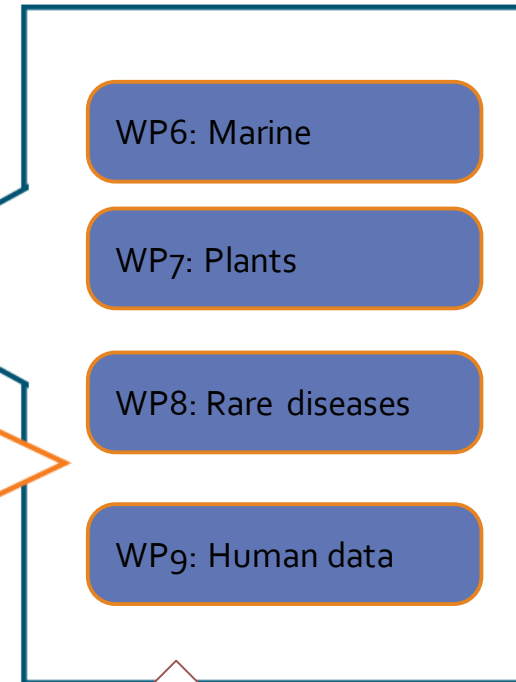 - EXCELERATE

EXCELERATE Structure "Work Packages"

Build ELIXIR platforms

WP1: Tools interoperability & registry
WP2: Community benchmarking

WP3: Data resources

WP4: Compute, Data access and exchange services

WP5: Interoperability backbone

WP11: Training

Use ELIXIR Services

WP6: Marine

WP7: Plants

WP8: Rare diseases

WP9: Human data

WP10: Node Capacity

WP12: Operation

WP13: Outreach and Industry

Embed and sustain ELIXIR

excelerate

EXCELERATE Work Packages – Early Results

- Tools: http://bio.tools
  - Available Tools, Benchmarking; World-wide

- Compute
  - Data Storage, Movement, Cloud, Identity Management
  - https://www.elixir-europe.org/documents/manual-how-connect-service-elixir-aai

- Training: http://tess.elixir-uk.org

Use Case: Human data

Mission: develop a long-term management policy for controlled-access human data and provide tools for data discovery and access

EGA as the central ELIXIR service for secure sharing of -omics data

Projects:

- Connect TraIT platform (NL) with EGA
- IMI OncoTrack – explore long-term storage and data-discovery for OncoTrack's data
- GA4GH Beacons – discovery service for genomics data in EGA
- Local EGA - Stores sensitive data locally Integrates metadata globally in EGA (part of ELIXIR Data Nodes network)

EUROPEAN GENOME-PHENOME ARCHIVE

elixir

Use Case: Rare Diseases research

- Mission: support the development of new therapies for rare diseases

- Collaboration with RD-CONNECT, BBMRI-ERIC, BBMRI-NL

- Actions:

  - Build ELIXIR portfolio of data resources and analysis tools rare disease research

  - Implement technical framework for comparison and standardization of services for rare-disease communities

  - Develop and deliver training courses and workshops for rare disease researchers
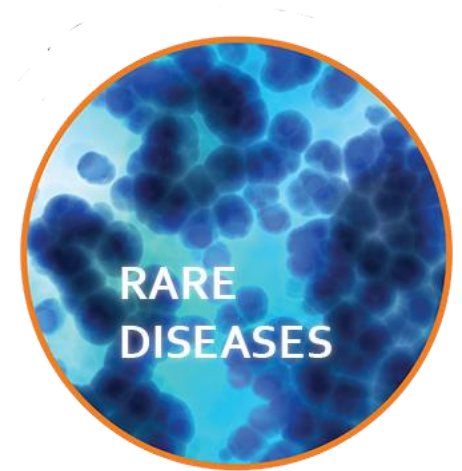
https://www.elixir-europe.org/documents/elixir-webinar-rare-disease-test-case-elixir-interoperability-backbone

Outline

- ELIXIR
  - Structure, Goals, Funding
  - Implementation
- **EGA**
  - **Details, Plans, Local EGA**
- ELIXIR Sensitive (Human) Data
  - Individual Projects
- Summary

About the EGA: Context and Metrics

- Launched 14th July 2008

- Role: **Secure** archive for **controlled** distribution of **consented** genetic & phenotypic data

- **7500+** users with access; **360+** submission accounts

- **Over 3.3PB** archived; ~**1800** datasets; over 700 studies

  - Genome data of over 100,000 people from 200 centres around the world

- **+200** contacts to Helpdesk/month

nature *International weekly journal of science*

Article

Nature **447**, 661-678 (7 June 2007) | doi:10.1038/nature05911; Received 26 March 2007; Accepted 11 May 2007

Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls

▾ The Wellcome Trust Case Control Consortium

nature *International weekly journal of science*

Review

Nature **458**, 719-724 (9 April 2009) | doi:10.1038/nature07943

The cancer genome

Michael R. Stratton[1,2], Peter J. Campbell[1,2] & P. Andrew Futreal[1]

ELIXIR pilot: EGA as a joint venture

EMBL-EBI ELIXIR Node
- Hinxton, Cambridge, UK

Spanish ELIXIR Node
- Centre for Genomic Regulation (CRG), Barcelona, Spain

www.ega-archive.org

# European Genome-phenome Archive (EGA)

- Joint service from the European Bioinformatics Institute (EMBL-EBI), UK, and the Centre for Genomic Regulation (CRG), Spain

- Secure and permanent archive for all type of genetic, -omnics and phenotypic data from humans

- ALL data consented for research use

- Data access managed through application to a Data Access Committee (DAC)

  - All data distribution is encrypted

  - User/Download specific encryption keys

www.ega-archive.org

EMBL-EBI

CRG
Centre
for Genomic
Regulation

EUROPEAN
GENOME-PHENOME
ARCHIVE

elixir

About the EGA: What is controlled access data?
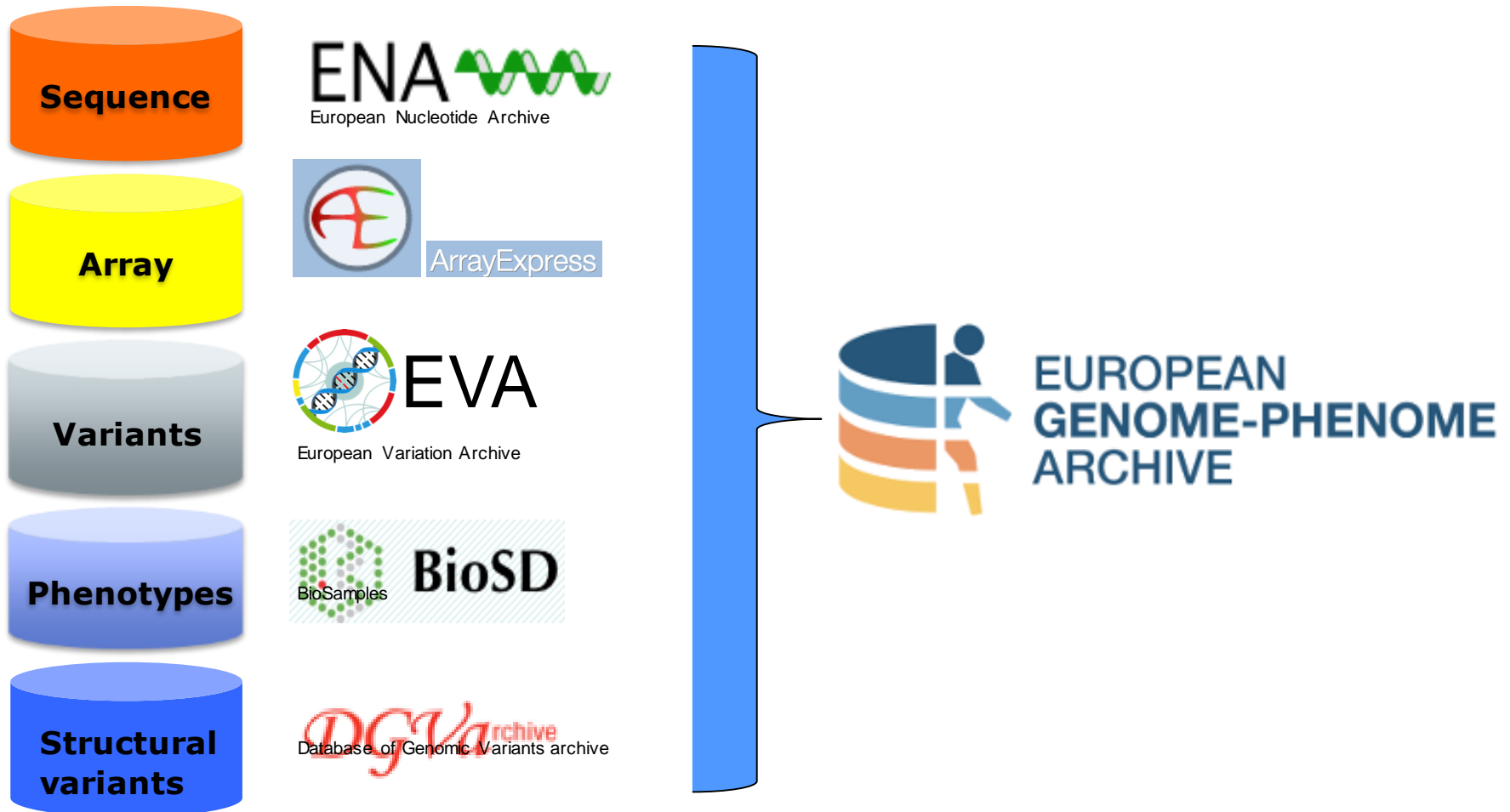
Open & public archives

Controlled access archive

Data Download Statistics

Monthly EGA Download Volume (TB)

Legend: Streaming HTTP, Streaming FTP, Staged Aspera
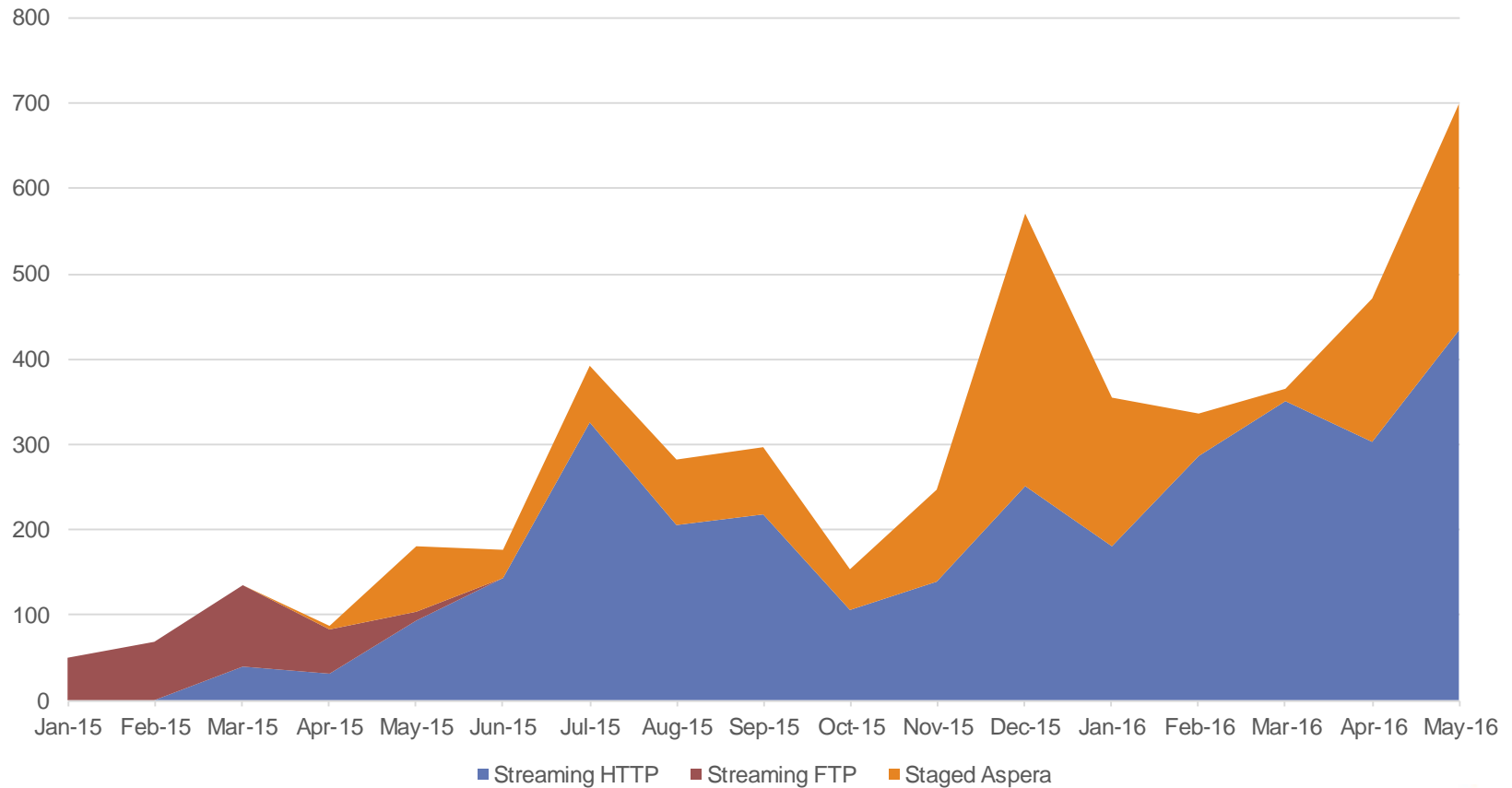
Download Locations
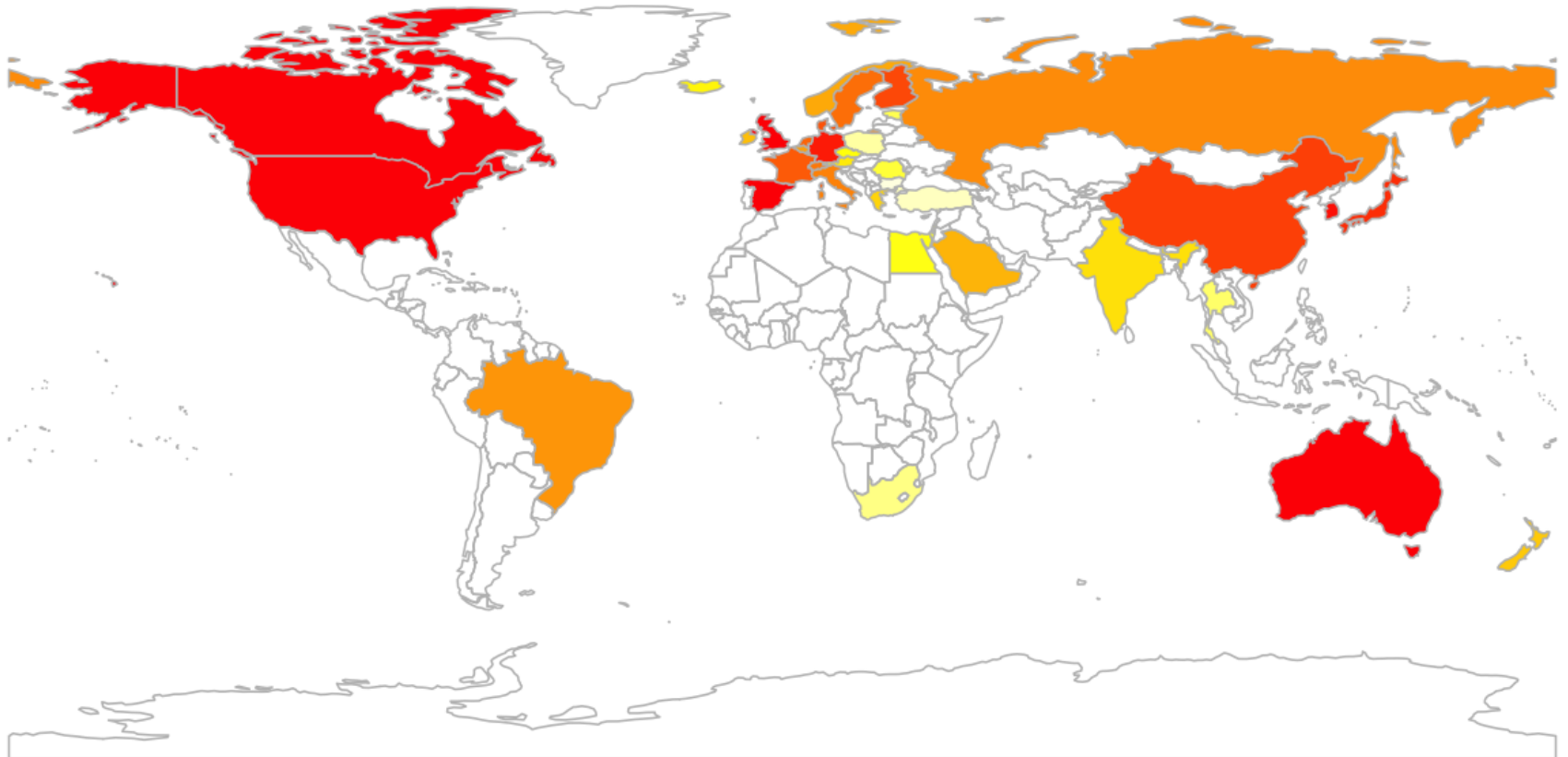
Strategic EGA Encryption Format(s)

- Submission: GPG Public Key, Node-Specific

- Archival: GPG CAST-5; AES-256

- Distribution: AES-128

- AES Encryption Allows Random-Access into encrypted Files
  - Encrypted Data can be used without need to Decrypt first
    - Encryption at Rest
- FUSE Layer (DEMO)
  - Source Code available [as part of "Local EGA"]:
  - https://github.com/elixir-europe/human-data-local-ega-fuse

Outline

- ELIXIR
  - Structure, Goals, Funding
  - Implementation
- EGA
  - Details, Plans, Local EGA
- **ELIXIR Sensitive (Human) Data**
  - Individual Projects
- Summary

EGA as an ELIXIR Service

- Strategic Integration and Expansion of EGA

- EGA Identity created during DAC approval of access application

- EGA Maintains Identity/Authorisation Database

- Link EGA Identity (EGA AAI) to ELIXIR Identity (ELIXIR AAI)
  - Subject to high Level of Assurance (LoA)
  - Enable access to EGA with ELIXIR Authentication
  - Integrate EGA into the ELIXIR Framework

  - Allows use of ELIXIR Resources
    - Allow Management of Access Permissions via ELIXIR REMS

EGA – ELIXIR Integration Use Case

- Data Access Permissions using ELIXIR REMS

  - Used by Finnish resources; integration project with EUDAT

  - ELIXIR EXCELERATE Integration with EGA

- https://rems-dev.elixir-finland.org/

- Benefits of REMS

  - Reduces **throughput times** of the application process

  - Provides easier **reporting/audit tools** for owners of the resource and the applicant

  - Increases **information security** also by relying on end users' home institutions usernames/passwords and federated authentication

EGA – ELIXIR Integration Use Case

- REMS hosted by ELIXIR Finland Node.

- AAI hosted by ELIXIR Czech Node

- EGA hosted by ELIXIR EMBL-EBI and ELIXIR Spain Nodes

- One Integrated User Experience

EMBL-EBI
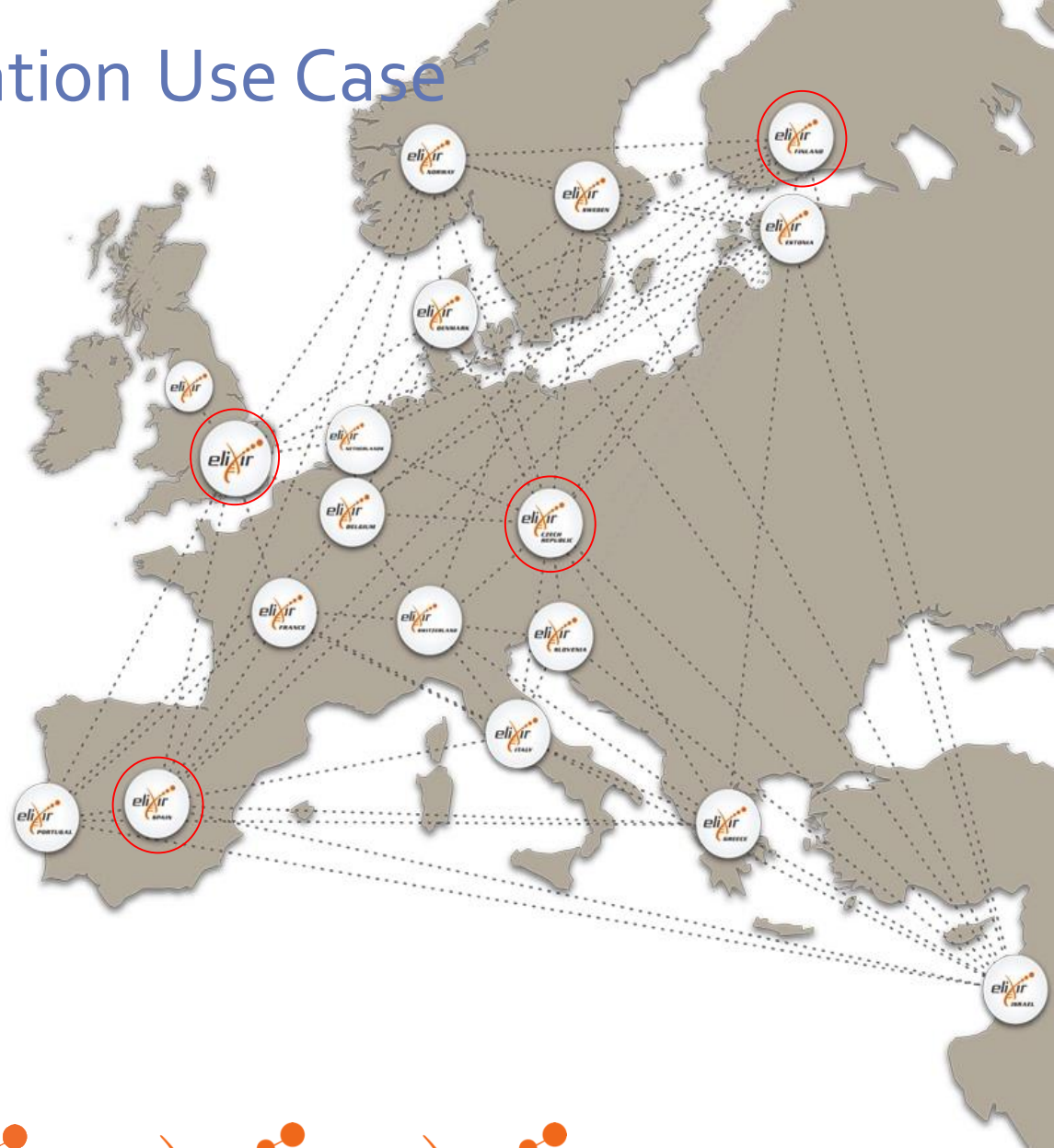
Local (Federated) EGA – "EGA-in-a-Box"

- Term: "Local EGA"

- Federate the EGA across ELIXIR Nodes
  - Handle Increased Volumes of Data (Submission & Distribution)
    - Load balancing / Resilience
  - Move Archive Location closer to Data Generation / Usage
    - Local Cloud Access / FUSE Layer
  - Be able to handle National/Institutional Data Policies
    - Data can't leave Country / Institute
  - Access Permissions Continue to be controlled by "Central EGA"

- Continue to act as a single Archive

Local EGA Demo

- Developed using Docker Containers

- Easy to deploy Demo Available

  - Creates a Demo EGA (on a VM)

  - Code Base & Instructions on GitHub (https://github.com/elixir-europe/human-data-local-ega)

  - Installation Walkthrough available as ELIXIR Webinar (https://www.elixir-europe.org/documents/elixir-webinar-local-ega-demo-setup-june-2016)
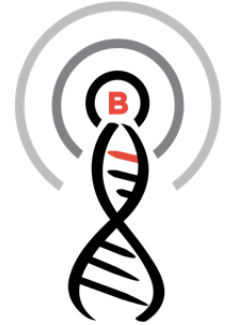
- All Source Code available on ELIXIR GitHub (https://github.com/elixir-europe)

Local EGA Demo

EGA Archive Metadata & Accounts

Access API

Download Client

Archive FUSE

Download FUSE

Access Decrypted Data after Download and Decryption

Direct Access to Archived Data

Access Encrypted Data after Download
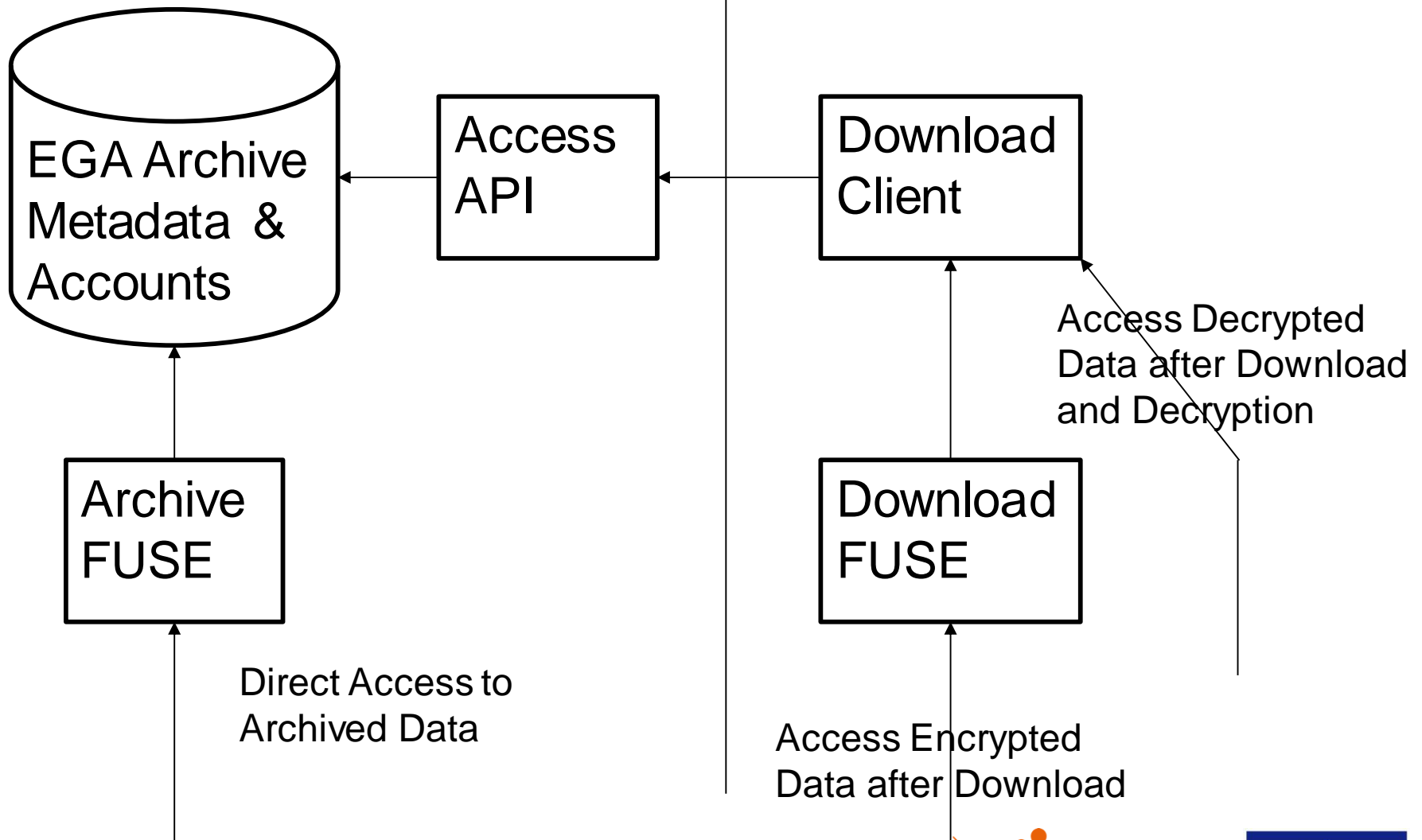
excelerate

ELIXIR Human (Sensitive) Data Pilot Projects

- ELIXIR GA4GH Beacons

- ELIXIR GA4GH Beacon APIs

- ELIXIR IMI OncoTrack

- ELIXIR TraIT

Findable - ELIXIR / GA4GH Beacon Pilot

- ELIXIR and many Node institutes member of GA4GH

- **GA4GH Beacons are a discovery service:**
  - which datasets include genomes with allele of interest

- ELIXIR pilot project with partners from the Netherlands, Sweden, Finland, France and Spain

- Complies to the GA4GH Beacon project standard and security working group policies

- **ELIXIR EGA Integration - Three objectives:**
  - Provide ELIXIR reference implementation on GA4GH Beacon **with 3 authorization levels**
  - Provide ELIXIR standards for data types
  - Provide an example on capacity and expertise build across ELIXIR Nodes to integrate national resourc service interface

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

elixir

Leverage data nodes network

- Beginning Beacon implementation in ELIXIR Nodes

- ELIXIR Nodes link to national policies and infrastructure

- ELIXIR Nodes provides a national framework for long-term resource management.
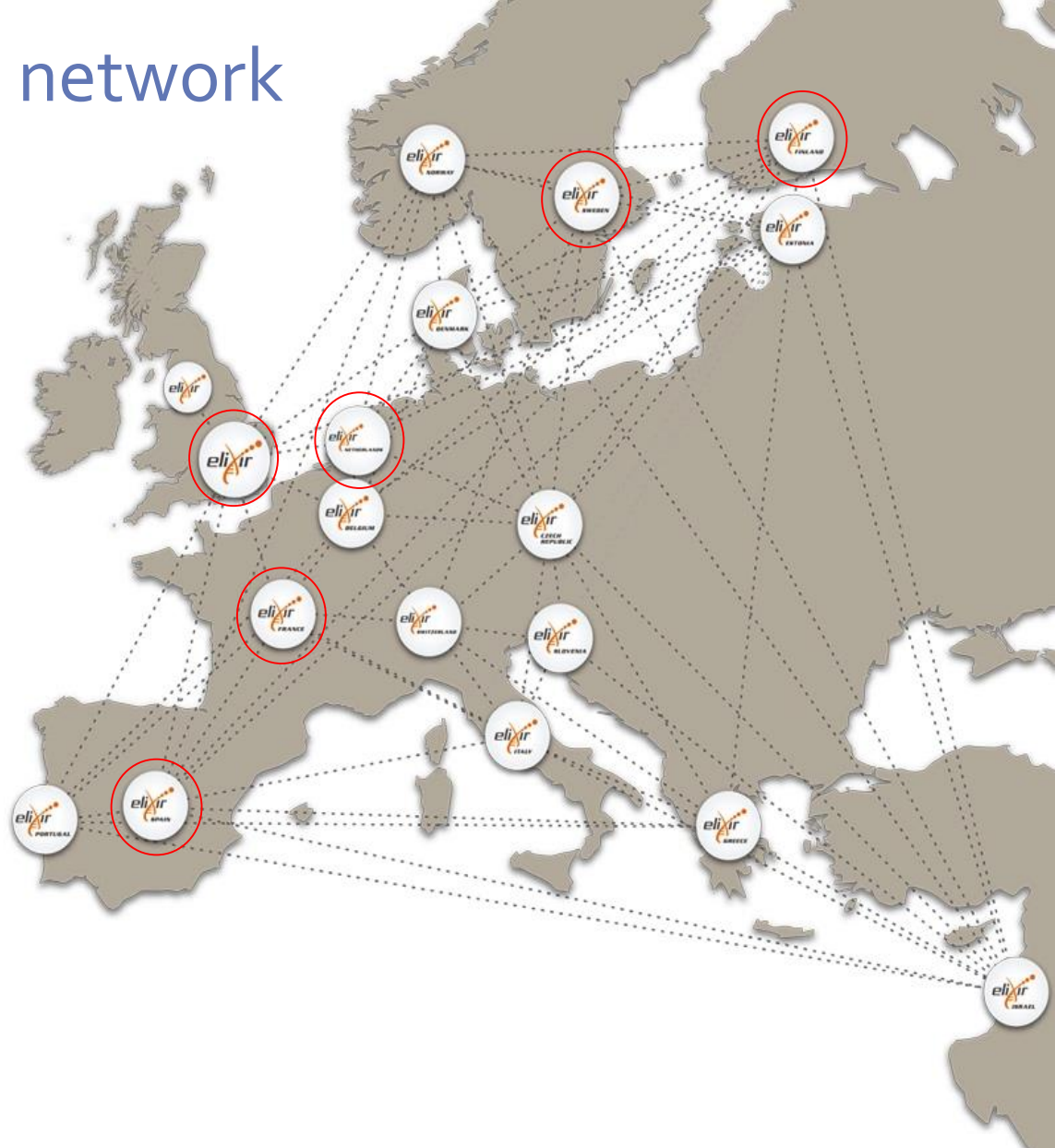
Beacon Demo

https://ega.crg.eu/beacon_web/#/

- Discoverability
- New Access Tier: "Registered"

Source Code is Available:

https://github.com/elixirhub/human-data-beacon

[Demo Screen]

EGA Beacon service

🔒 ega.crg.eu/beacon_web/#/

Note: The EGA archives a large number of datasets, some of which are `Publicly` available. If you have an account on this website, you can access to `Registered` datasets. To access `Controlled` datasets can be done by contacting the relevant Data Access Committee (DAC), whose details are displayed on the Dataset description page under "Who controls access to this dataset" (click on the Dataset ID to go to the Dataset page). Once you have access to a dataset, you will be able to query it at EGA Beacon.

EGA ID ⌄	Short title ⌄	Access type ⌄
EGAD00001000433	This sample set comprises cases of schizophrenia with additional cognitive measurements, collected in …	CONTROLLED
EGAD00001000614	This sample set of UK origin consists of clinically identified subjects with Autism Spectrum Disorders, mo…	CONTROLLED
EGAD00001000443	The sample selection consists of subjects with schizophrenia (SZ), autism, or other psychoses all with m…	CONTROLLED
EGAD00001000740	Low-coverage whole genome sequencing; variant calling, genotype calling and phasing	PUBLIC
EGAD00001000613	The MGAS (Molecular Genetics of Autism Study) samples are from a clinical sample seen by specialists …	CONTROLLED
EGAD00001000430	Two groups of samples with diagnosis of schizophrenia or schizoaffective disorder in the UK: cases with …	CONTROLLED
EGAD00001000434	The BioNED (Biomarkers for Childhood onset neuropsychiatric disorders) study has been carrying out d…	CONTROLLED
EGAD00001000437	The Tampere Autism sample set consists of samples from Finnish subjects with ASD (autism spectrum d…	CONTROLLED
EGAD00001000439	The entire sample collection consists of 2756 individuals from 458 families of whom 931 are diagnosed …	CONTROLLED
EGAD00001000442	Samples from three sources: the Genetics and Psychosis (GAP) set consists of samples from subjects w…	CONTROLLED
EGAD00000000028	Procardis study for coronary artery disease. GWAS study. 3352 cases - 3145 control	REGISTERED
EGAD00010000300	Summary statistics from Haemgen RBC GWAS (Anemia)	REGISTERED
EGAD00000000029	Aggregate results from a case-control study on stroke and ischemic stroke. 19602 samples.	REGISTERED
EGAD00000000115	WNT-signaling and Dupuytren's Disease. GWAS analysis. 856 cases - 2836 control.	CONTROLLED
EGAD00001000435	These samples are a subset of a nationwide collection of Finnish autism spectrum disorder (ASD) sampl…	CONTROLLED

ELIXIR – IMI OncoTrack

- Academic – Industry Translational research project
  - Goal of identifying biological markers that will help our understanding of tumors

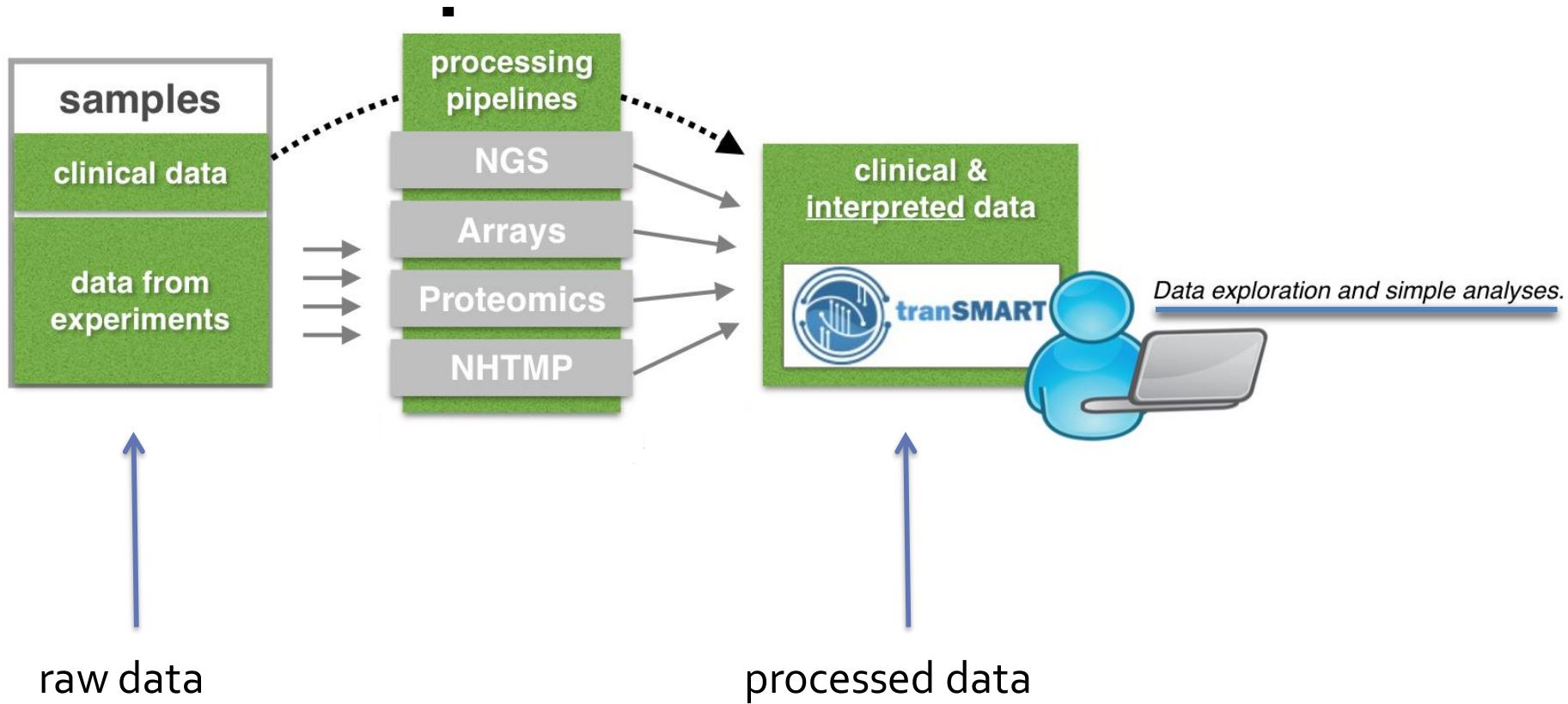- Scoping study to begin understand long-term knowledge management requirements
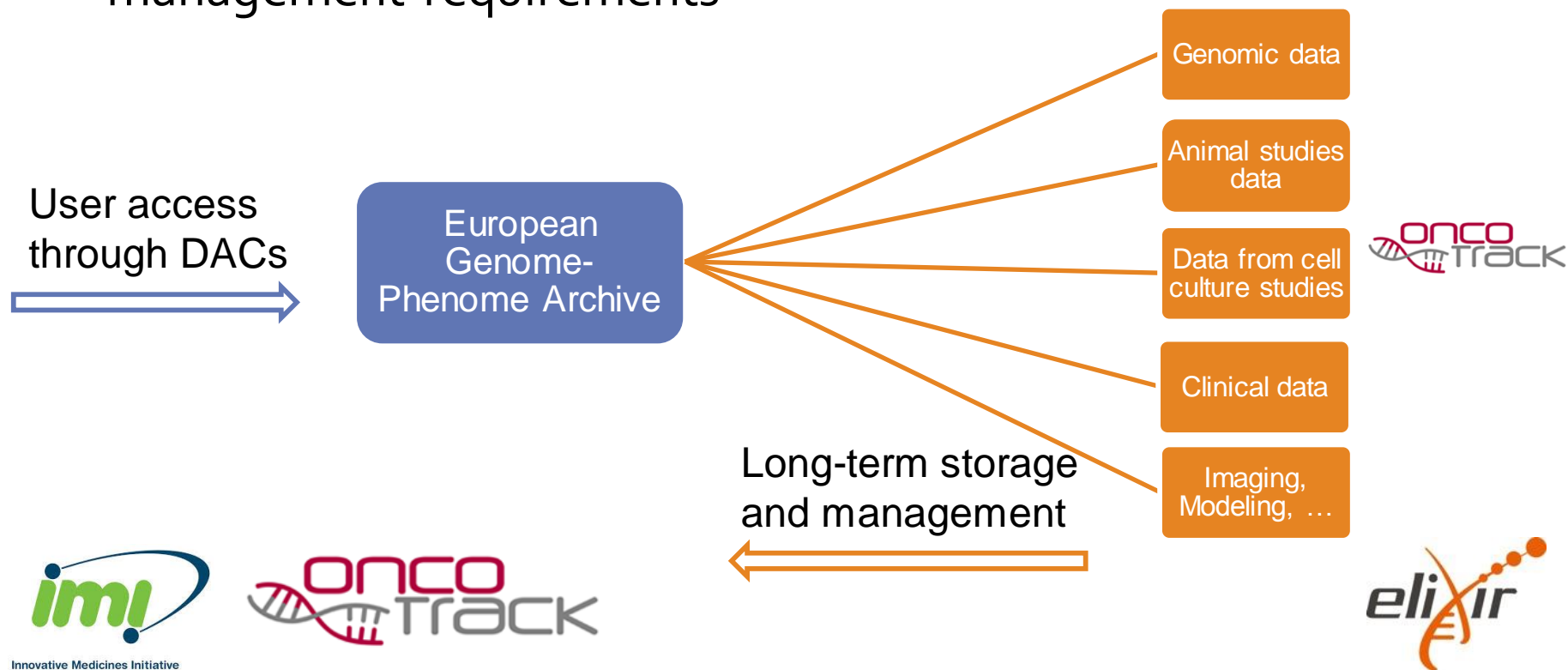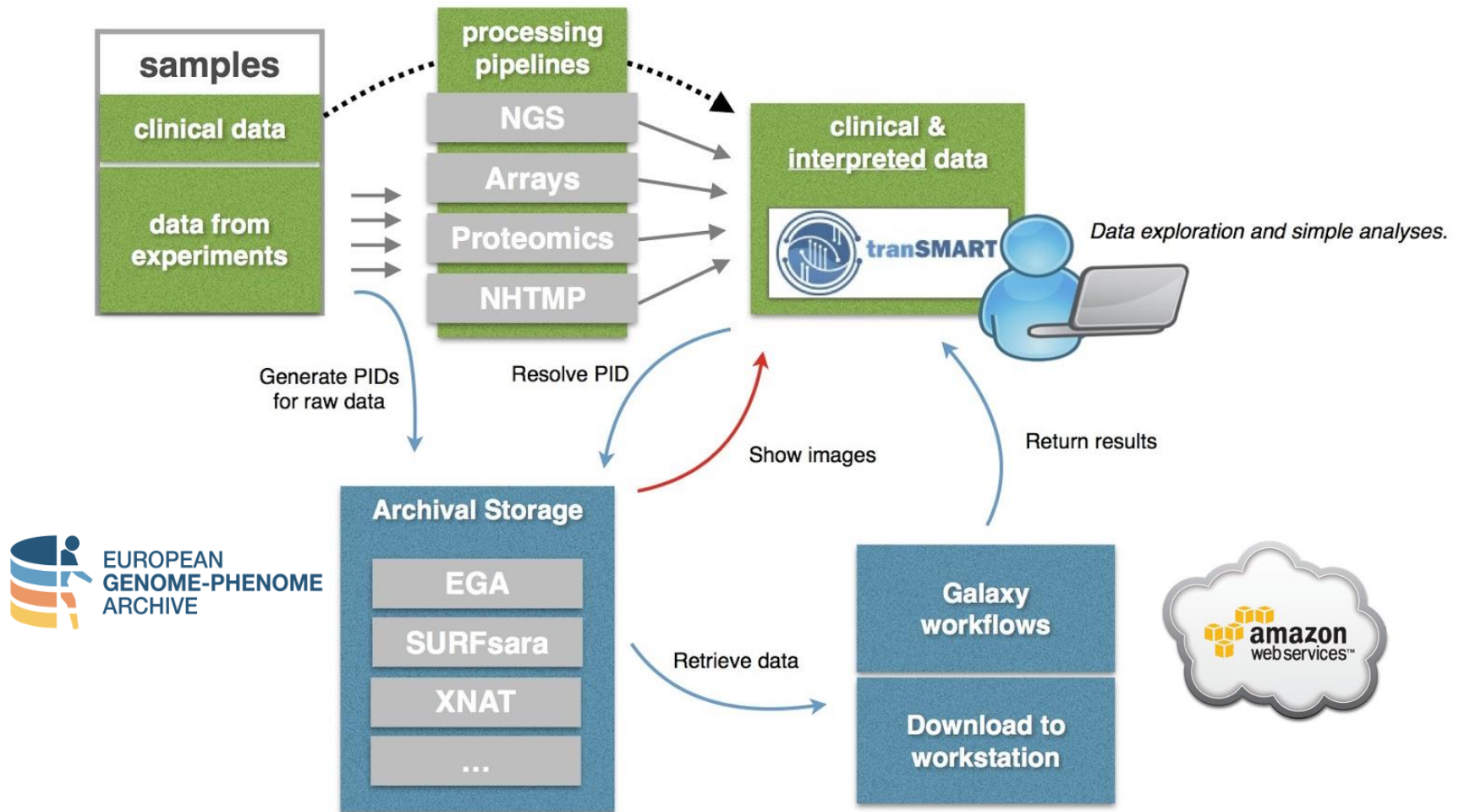
ELIXIR – TraIT

- Aim: To develop and demonstrate integration of

1. Processed, Clinical and Translational data resources
   a. Increase integrated data usage and discovery
2. Large scale data archived at EGA
   a. EGA as a sustainable long term storage solution
3. Secure data analysis workflow in a Galaxy framework
   a. Investigate cloud / platform integration
4. Ensure compatibility with sample BioBanks

Molecular profiling: work flow

Molecular profiling: **raw data**

Elixir TraIT: Re-Usable End-to-end exploration of data integration

Source: https://wiki.galaxyproject.org/Events/GCC2015?action=AttachFile&do=get&target=S1_T2_Hoogstrate_TraIT.pdf
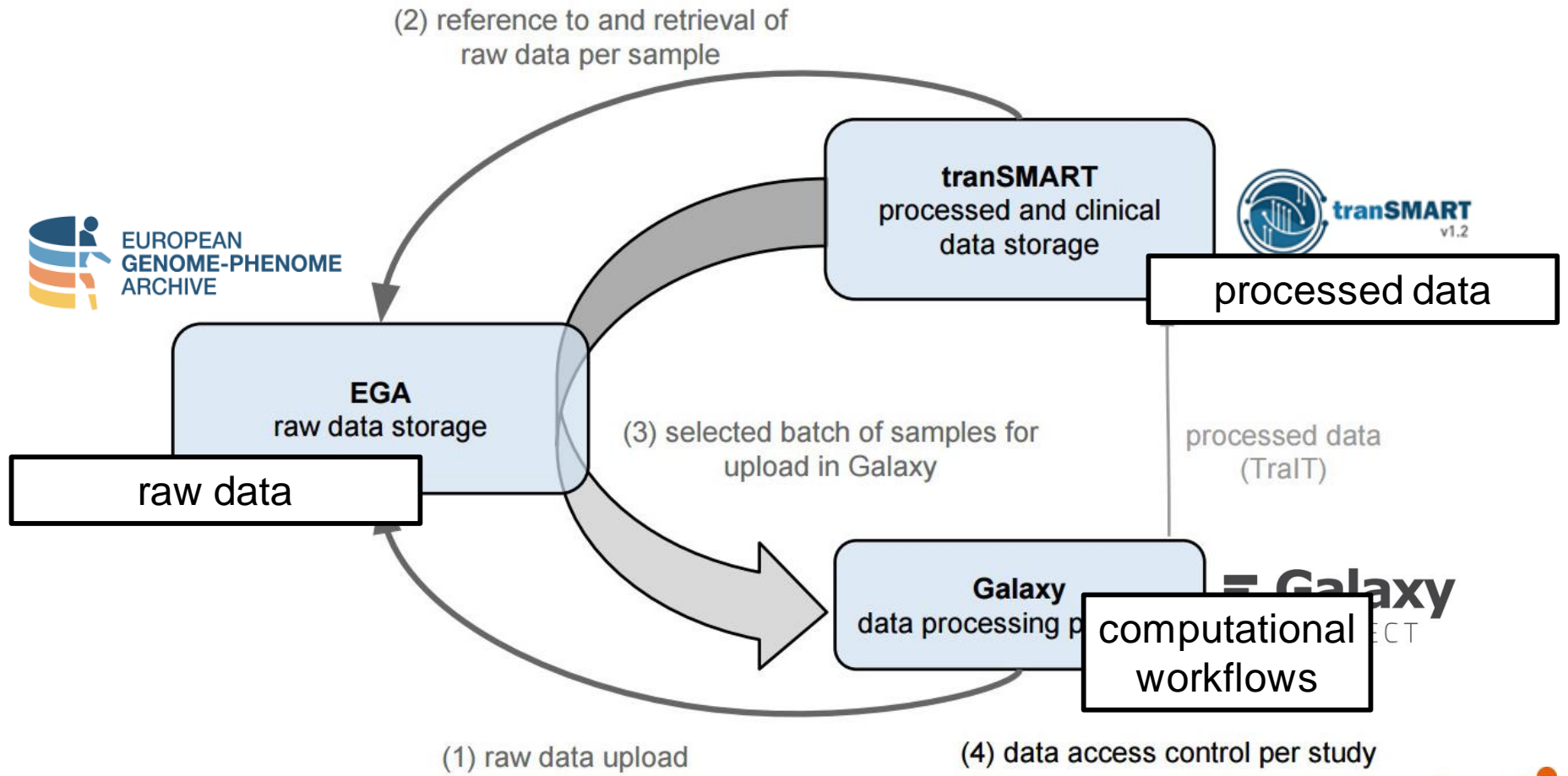
Outlook: Strategy & Features

- Access to EGA Data from ELIXIR

    - Enhance Discoverability

    - Standards Compliance (API & Metadata)

    - Broaden Scope of Archived Data

    - Data Access Agreements

- Broaden Archive Locations

    - Local EGA

- Provide More Fine-Grained Access to EGA Data

    - Random-Access to Local Data via FUSE

    - Random-Access to Remote Data via API

- Cloud Access

Outline

- ELIXIR
  - Structure, Goals, Funding
  - Implementation
- EGA
  - Details, Plans, Local EGA
- ELIXIR Sensitive (Human) Data
  - Individual Projects
- **Summary**

Summary

- ELIXIR: European Federated Framework of Services

- EGA: ELIXIR's Sensitive Data Strategy

  - FUSE Layer enables Encryption at Rest & Direct Archive Access

  - REMS extends EGA into ELIXIR framework

  - Local EGA extends Secure Archive across Elixir Nodes and Institutes

  - GA4GH Beacon improves Data Discoverability and engages with Global Community for Standards and Formats

  - OncoTrack explores extending EGA Data Model an Industry cooperation

  - TraIT explores EGA as Raw Data storage in Live Analysis Pipeline

Outline

- ELIXIR
  - Structure, Goals, Funding
  - Implementation

- EGA
  - Details, Plans, Local EGA

- ELIXIR Sensitive (Human) Data
  - Individual Projects

- Summary

ELIXIR Gateway on F1000Research

- Online platform to capture ELIXIR's publication output, research articles transparently peer-reviewed though F1000R

- ELIXIR Editorial board:

  - Chair: Niklas Blomberg, Members - ELIXIR Heads of Nodes: Arlindo Oliveira, Barend Mons, Inge Jonassen, Bengt Persson

- Latest articles:

  - Moreews F et al. *BioShaDock: a community driven bioinformatics shared Docker-based tools registry*

  - Eijssen et al*. The Dutch Techcentre for Life Sciences: Enabling data-intensive life science research in the Netherlands*

  - Bousfield et al. *Patterns of database citation in articles and patents indicate long-term scientific and industry value of biological data resources*

Papers: http://f1000research.com/channels/elixir-reports
Posters and slides: http://f1000research.com/channels/elixir/elixir-posters

Keep in touch ELIXIR Newsletter

Sign up:

https://www.elixir-europe.org/news

ELIXIR webinars

- Regular series of updates from ELIXIR

- Typically 3rd Wednesday of month

- 30 minute talk: 30 min questions

- Examples webinars:

  - ELIXIR registry - Kristoffer Rapacki

  - Data and Software carpentry - Aleksandra Pawlik

  - Marine metagenimics: towards user centric services – Nils Peder Willassen

- Run using WebEx - for live or recorded viewing

- Full programme: www.elixir-europe.org/webinar

THANK YOU!

Alexander Senf, PhD -- asenf@ebi.ac.uk

www.elixir-europe.org

@ELIXIREurope

/company/elixir-europe

ELIXIR and European Union

- ELIXIR Hub uses the legal model of EMBL – it is an international treaty organisation, separate from the EU.

- Result of the UK referendum on the EU:

    - Will not affect ELIXIR Hub operations

    - Will not affect UK or EMBL-EBI membership in ELIXIR

    - Will not affect ELIXIR's access to EU funding

    - Is not expected to affect ELIXIR UK node in the short term