



BBMRI-ERIC

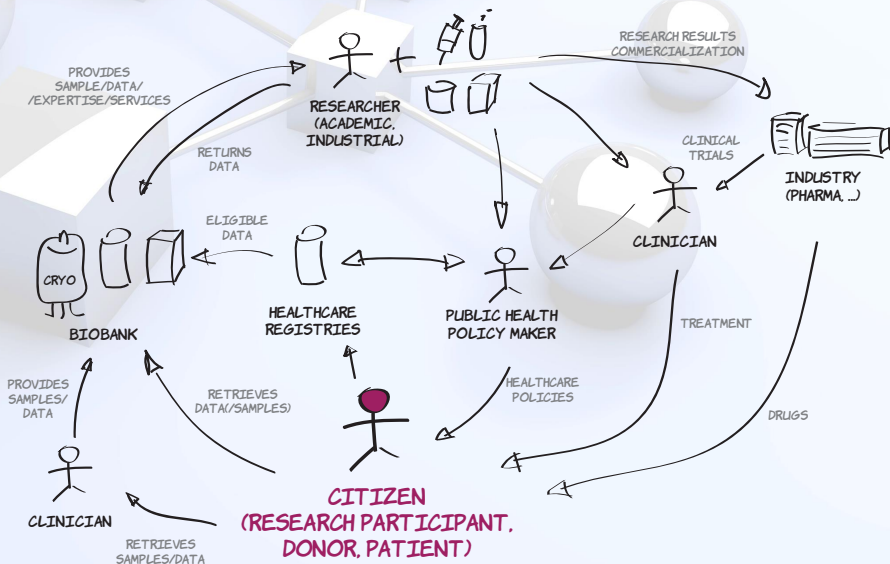
Biobanking and
BioMolecular resources
Research Infrastructure

BBMRI-ERIC and AAI

Assoc. Prof. RNDr. Petr Holub, Ph.D.
IT & Data Protection Manager @ BBMRI-ERIC,
CIO of BBMRI-ERIC CS IT

CORBEL and AARC/AARC2 AAI Workshop,
Paris, 2016-05-31

What is BBMRI-ERIC



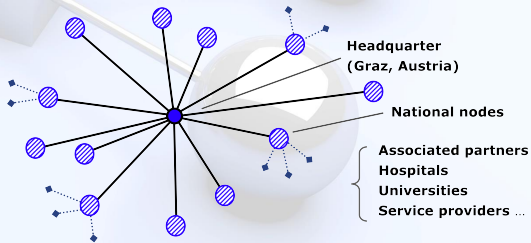
What is BBMRI-ERIC

- ▶ An infrastructure that provides/facilitates secure and privacy-protecting access to key resources in order to support biomedical research and to support healthcare advancement:
 - **biosamples** from biobanks,
 - related **data**: clinical, omics, phenotypes, etc.,
 - **expertise** and other **services** (e.g., sample & data hosting),
 - biomolecular resources.

biobanks := samples + data + expertise + services;

What is BBMRI-ERIC

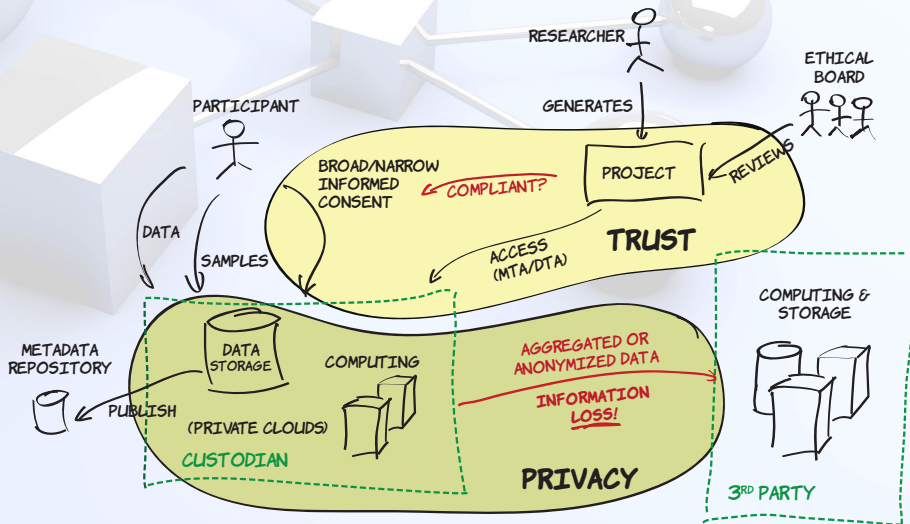
- ▶ Hierarchical distributed architecture ≡ “hub-and-spokes architecture”



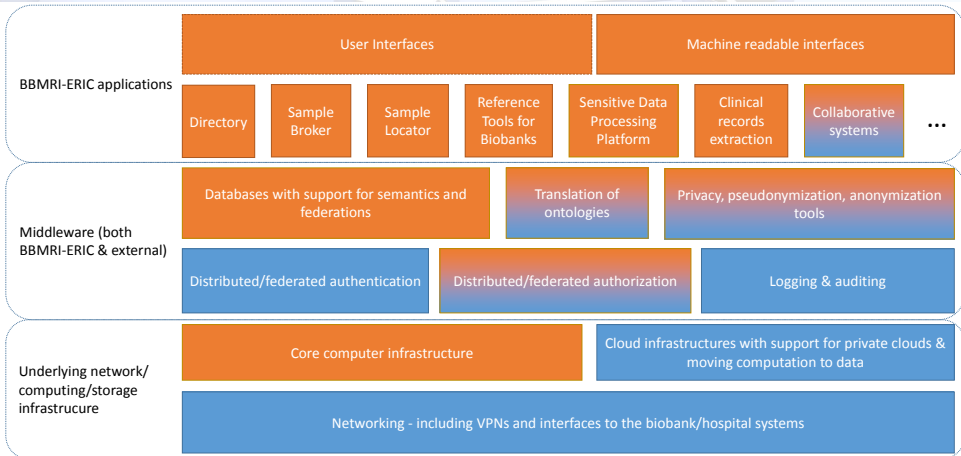
⇒ federated IT architecture

- ▶ Subject to regulatory frameworks: privacy-protection, health,...
 - e.g., upcoming General Data Protection Regulation.

IT Architecture of BBMRI-ERIC

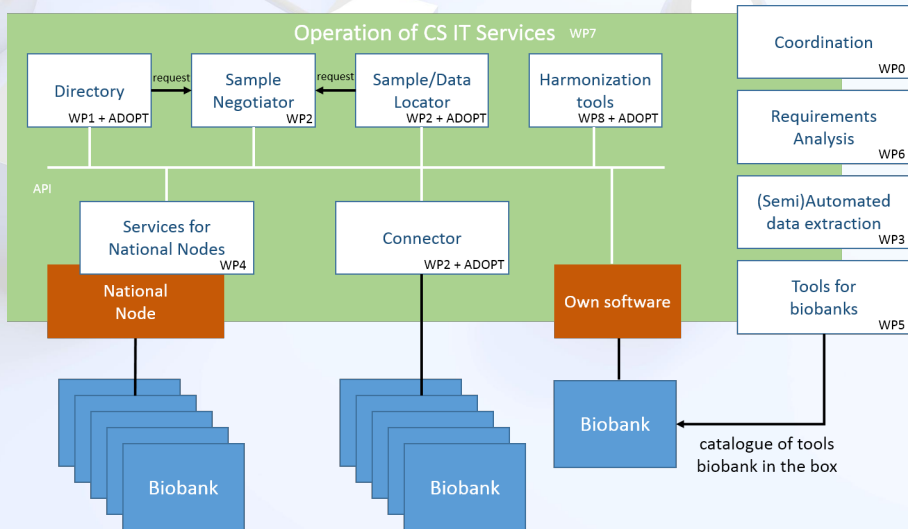


IT Architecture of BBMRI-ERIC

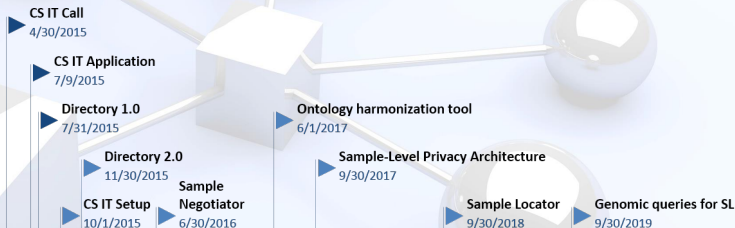


orange ...BBMRI-ERIC own components; blue ...components expected from other infrastructures.

- ▶ Most of the IT services should be implemented via BBMRI-ERIC Common Service IT
 - formal way to organize the member countries contributing to the IT,
 - official start November 1, 2015 (effective January 1, 2016),
 - ADOPT BBMRI-ERIC acts as booster to CS IT core budget,
 - acts as coherent development ecosystem:
 - consistent set of tools implementing the whole workflow of BBMRI-ERIC IT services,
 - (running Scrum of scrums together :-)).



BBMRI-ERIC CS IT



2015 | 2015 | 2016 | 2017 | 2018 | 2019 | 2020 | **2020**

Today



Status of AAI (1)

- ▶ What do we need from AAI:
- ▶ Authentication:
 - identity verification (vetting)
 - LoA 1–2 depending on service
 - authentication instances
 - LoA 1–3 depending on service
 - federated architecture
- ▶ Authorization
 - matching of informed consent & project as a part of initial authorization decision
 - person+project identity for authorization decisions
- ▶ EGI-Engage M6.2: BBMRI-ERIC Security & Privacy Requirements

Status of AAI (2)

► Summary of minimum requirements:

Table 8: Minimum requirements for basic data types. Non-personal data is used to denote data the does not contain any traces of privacy-sensitive data (e.g., data about operation of the biobank storage systems).

	raw (non-deidentified)	pseudonymous	practically anonymous	non-personal
<i>Authentication and authorization</i>				
Identity verification	LoA ≥ 2	LoA ≥ 2	LoA ≥ 0	open
Authentication instance	LoA ≥ 3	LoA ≥ 2	LoA ≥ 0	open
Assessing project & informed consent compliance	not available for research	MANDATORY	RECOMMENDED	–
Restricted access	high security	high security	medium-low security	open
DTA/MTA	REQUIRED	REQUIRED	RECOMMENDED	open
<i>Authentication and authorization</i>				
Access log archive since last access	≥ 10 years	≥ 10 years	≥ 3 years	–
<i>Data transfers and storage</i>				
Encrypted storage	REQUIRED	REQUIRED		
Encrypted transfers	REQUIRED	REQUIRED		

Status of AAI (3)

- ▶ Current plan for implementation:
 - hookup BBMRI-ERIC into eduGAIN – done
 - pilot *per se* – international organization headquartered in one country
 - develop BBMRI-ERIC Identity
 - piloting withing AARC and GÉANT VOPaaS
 - implemented by a Proxy IdP with various backends
 - identity linking/merging
 - use of BBMRI-ERIC National Nodes for registration of “homeless” or “effectively homeless” (insufficient LoA @ home) users
 - become one of pilot applications for AARC2
 - close collaboration within CORBEL WP5 – Access
 - collect, analyze, implement needs of BMS infrastructures participating in CORBEL
 - collaboration with ELIXIR

Status of AAI (4)

- watch closely for STORK successor(s)
 - government-backed identity verification (vetting) is important feature
 - let's hope for eIDAS

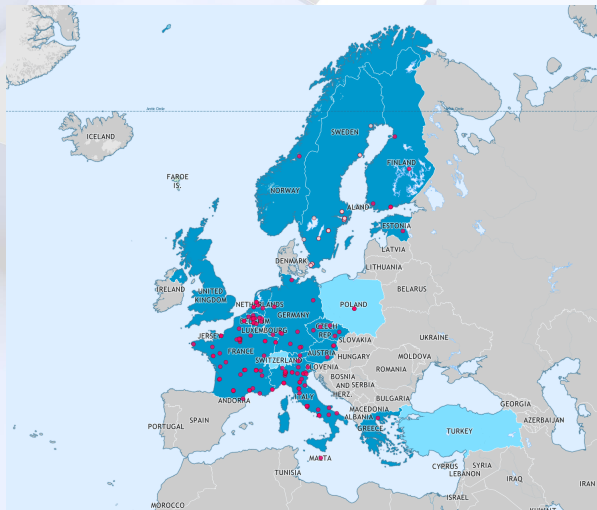
Status of AAI (5)

- ▶ REMS (BBMRI/ELIXIR FI)
 - partial support for sample/data access negotiation,
 - experiences with pilot deployment in THL Biobank (FI),
 - explored now as a part of initial work on BBMRI-ERIC Negotiator.
- ▶ Organizational aspects – interdependencies of infrastructures
 - SLA/SLD problem: what if an infra (to be used) is not dependable enough for the other?
 - what if an infra changes its policy?
 - what if there is an issue of conflicting business models?
- e.g., services are free only for members: and the members of two infras are not 100% identical – this may happen even during runtime

Challenges for AAI

- ▶ Consistent information about LoA (identity verification | auth instance) from federated AAI
 - legal review of LoAs – have to withstand hearing at court
- ▶ People ↔ country mapping
 - members of ERICs are countries
- ▶ Dealing with less than 100% geographical infrastructure overlap
 - especially if money transfers are expected by either/both infrastructures, and the consumer infrastructure is expected to provide services for their members for free

Challenges for AAI



Map based on the BBMRI-ERIC Directory 2.0 as of Dec 23, 2015.

► Full members:

- Austria
- Belgium
- Czech Republic
- Estonia
- Finland
- France
- Germany
- Greece
- Italy
- Malta
- Netherlands
- Norway
- Sweden
- United Kingdom

► Observers:

- Poland
- Switzerland
- Turkey
- IARC

Challenges for AAI

- ▶ Collaboration with industry
 - BBMRI-ERIC infrastructure has industrial users: commercial research brings drugs to market
- ▶ Dealing with “homeless users”
 - big institutions (e.g., Pfizer :)) will not deploy full-scale IdP just because of a few users procuring samples/data
- ▶ Collaboration with non-European countries
 - collaboration with Asian countries
 - collaboration with Africa (e.g., B3AFRICA project)
- ▶ Affiliation of people to projects
 - issue of bootstrapping a project in trustworthy way
- ▶ IdP ↔ SP attribute access negotiation simplification

Challenges for AAI

- ▶ Flexibility of AAI to react to changing needs of users
 - infrastructure (customer) induced
 - induced by regulatory frameworks

Time Line for AAI

06/2016 **BBMRI-ERIC Identity with LoA 2/2 for Negotiator**

- ▶ ongoing implementation with AARC/VOPaaS

09/2016 **Security toolset release for BBMRI-ERIC (EGI-Engage D6.11)**


- ▶ includes working AAI
- ▶ integration of federated AAI into BiobankCloud

06/2017 **BBMRI-ERIC Identity with LoA 3/2 for Locator**

- ▶ ideally with eIDAS backend, if available at that time

Privacy & Security Requirements of BBMRI-ERIC IT services

- ▶ Initial version of privacy & Security requirements published as EGI-Engage Milestone M6.2 document:
<https://documents.egi.eu/document/2677>
 - requirements are expected to be kept updated as our understanding evolves, regulatory frameworks are updated, and technologies are becoming available
 - expected update: October 2016 – part of Security & Privacy Architecture



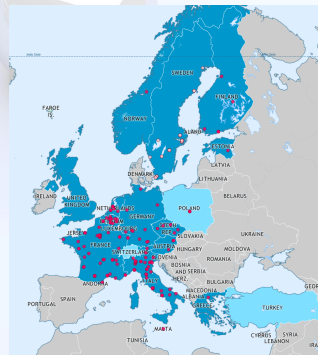
Thank you for your attention!
Q?/A!

<http://www.bbmri-eric.eu/>
petr.holub@bbmri-eric.eu

What is BBMRI-ERIC

- ▶ European Research Infrastructure Consortium to facilitate access to high-quality biobanks and biomolecular resources
 - legal entity on European level,
 - est. 3 December 2013.

BBMRI-ERIC is today the largest health-oriented ERIC ever launched in Europe.



IT Architecture of BBMRI-ERIC

- ▶ Modular architecture with components interconnected by well defined interfaces
 - replaceable and reusable, well-defined (small) components,
 - standardized in ideal case, well-defined at least
 - this is critical as some components may need to be implemented by the commercial companies (e.g., components of hospital information systems).



Collaboration with ELIXIR

- ▶ AAI
 - collection of needs of all the BMS infras
 - using CORBEL WP5 as a framework
 - pilots to AARC2 with ELIXIR
- ▶ Harmonization of ontologies
 - focus on BBMRI-ERIC on biobank-related ontologies: phenotyping, clinical, biobanks, ...
 - using CORBEL WP6 as a framework
- ▶ Software development best practices
- ▶ GA4GH-ELIXIR Beacons

Examples of BBMRI-ERIC Use Cases

► Aggregate view of the infrastructure

- Q – bio/med researcher: **“What biobank could have samples relevant for my research?”**
- Q – bio/med researcher: **“What biobank is capable of hosting my samples?”**
- Q – biobanker: **“What biobanks are similar to ours?”**

⇒ **BBMRI-ERIC Directory**

- currently in non-public beta version, covering more than 500 biorepositories,
 - currently largest repository contains **30,000,000+** samples,
 - includes even a few smaller non-human sample collections (but health focus),
- Directory 2.0 – released December 2015

Examples of BBMRI-ERIC Use Cases

- ▶ Facilitate access to the samples and data
 - Q: “**I need n samples with ... specifications**”
 - researchers do not know what exactly they need
 - in terms of the material type and sample quality for given experiment
 - multi-round negotiation between researchers and biobankers (resource providers in general)
 - ... while having hundreds or thousands of biobanks
 - biobankers are overloaded with fuzzy requests
 - biobankers are willing to release samples only for certain purposes
 - Q: “I would like to have these 20 samples from this great cohort of **100,000** participants, please.”
 - A: “**NO!!!**”

⇒ **BBMRI-ERIC Sample/Data Negotiator**

Examples of BBMRI-ERIC Use Cases

- ▶ Access to sample-level information: browsing, searching
 - Q: “I need to see what sample types are available in my research field in order to develop new research projects.”
 - BBMRI-ERIC is committed to ensuring privacy
 - differential privacy approach
 - famous attacks on privacy: attack on Massachusetts Group Insurance Commission by dr. Sweeney, attack on Netflix user DB by Narayanan and Shmatikov
 - k -anonymity: each record is undistinguishable from at least $k - 1$ other records \implies dimensionality curse,¹ datasets are sparse in reality
 - k -anonymity can still leak information $\implies l$ -diversity, t -closeness

¹ AGGARWAL, Charu C. On k -anonymity and the curse of dimensionality. In: *Proceedings of the 31st international conference on Very large data bases*. VLDB Endowment, 2005. p. 901-909.

Examples of BBMRI-ERIC Use Cases

- ▶ Access to sample-level information: browsing, searching
 - Q: **“I need to see what sample types are available in my research field in order to develop new research projects.”**
 - disclosure filters
 - not only privacy protection,
 - also protection of resources based on biobankers' policies,
 - specific support needed for rare diseases
 - amplified problem of patient identification,
 - need for cross-biobank patient identification.

⇒ **BBMRI-ERIC Sample/Data Locator**

Examples of BBMRI-ERIC Use Cases

- ▶ Access to data only

- Q – bioinformatics: **“I need access to the clinical/omics data for my research.”**

⇒ **BBMRI-ERIC Sample/Data Locator**

⇒ **BBMRI-ERIC Platform for Sensitive Data Processing**

- BiobankCloud, Mosler/TSD, etc.

- ▶ Measuring impact of bioresources

- Q - biobanker, funding organizations: **“We need to know the impact of a bioresource.”**

⇒ **BRIF** now adopted by BBMRI-ERIC

- BioResource Impact Factor

Principal Components

- ▶ **BBMRI-ERIC Directory**
 - aggregate information about available resources: biobanks & collections,
 - even achieving agreement on such minimum data structure has not been simple :) – ongoing updates to MIABIS 2.0 standard,
 - beta version of BBMRI-ERIC Directory already used by pilot users as of May 2015.
- ▶ **BBMRI-ERIC Sample/Data Negotiator**
 - brokering of samples between researchers and biobankers,
 - efficient $M : N$ communication tool for large M and N .
- ▶ **BBMRI-ERIC Sample/Data Locator**
 - federalized architecture with distributed queries,
 - privacy and security by design to avoid vulnerability to privacy attacks.

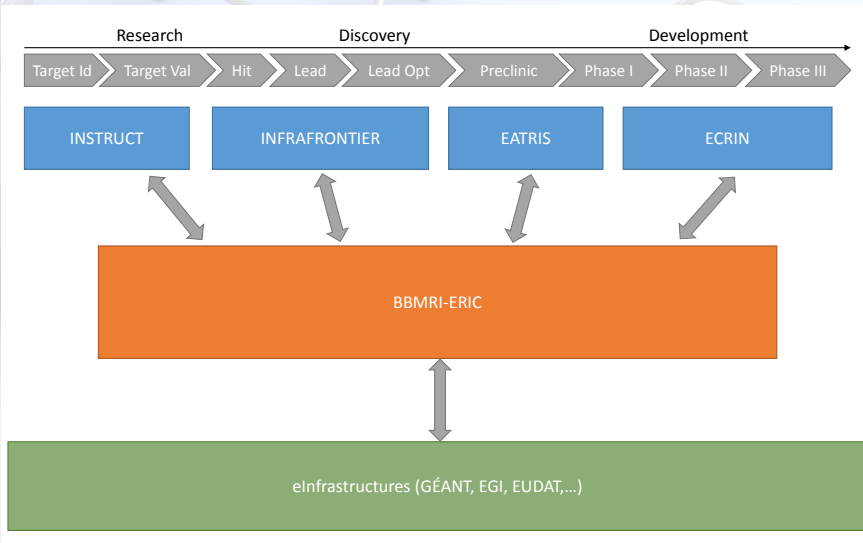
Principal Components

- ▶ Tools to support national-level and local-level infrastructures
 - reference tools for biobanks and national nodes to connect to the European infrastructure,
 - registry of BBMRI-ERIC endorsed tools.
- ▶ Data harmonization service + metadata registries
 - ontologies registry, translation/harmonization recipes.

Principal Components

- ▶ Extraction of structured data from unstructured clinical records
 - this is one of the major problems which limits performance of biobanks at the moment,
 - involves complex natural-language processing and machine learning,
 - *language and region specifics* \implies generating data in different ontologies and different structures
 - accompanying data often comes from health care systems.

Collaboration with Other Infrastructures



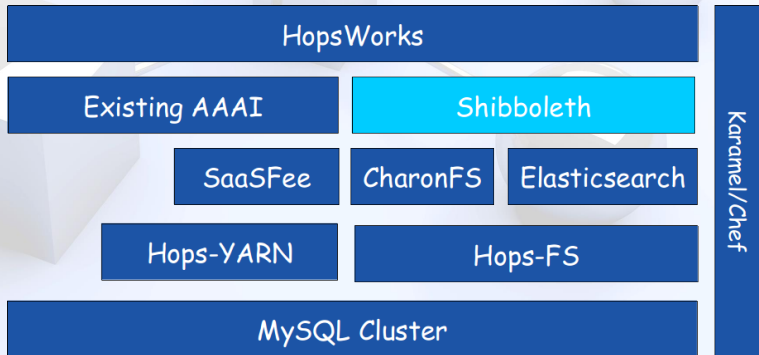
Status of Cloud Computing (1)

Default is private clouds in biobanks: but can we go beyond that?

- ▶ Private clouds piloted by BiobankCloud
 - focus on solving multi-tenancy problem (person+project)
 - prototyped with Apache jclouds® interfaces
 - support for distributed encryption to store data beyond biobanks

Status of Cloud Computing (2)

► BiobankCloud architecture



Status of Cloud Computing (3)

► BiobankCloud architecture

Karamel API (YAML)

Chef

VMIs

JClouds
Create VMs

BitTorrent
Reduce Install Times

ssh

AWS

GCE

EGI

OpenStack

Bare Metal

Status of Cloud Computing (4)

- ▶ **BBMRI Competence Center in EGI-Engage**
 - basic scenario: private cloud based on EGI-Engage platform for BiobankCloud processing genomics data,
 - later phase: explore what is possible beyond that.
- ▶ **Trusted/secure data sharing platforms**
 - collaboration with TSD, MOSLER/TSD 2.0, and others,
 - known to work in some legislative frameworks (e.g., Nordic countries).

Status of Cloud Computing (5)

- ▶ Use of 3rd party providers
 - extending notion of “private clouds” to ingest contracted clouds: under what conditions?
 - impact of GDPR – responsibility is now both with data owner and data processor
 - what is the impact in various legal frameworks – GDPR actually does not harmonize it
 - what level of certification will be required if acceptable at all?
 - now looking into ISO 27001/27018 certifications
 - exploring also as a part of PhenoMeNal
 - input for European Open Science Cloud
 - ... if it will also become a cloud in technical sense :)

Time Line for Clouds

09/2016 **Security toolset release for BBMRI-ERIC**
(EGI-Engage D6.11)

- ▶ integration of federated AAI into BiobankCloud

08/2017 **Evaluated cloud environment and demonstrator of analysis workflow for biobank studies**

- ▶ demonstrator
- ▶ minimum: private cloud using EGI cloud stack inside BBMRI.{cz,nl,se} biobanks

A Few Further Notes on Clouds

- ▶ And some more general notes... *if* there is future cloud marketplace (e.g., as a part of EOSC)
 - research institutions must balance CAPEX/OPEX,
 - research institutions or downstream research infrastructures must be given access to plurality of services (incl. brokering services),
 - cloud brokering/marketplace initiatives need to
 - remain neutral and lightweight,
 - be non-competing with upstream providers and downstream users,
 - be standard-compliant and thus also subject to competition.
 - clarify role of academic vs. commercial cloud providers