Coordinated Research Infrastructures Building Enduring Life-science services - CORBEL -

Deliverable D6.4
Report on models and pilot designs for sustainable scalable cloud-based provision of data access and compute across infrastructures

WP6 – Data access, management and integration

Lead Beneficiary: Morris Swertz (UMCG)
WP leader: Helen Parkinson (EMBL-EBI), Carole Goble (UNIMAN)
Contributing partner(s): UMCG, CSC, BBMRI-ERIC, Lygature, UNIMAN, EMBL-EBI

Contractual delivery date: 28 February 2017
Actual delivery date: 8 March 2017

Authors of this deliverable: Morris Swertz, Fleur Kelpin, David van Enckevort, Ilkka Lappalainen, Mikael Linden, Tommi Nyrönen, Petr Holub, Jan-Willem Boiten, Anna Leida Molder, Helen Parkinson

Contributors to this deliverable: see table next page

Grant agreement no. 654248
Horizon 2020
H2020-INFRADEV-1-2014
Type of action: RIA

Contributors:

Contributor	***Affiliation***
Jason Swedlow	EUROBIOMAGING
Gianluigi Zanetti	BBMRI.IT/CRS4
Stefan Klein	Euro-BioImaging
Steven Newhouse	ELIXIR / EBI
Dylan Spalding	ELIXIR / EBI
Pieter Neerincx	BBMRI.NL / UMCG
Robert Reihs	BBMR.AT / UGraz
Jarno Laitinen	CSC, ELIXIR Finland
Luděk Matyska	ELIXIR CZ (Masaryk University)
Alexander Vasilenko	MIRRI (VKM)
Renzo Kottmann	Jacobs University
Katja Herzog	EU-OPENSCREEN
Luca Pireddu	CRS4 and BBMRI-IT
Matthew Viljoen	EGI Foundation
Antti Pursula	ELIXIR.FI / Tryggve project
Miroslav Bartosek	Masaryk University
Michal Procházka	Masaryk University, ELIXIR CZ
Bob Jones	CERN
Wojtek James Goscinski	Monash University
Paul Bonnington	Monash University
Pablo Roman	ELIXIR
Roland Pieruschka	EMPHASIS/ Forschungszentrum Jülich
Irene Nooren	ELIXIR.NL / SURF
Christian Ohmann	ECRIN
Steve Canham	ECRIN

Andreas Scherer	EATRIS
Nick Juty	EBI
David Smith	MIRRI / CABI
Serena Battaglia	ECRIN
Wolfgang Kuchinke	ECRIN

Content

Executive Summary

The mission of CORBEL is to facilitate joint operation of research infrastructures of Europe and to provide access to bioscientific resources in a standardised fashion. Task 3 in Work Package 6 addresses the **secure data management and compute needs of service providers dealing with data that needs to be access controlled, for example human identifiable data** such as genome sequences and related personal meta-data, and dealing with users acting in different roles.

As first deliverable we here **report on models and pilot designs for sustainable scalable cloud-based provision of data and compute across infrastructures**, providing guidance to BMS infrastructure development. Therefore we have first surveyed use cases and needs of BMS infrastructures and their users. Subsequently we surveyed existing models for the provisioning of data access and compute. Finally, we have shortlisted a series of pilot designs to inform next steps in the joint development across BMS infrastructures and e-infrastructure providers, in particular within deliverable 6.5 of CORBEL but also in ongoing projects EXCELERATE, EGI-engage, BBMRI-ERIC ADOPT, etc.

The following main concerns were identified:

- Enable data discovery without disclosing identifiable data
- Enable request workflows to grant access
- Establish identity and attributes of the user
- Managing access control across systems
- When data cannot leave the premises
- How to determine if a facility is 'secure enough'

The following existing models were identified:

- Search catalogues with anonymous aggregated data
- Data request followed by download data access
- Closed analysis environments where data is brought together centrally for analysis
- Federated analysis where analysis travels to the data and data is kept locally

In addition the following pilot designs are being proposed:

- Data discovery without disclosing identity (such as beacons)
- Secure cloud extension across providers
- BMS infrastructure in a box and
- Federation of request workflows.

We recommend that in the coming years in CORBEL we will evaluate these models and pilots, share security issues, discuss technical implementation challenges and develop shared materials and best practices, for example:

- Evaluation of beacon query extensions (with ELIXIR implementation study, 12/2017)
- Federated request workflows and piloting of local EGA in CORBEL (D6.5 month 48)
- OpenStack and EGI fedcloud extension (with EGI and EXCELERATE WP4)
- Web cloud extension (TranSMART, MOLGENIS, XNAT) in CORBEL (D6.5 month 48)
- Evaluation of BIBBOX for other BMS infrastructures (with BBMRI-ERIC)

Introduction

This is the first report of task 6.3 of the CORBEL project. This task addresses secure data management needs of service providers dealing with data that needs to be access controlled, i.e., the needs of research infrastructures and their service providers. Here we present deliverable D6.4 **report on models and pilot designs for sustainable scalable cloud-based provision of data access and compute across infrastructure**s, providing guidance to the BMS infrastructure development within and beyond the CORBEL partnership.

This report focusses on the data access and compute needs of 'consented' data, i.e., human subject research data having suitable consent from the data donor (patient/participant) to be used in a research context (see appendix for an overview of terminology used). While other types of sensitive data, such as patient data within the context of health care or commercially sensitive data, might benefit from the same models and pilot designs we will not address their needs explicitly. This report provides essential information for future CORBEL and partner tasks that aim to deliver a federated authentication and access solution to data service provides selected by the project who support BMS ESFRI data management, analyses, deposition and distribution and to improve interoperability with European e-infrastructures and leverage existing investments of these capacities within the biomedical and life science domain.

Match to CORBEL objectives

Many BMS infrastructures engage with human sensitive datasets that are consented for research, such as genome sequences and related personal meta-data/phenotypes, radiology images, lifecourse information and clinical endpoints. Authentication for data access takes place up to a level of assurance requested by the service providers or other stakeholders (e.g. data access committees) that can authorize access based on application requests or researcher affiliations. The CORBEL project will implement pilots driven by (WP3) use cases to enable and streamline use cases in this space, e.g., to demonstrate controlled data flow in multi-center personalised medicine studies that combine IMI/eTRIKS tranSMART and BBMRI MOLGENIS biobanking platforms (bridging these where appropriate). The secure data access services developed will not be case specific, thus the potential scope of applications also includes human sequences, compound screening and high-throughput imaging data.

This report supports the following CORBEL objectives (as described in CORBEL DoW):

Secure data access services - deliver federated authentication and data access services that implement the project's ELSI requirements (WP7) and that can be integrated in domain specific platforms and underlying e-Infrastructures. Services include **local and federated data access modes for managed data, identity management for users and dataset owners, and access to/from secure computational (grid/cloud) environments in which to perform analyses on such data.**

Secure access best practices – produce a generalized best-practice document (e.g. about implementing access management using Security Assertion Markup Language for authentication and authorization) targeted to biomedical data service providers. This document will be more broadly disseminated e.g. as **recommendation for the existing European trust networks (e.g. eduGAIN) who support the biomedical and life science community**. To ensure coherent progress the task will collaborate closely with the H2020 AARC and AARC2 projects (see appendix).

Streamlining access applications – potential dataset users have to repeatedly apply for data access using similar yet different data specific forms. Service providers need to track the users and their affiliations over the time data access is granted. CORBEL aims to develop a common method using existing AAI tools to create a process for data service providers to identify research data users and authentication and authorization attributes associated with users using standard data security exchange formats compatible with the existing e-Infrastructure solutions. This will **streamline user management processes, data request events, and data access review and provision groups among biomedical data service providers** as well as bring tangible benefits to users by reducing red tape.

Federated authentication - current identity federations in Europe belonging to GÉANT eduGAIN vary in the extent that the identity attribute information can be trusted, and the number of attributes about users that are shared by the user home organisation to scientific service providers. Communicating the level of trust/assurance about the quality of user identification information provided to the scientific service providers, as well as the number of user attributes (e.g. home institution and email) are appropriate in specific controlled-access data management contexts. **CORBEL will engage with partners, TERENA, EGI and relevant international activities (e.g. Global Alliance for Genomics Health GA4GH, Nordic e-Infrastructure Collaboration NeIC, GEANT AARC/AARC2 projects)** to report on current state of implementation of federated authentication within the biomedical community. CORBEL will then use the existing knowledge to map levels of trust and verification needed within a heterogeneous network of BMB RI data service providers and make a recommendation how to achieve improvements in secure data access in collaboration with e-infrastructure providers.

Deliverable goals and scope

This report aims to support CORBEL goals by identifying **existing and emerging solutions that address secure data management needs of service providers dealing with consented human subject data that needs to be access-controlled**.

Examples of scope include:

- genome sequences and related potentially privacy sensitive metadata
- streamlining access request applications for requesting and granting access to data
- federated authentication, authorization and data access services
- access to/from secure computational (grid/cloud) environments
- expressing different levels of trust and user attributes

This report will summarize needs and use cases, survey existing solutions, and shortlist pilot ideas for data access that integrate existing solutions, implemented within the time frame of CORBEL, and as basis for BMS infrastructure implementation. The focus will be driven by (WP3) use cases, e.g., to demonstrate controlled data flow in multi-center personalised medicine studies. Privacy enhancing systems that are applied to the data are out of scope for this deliverable, i.e., technologies to de-identify, anonymise or pseudonymise data.

Approach

The objective of this report is expressed in the following research question: What are current models and pilot designs for sustainable scalable (cloud-based) provision of sensitive data access and compute within and between BMS infrastructures?

We therefore have surveyed expert users and providers from the BMS communities (see contributors) to identify requirements and use cases. We used an open question survey/interview to ask the BMS infrastructures and the supporting e-infrastructure providers the following sub-questions:

1. What are current and planned use cases and scenarios for access and compute across infrastructures for your community in the context of sensitive data? *Please give a short description of each use case and for each their most important requirements for access.*

2. What are current access models that you use/provide to serve data access and compute needs? *Please give a short description of current access models that are available, what systems and/or tools are used, the funding/business/model that pays for these, and to what extent they address the requirements of the use cases*

3. What opportunities, if any, do you see to enhance current models or to pilot new access models currently not yet available and how do you expect these will compare to current state of the art?

Subsequently we surveyed existing infrastructure and tool providers (see contributors) within the context of access to sensitive data, again complemented by literature research. In addition we have had 2 international workshops with strong representations of both e-infrastructure and BMS user communities (see contributors) complemented by literature research. Based on these we defined BMS needs and pilot designs as a basis for joint implementation of interoperable data access models in BMS infrastructures and supporting e-infrastructures to be used beyond this report. Below, each of these elements is described in detail.

Data access needs within BMS infrastructure communities

This section summarizes the use cases and needs collected from each BMS infrastructure. In the next section we will integrate these results. We received input from BBMRI-ERIC, EATRIS, ECRIN, ELIXIR, Euro-BioImaging, Instruct and MIRRI.

BBMRI-ERIC

BBMRI-ERIC is an infrastructure that provides/facilitates secure and privacy-protecting access to key resources in order to support biomedical research and to support healthcare advancement. Managed resources include:

- biosamples from biobanks
- related data such as clinical, omics, phenotypes, etc.
- expertise and other services (e.g., sample & data hosting)
- biomolecular/omics data.

The major goals of BBMRI-ERIC are:

- to increase use of material and data stored in European biobanks, while adhering to strong privacy protection of patients and donors contributing the material and data;
- to improve quality and traceability of the material and data in European biobanks, addressing the alarming recent publications demonstrating that large portions of biomedical research are not reproducible, which has even been demonstrated specifically for the process of generating data from samples;
- to improve data harmonization and contribute to the standardization processes
- to contribute to the ethical, legal, and social issues, with particular focus on cross-border exchanges of human biological resources and data for research use.

Figure 1 below, from [Holub16], shows the perspective of BBMRI-ERIC IT on systems for sensitive data. Orange components are assumed to be built by BBMRI-ERIC, blue components are expected from other e-Infrastructures. Orange-blue components are assumed to be developed jointly with other e-Infrastructures and/or with the national nodes, e.g. BBMRI-NL, BBMRI.FI, etc.

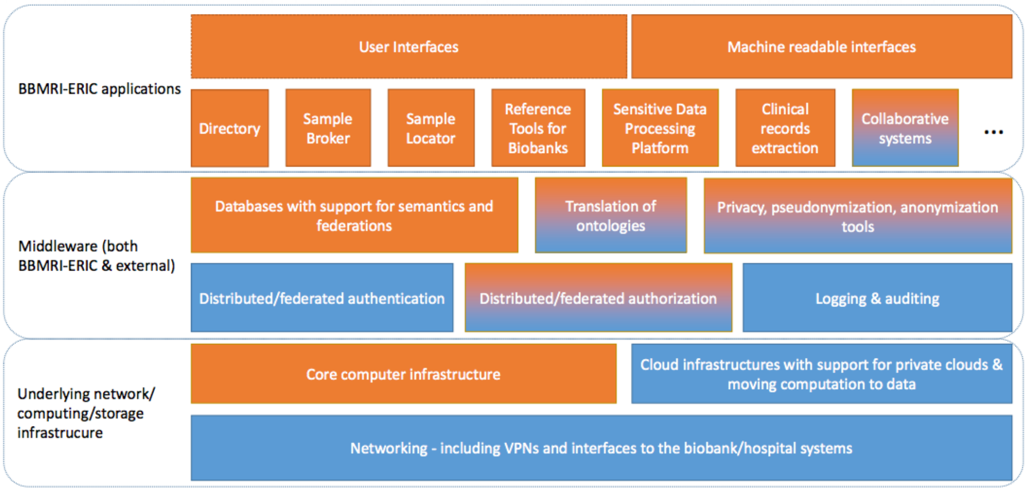

BBMRI-ERIC applications
User Interfaces
Machine readable interfaces
Directory
Sample Broker
Sample Locator
Reference Tools for Biobanks
Sensitive Data Processing Platform
Clinical records extraction
Collaborative systems
...
Middleware (both BBMRI-ERIC & external)
Databases with support for semantics and federations
Translation of ontologies
Privacy, pseudonymization, anonymization tools
Distributed/federated authentication
Distributed/federated authorization
Logging & auditing
Underlying network/ computing/storage infrastrucure
Core computer infrastructure
Cloud infrastructures with support for private clouds & moving computation to data
Networking - including VPNs and interfaces to the biobank/hospital systems

Figure 1. BBMRI-ERIC systems overview

A typical scenario starts with a user searching for samples or data or biobanks to start a collaboration (see the Directory and Sample Broker/Locator components). Before accessing samples or privacy-sensitive data (data that is personal and not anonymous), the user must submit a project that undergoes ethical evaluation, and only users with approved projects will be supported. The users then request the samples and/or data and negotiate with biobankers responsible for those samples/data. Once the user's request is approved, the user signs an MTA (Material Transfer Agreement) and/or DTA (Data Transfer Agreement) and the samples / data are given to the user. Depending on the type of the request, the biobank can transfer either anonymous data or pseudonymous data with strong-enough MTA/DTA that prevents recipients from any re-identification attempts.

When processing privacy-sensitive data, it might be required that non-deidentified data never leave the biobank. In case of multi-center studies across multiple biobanks such data is often integrated via a mutually trusted compute environment. For example, in BBMRI-NL there are dedicated compute cluster environments set up around Genome of the Netherlands and Transcriptome of the Netherlands projects in Groningen and Amsterdam. Alternatively, federated approach to the analysis can be used, which means that the processing of pseudonymous data or even non-deidentified data takes place inside the biobank and only the aggregate anonymized data are sent out to the researcher; this has been previously described and demonstrated, e.g., using DataSHIELD [DataSHIELD14]. An extended version of this scenario is targeted by the Sensitive Data Processing Platform component in the software stack diagram. Another specific aspect of the BBMRI-ERIC infrastructure is the heterogeneity of data that are coming into the biobanks and that need to be mapped into consistent data sets. Therefore BBMRI-ERIC works with federated databases with semantic data support and translation of ontologies, which is being addressed in Task 6.2.

Privacy and security are fundamental concepts that must be built into all BBMRI-ERIC IT services by design, as trust and transparency are the key elements of medical research infrastructures dealing with privacy-sensitive human data, see [Holub16] for an extensive analysis.

EATRIS

EATRIS ERIC is a distributed research infrastructure which offers a network of service providers and expertise around translational research with a joint mission to bridge the 'translation gap' between research and medical products, such as drug discovery, vaccines and medical devices. Objectives are to facilitate collaboration, and development of tools to support this. Overarching theme is personalized medicine, moving away from 'one size fits all' treatment towards specific (personalized or stratified) treatments based on detailed patient profiles.

Users receive access to relevant services provided by EATRIS member facilities, which are high-quality academic centers. Access models include simple bilateral (contractual) agreements between the user and the service provider, wherein the degree of use is specified. It will therefore be decided in a case-by-case scenario between the user and the provider as to how samples and data are dealt with, and how potentially sensitive information are handled (data storage, data pseudonymisation, etc).

Scenarios may range from long-term storage of data at the provider site, at an external site, short-term storage, or no storage at all. In specific cases of doubt, EATRIS sites and EATRIS coordination and support may consult experts in data stewardship for further advice. Issues include secure multi-center acquisition, processing and sharing of clinical data (same as ECRIN), secure and efficient de-identification and sharing of clinical images (same as Euro-BioImaging), secure multi-center biosample logistics (the same as BBMRI), efficient creation, analysis and dissemination of molecular data (ELIXIR), tools to integrate and analyse this data in research project (eTRIKS) and a common IT foundation to support all of above (CORBEL).

So far there are no use cases for infrastructure-provided data access / compute access scenarios. There is no specific data access module in place. Data access largely depends on the EATRIS partners and the user. Possible scenarios include the use of specific software, hardware, external partners with expertise in data stewardship (e.g. ELIXIR), web application (e.g. Webmicroscope), etc. Access and use of the various options depend on the project requirements and can be adjusted to meet the user's need.

ECRIN

ECRIN does not currently provide any computing infrastructure. The prime scientific IT need of trials units, for management of their trial data, is organised by each unit independently – using IT infrastructure that they may manage themselves but which is more normally supplied by their parent university or hospital as an IaaS or PaaS service. The software used may be commercial (probably about 60%), open source (about 20%) or built in-house as a proprietary system (about 20%). A survey of all ECRIN centers in 2009 yielded 59% commercial solutions, 6% open source and 32% proprietary ones [Wolfgang Kuchinke, Christian Ohmann, Qin Yang, Nader Salas, Jens Lauritsen, et.al. Heterogeneity prevails: the state of clinical trial data management in Europe - results of a survey of ECRIN centres. Trials 2010, 11:79. DOI: 10.1186/1745-6215-11-79]. The situation becomes more complex, because the leading investigator (The so-called sponsor) decides about what software solutions are used. Thus, in addition to clinical data management solutions, software for adverse events reporting, clinical sites management, patient recruitment and clinical trials management may be used. A small but growing number of units are using systems that are completely externally hosted, usually by the system vendor, i.e. as a SaaS system.

There are, however, some problems with this arrangement, and a central service housed within public infrastructure, providing secure access to sensitive data, using a variety of existing systems to process the data, could be of interest to many clinical research units and groups. A growing use case is for the long-term storage of data, both for general archiving purposes and as a prelude to possible sharing of that data with others. An increasing number of data repositories have been developed, some specifically designed to manage sensitive data. It is not clear at this point if existing and planned repositories will be sufficient, or if an additional repository for trial data, specialising in European non-commercial studies, could have a useful role to play. A further, associated need is to support *in-situ* re-analysis or meta-analyses of datasets, involving the temporary import of data from different repositories into a specialist computing system and the provision of a suitable analysis system (e.g. an R environment).

The ECRIN pilot of EUDAT developes a repository for the secure, transparent and GCP (Good Clinical Practice) compliant storage of clinical trials data (https://www.eudat.eu/communities/the-use-of-the-eudat-repository-to-store-clinical-trials-in-a-secure-and-compliant-way). For such a safe and accessible storage the B2SAFE and B2SHARE services of EUDAT are used and adapted. In addition, an authentication service (AAS) manages the access rights for users; linking to metadata through B2FIND and the data type registry is available. The pilot is developed in cooperation with the "EUDAT sensitive data group" (https://eudat.eu/a-eudat-working-group-on-sensitive-data-management), which supports the creation of restricted access areas inside B2SHARE for sensitive data. It is planned to enable a transfer of sensitive data from the restricted access area of B2SHARE into a secure analysis area (e.g. ePouta) to enable sensitive data analysis without the need to download all data by the data user.

Finally, because of the wide variety of locations where trial data, and associated documents like trial protocols, are stored – there is a pressing need to a) agree a common metadata system for clinical research data objects [Canham16], and b) develop a repository system to house that metadata and allow its easy search by humans and machines alike.

ELIXIR

ELIXIR is an intergovernmental organization consisting of national Nodes based in the member states and a Hub located in Cambridge, UK.ELIXIR was established to coordinate the national resources so that they form a single sustainable infrastructure supporting European life science requirements. The coordination is done through five distinct platforms focusing on data, tools, compute, interoperability and training. The compute platform and inteoperability platforms are most relevant for this deliverable.

The ELIXIR compute platform roadmap includes support for large scale data transfers, storage and compute (http://bit.ly/elixirtech) services. The researchers are authenticated using ELIXIR identity based on social or academic identity management architectures. ELIXIR Identity can be enriched with authorization attributes that provide access to data or services. The ELIXIR interoperability platform provides standards, services and best practice for identifiers, workflows, semantics and and is driven by CORBEL use cases and requirements.

Euro-BioImaging

In the Euro-BioImaging community we see the following sensitive data scenarios:

1. Multi-center studies increasingly collect medical imaging data, such as magnetic resonance imaging (MRI), computed tomography (CT), positron emission tomography (PET), or ultrasonography (US). These data are typically acquired in the “DICOM” format, and stored in the clinical picture and archiving communication system (PACS) at the participating institutions. For central review and analysis, it is required to gather all data into a central research archive, which can be accessed by the researchers involved. Hereby, proper de-identification of the imaging data is crucial; this de-identification has to take place at the institution itself, prior to uploading it to the central research

archive. The central research archive should have flexible mechanisms for controlling access (read/write) by the different users.

2. Single-center clinical and population studies that collect medical imaging data have similar needs as multi-center studies described above. Systematic storage and management of imaging data in a secure research archive, possibly hosted within the firewall of the institution, is demanded.

3. Once medical imaging data (be it from a single-center or multi-center study) is stored in a research archive, procedures for efficient quality control, automated image processing, and extraction of quantitative imaging biomarkers (e.g. hippocampus volume as a biomarker for Alzheimer's disease; knee cartilage thickness as a biomarker for osteoarthritis; computational radiomics features for assessment of tumor properties) are needed. Image processing may require heavy computational resources. Therefore, scalable or cloud-based solutions, integrated seamlessly with the data storage, are required for this purpose. Results of image analysis should be uploaded again to the central imaging archive, while retaining a link to the original imaging data for provenance reasons.

4. To support use case 3.4 (see WP3 CORBEL), quantitative imaging biomarker data must be brought together with clinical and genetic data, to enable a joint statistical analysis. Such an integration platform should be secure, and access rights should be controlled per user. Basic statistical analyses should be supported by a web-based platform, and for more advanced analysis the user should be able to download selected subsets of the data to his/her own environment.

INSTRUCT

Instruct is a distributed infrastructure addressing the field of Structural Biology. As such, it focuses in bringing spatial three-dimensional information to the understanding of biological processes, often reaching atomic or quasi-atomic resolution. Its ERIC status has already been approved and publication in the Official Journal of the European Union is expected to happen in early 2017.

The range of technologies and applications in which Instruct is involved is very large and diverse. Most often, data are not patient-related, although increasingly there are areas of work in which personal data is involved, such as Nuclear Magnetic Resonance in metabolomics and some applications of X-ray microscopy, with data sizes in the few GB's. Long term data preservation is very important, as it is indicated in Instruct Data Management Plan, although a general approach is still under study, considering that, for some technologies, each instrument may produces several TB's per day.

MIRRI

MIRRI is in the process of establishing its legal entity at the end of its preparatory phase, it is close to having 11 countries on board but as yet it does not have a centralised data system. mBRC (resource centre) partners do contribute to centralised systems such as straininfo.net www.straininfo.net/ , CABRI www.cabri.org/ and the WDCM tools www.wdcm.org/ e.g. Global Catalogue of Microorganisms http://gcm.wfcc.info/. DSMZ are running a system they hope to expand to other resource holders, BacDive http://bacdive.dsmz.de/; therefore there are efforts but not at the Research Infrastructure level yet.

The MIRRI partners tend not to deal with patient information, the sensitive information around microbiology may relate to areas such as biosecurity, customer data etc. not the focus of CORBEL's focus on sensitive data. General guidance for Biological Resource Centres is given in OECD BRC guidance documents [OECD07], in particular *"BRCs should introduce appropriate measures (protocols, tools and standards) in their own informatics systems to assure reasonable security of information. There are existing systems, e.g. authentication by user ID and password, encryption, encryption of messages and restriction of IP addresses that may provide the basis for such measures. Backup-files should be stored in secure cabinets."*

Individual institutions in MIRRI are requested to follow such guidance, this will be mandatory through the partner charter when MIRRI is fully established. The focus is on an integration environments with two key tasks:
1. to make microbial culture collections (CC) catalogues data visible and accessible from Life Science databases,
2. to make Life Science databases records visible and accessible from CC aggregated catalogue, in two formats: a. for human access, b. for computer programs.

Data access use cases and concerns

This section integrates recurring needs and use cases named across the BMS infrastructures in the context of secure access to sensitive data (i.e. sub-question 1). We ordered the concerns following a research data cycle in case of reusing existing sensitive data, i.e., first users need to find relevant data, then they need to request access, then access must be granted and a suitable environment for data reuse must be made available:

Enable data discovery without giving access

The first concern is to enable potentially interested researchers to find the data. As a BMS infrastructure there is a desire to increase the (re)use of data. However, if data is under controlled access then interested users may not be able to search the data directly. Therefore, as a BMS infrastructure one wants to enable data discovery without giving access to the individual level data.

The most commonly used data access method to serve this need is to extract summary level data from the individual datasets and collect those in a central place. While this optimizes the protection of data privacy, it limits searches to pre-computed counts only as there is no 'live query' capability. As a consequence, much time might be lost in unnecessary communication between data users and data providers to elucidate if actually relevant data is available.

An example model to serve this need is implemented in the BBMRI-ERIC directory. Here interested researchers can search metadata from 500+ data and sample collections. When a collection that might contain one or more items of interest is found, then the researcher must engage in direct communication with the biobank to ask how many relevant items are actually available. To also ease this phase of the data discovery, BBMRI-ERIC is developing the 'negotiator' system where researchers can easily engage in conversation with multiple biobanks at once.

The method of precomputing the search queries has limitations in the capability to find specific and rare data/samples. Therefore alternatives are being piloted where users can search within the data, but without disclosing privacy sensitive information. For example, by fragmenting the data sufficiently, like in the Beacons (see below) or by increasing the granularity of the data such that no individual records can be disclosed (as for example in the Dutch pathology database where can be searched in 30 million samples but no query results in a count of '1').

Request workflows to grant access

When a researcher discovers relevant samples the next step is to acquire permission to access the data. As BMS data provider of consented data the provider has the obligation to establish if the data user will actually adhere to the permissions given by the data donor in the informed consent. Therefore, each BMS data provider has implemented / is implementing data request procedures where the data user needs to fill in some form specifying how and for what purpose the data will be used, which is subsequently approved by a data access committee (DAC). Currently, data request is often implemented per data provider ranging from simple email + Word document forms up to advanced online data request workflows. However, because these request workflows are mostly implemented differently, researchers in need of data from many providers must fill in the request multiple times.

An example of such model is based on the ELIXIR REMS software [REMS1], [REMS2]. It supports an electronic data access application process for granting secondary research access on sensitive data governed by a Data Access Committee (DAC). The tool provides federated authentication of each applicant, a customizable application form, and communication between the DAC and the applicants. Once access has been approved REMS can feed data access authorizations as part of the authentication process to the service providing the requested data. REMS is currently used for managing data collected from the Finnish National Institute for Health and Welfare biobank samples [THLDA]. Future work is to evaluate success of REMs in EGA and extend it to Biobank request workflow integration.

Identity and attributes of a user (authentication)

When permission is granted to access the data, the next task is to technically enable the access. However, before authorization there needs to be insurrances on the identity of the person. Research is increasingly international, and as a consequence, most BMS data providers typically don't know the person requesting the data which makes granting access not a trivial task. Because before granting access the data provider must have assurance that the user requesting access is actually the person she claims to be (authentication) and also that she indeed has the credentials she claims to have, such as the organisation (s)he is part of (attributes).

Authentication might be a slightly confusing term [Holub16], as it needs to comprise two equally important steps: (a) registration process, which binds the virtual identity to the physical identity of the person (e.g., by visiting in the registration office with a government-issued ID card while creating the virtual identity), and (b) authentication instance, which is verification of the person's virtual identity (e.g., a person proves possession of her virtual identity using a password).

In addition, granting access (authorization, see next section) may often depend on the attributes that can be attached to the individual asking access. For examples, the following attributes are often used:

- institutional affiliations/roles which assert the user has a certain relation to the given organization, e.g., an employee, a student, or a faculty member of an educational institution,
- project affiliations/roles which assert the user has affiliation to a project or even more specifically that the user has a certain role in a project,
- group affiliation, where it is possible to describe adherence of the user also to any other virtual group or subgroup (which is a generalization of institution/project roles).

Currently, there are only limited examples of EU wide authentication systems that also provide attributes.

One well known example is eduGAIN, a service of identity federations around the world, simplifying access to content, services and resources for the global research and education community (see the pilot section below). This service is the basis for many of the identity services within the pilots currently underway in BMS infrastructures and is of key importance for future integration of BMS services. However, there is still much work needed to translate services like these to use in systems for research data access and also providing sufficient confidence that the attributes provided are up-to-date.

Managing access control (authorization)

When access permission is granted, the next challenge for many BMS service providers is to actually technically enable the access which translate to ‘authorization to access particular systems’. These might be implemented in various ways, depending on the facilities available and the technical capabilities of the service. For example, using SFTP for download, or providing access to a computer cluster or online web application to analyse data in place. Often, multiple IT systems are involved making access difficult.

Ideally, BMS owners would want to be able to delegate authority for granting access to particular individuals (e.g. ‘data managers’) and enable these individuals to give (or revoke) access to all systems via a centralized authorization system. Several technical solutions are currently being piloted for federated authentication and authorization.

An example of this model is the ELIXIR Authentication and Authorization Infrastructure (AAI) allows single sign-on to services across ELIXIR. Here the Perun system for identity and group management is used (see Appendix) which then can be used by connected systems/applications as basis for allowing/preventing access. Another example coMANAGE which is implemented in the collaboration of the Dutch SURF foundation, BBMRI and the Dutch university medical centers. In any case, implementation of centralized access control management systems is a lengthy activity because it must be connected to many local access control systems. At the same time, the need for these systems increasingly arises from practical use cases around biomedical sensitive data, since researchers want to use multiple applications, data sets, and compute resources in a concerted manner.

When data cannot leave the premises

Some BMS data providers have the limitation that data can only be analysed as long as it stays within the (fire)walls of the data provider organization. For example, some biobanks such as Lifelines (http://lifelines.net) and the UK 100,000 genomes project require that no data leaves the biobank's IT systems unless explicitly consented. For example, under a specific MTA some data has been shared into a central BBMRI-NL analysis environment shared by multiple biobanks to enable multi-center analysis. When transnational access is required then many more data providers have limitations with respect to data leaving the country, also depending on national legislation.

Therefore several protocols and technical solutions have been developed. On one hand there is the 'closed box' model where data can only be analysed using a central IT facility, but with technical measures in place to prevent downloading. For example, Lifelines uses a remote Windows desktop solution that allows researchers to perform their analysis on site (with downloads made impossible).

Alternatively, there are solutions that enable distributed analysis, i.e., enable federated analysis where the analysis algorithm is sent to each dataset separately as compared to first bringing all data together for analysis. This has the large added advantage that it allows data to stay within e.g. a country's jurisdiction and enables sharing when it would not otherwise be shared. There are also bandwidth issues for moving around large datasets that makes this scenario attractive in case of large data with small computational demand. A potential drawback of this method is that it constrains analysis freedom, thus hampering research progress in order to increase assurance that no undesirable data access happens. A well-known example of this solution is DataSHIELD [DataSHIELD14]. The latter model requires very precise data standardization of data models across the participating centers, since manual inspection of potential deviations is not possible anymore. In both models, there is usually an audit trail to ensure logging of (unwanted) behavior.

Finally, an interesting pilot approach to address these concerns is the concept of 'local EGA' (see appendix 2). Traditionally the European Genome-phenome Archive (EGA) is a central data repository. However to also serve BMS data providers requiring data to stay local or within a trusted analysis environment pilots are underway to install EGA locally so that data doesn't need to be submitted centrally. Otherwise, the local EGA functions identical to the centralized EGA enabling central cataloguing of the dataset and basis for federated request workflows. When data request is granted users could be directed to the local EGA installation.

How to prove that a facility is ‘secure enough’

When the IT needs are larger than a BMS data provider can manage there is often a desire to outsource the facilities for data access. Or alternatively, when multiple BMS data providers want to pool their data they also need to sent their data to IT services hosted by a third party. However, then BMS loses control over the IT infrastructure while still being responsible for the security of data access. Therefore BMS providers are concerned that a service might not be ‘secure enough’. However, it is unclear what conditions must be met and what certification helps to prove this. For example, ISO/IEC 27001:2013 provides a set of rules and procedures but these might be more advanced what many biobanks would require from their local IT department. Meanwhile, the concern of security of the facility can be alleviated by reducing the sensitivity of the data using privacy privacy enhancing technologies such as pseudonymization and anonymization, as e.g. described in ISO 29100.

Current models for sensitive data access and compute

This section summarizes commonly used models that are used in the BMS infrastructures to serve data access and compute needs (i.e. answer to sub-question 2). Pilot designs are described in the next section.

Central catalogues with aggregated/summary data

As described above central catalogues provides tools for findability of data or samples by providing summary data and metadata about the collections of interest. Confidentiality of sensitive data is preserved through two key attributes of a catalogue:

1. The data is highly aggregated
2. The data is generalized

Through the use domain specific minimal information models a catalogue can provide relevant information to find relevant data sets, to which access can be attained through one of the methods described below. However catalogues by nature also have some limitations. They cannot answer detailed questions about the data or provide accurate counts on availability. For example the BBMRI-ERIC directory describes biobanks and their collections based on the MIABIS Core minimal information model [MIABIS16][DIRECTORY].

Data access committee followed by data access

Another often used model is where data access is requested at a data access committee (DAC) followed by the ability to download the data.

For example, the European Genome-phenome Archive (EGA) is a service and database for permanent archiving and sharing of all types of personally potentially identifiable genetic and phenotypic human data resulting from biomedical research projects [EGA2015]. The EGA is a shared resource through a collaboration between the European Bioinformatics Institute (EMBL-EBI) and the Centre for Genomic Regulation (CRG). The current EGA model between EMBL-EBI and CRG shares responsibilities for data submission APIs, data access APIs and tools, helpdesk operations, and data federation between the two sites. The EGA includes major reference data collections for human genetics research, such as UK10K (http://www.uk10k.org/), Wellcome Trust Case Control Consortium (https://www.wtccc.org.uk/), RD-Connect (http://rd-connect.eu/), and International Cancer Genome Consortium (http://icgc.org/). The EGA contains exclusive data collected from individuals whose consent agreements authorise data release only for specific research use or to bona fide researchers. In all cases, data access decisions are be made by the appropriate data access committee (DAC) and not by the EGA. The DAC will normally be from the same organisation that approved and monitored the initial study protocol or a designate of this approving organisation. EGA's processes are shown in Figure 2.

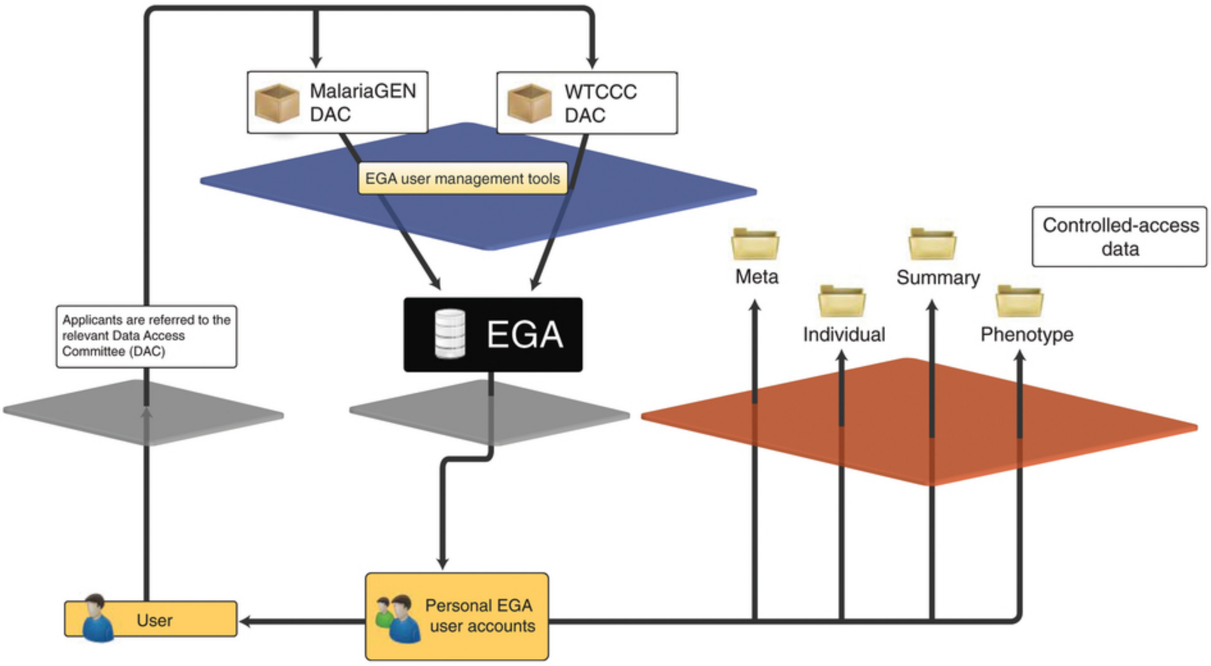

MalariaGEN DAC
WTCCC DAC
EGA user management tools
Controlled-access data
Meta
Summary
Individual
Phenotype
Applicants are referred to the relevant Data Access Committee (DAC)
EGA
User
Personal EGA user accounts

Figure 2. The EGA model, reproduced from Lappalainen et al 2015

Centralized data analysis workspace

Traditionally, large multi-center research consortia address their data access needs by physically bringing data together in one place. Direct data access is given to all partners within legal boundaries of the consortium/access agreement. This is still by far the most common approach of data sharing for several reasons: it allows for direct sharing of compute and analytics resources within the consortium, it simplifies data quality control, it allows central management of user access and technical support, and can be implemented using proven-technology.

Examples within the CORBEL WP3 context include the tranSMART repository for genomics biomarkers, the MOLGENIS for integration and analysis of biobank and geno-pheno datasets, and the XNAT image archives, although most of these technologies would in principle also support a more federated approach. Example within the BMS partners include BBMRI-NL where more than 20 biobanks have brought metabolomics, genomics, transcriptomics data together within one computer cluster enable a single access point to all these integrated data and simplifying access procedure to one DAC.

While the simplicity and ability to use proven technology is the clear advantage of this approach, it also has some drawbacks that need to be addressed carefully:

- large central data collections are also sensitive to major data leaks and therefore tend to require a higher level of data protection and data security;
- the managing entity must be trustworthy (and be able to provide evidence for that) for all partners in the consortium before they would be inclined to hand out sensitive data;
- the transfer of sensitive data usually requires a legal agreement to establish roles and responsibilities (e.g. in the form of data processing agreements) and to establish the rights and obligations on the data (e.g. in the form of a data transfer agreement);
- user administration tends to be rather complex requiring dedicated processes and/or procedures for initial user identification (is the user really the individual he/she claims to be?), requesting access to the data (who is allowed to request access, and how does the service provider verify that this request is legitimate?), and how to keep the user administration up-to-date (who is notifying the service provider when a user left the organization or project?).

In practice, the situation may be even more complex because the centralised and federated use cases are not strictly separated. For instance, the tranSMART use case for imaging and genomics biomarkers (use case 3.4 in WP3) utilizes a central repository for the clinical data and the conclusion data for imaging and genomics experiments, but links back to other databases that may be federated, e.g. for images (XNAT), raw genomics data (e.g. EGA at EBI), and biosamples (e.g. using the MOLGENIS catalogue). An additional complicating factor may be the federation of the underlying compute facilities: it can be foreseen that scale-out to cloud compute resources will be required for some of the analytics in these central data repositories, which adds an additional layer of complexity to the user authentication/authorization technology. Nevertheless, this hybrid implementation model offers the best of both worlds from the functionality perspective and is therefore likely to stay around. As a consequence, it is recommended to select a pilot use case with such a hybrid central/federated implementation within CORBEL to provide a best practice implementation of the data access processes/procedures.

Federated data access via distributed (meta)analysis

This is the general alternative to the centralized approach. Instead of bringing the data centrally for analysis, the data is kept distributed and a central analysis protocol is executed in a distributed way on each partial dataset and the results are integrated.

A well-known example includes GWAS meta analysis which is the analysis of genotype and phenotype associations that are both large and very privacy sensitive (in particular the DNA data). To reach sufficient statistical power there is a need to reach 100,000s of samples that almost no individual laboratory has. Meanwhile the computational demands are relatively modest and the data very standardised which makes it possible to define the analysis protocol centrally and then send the procedure as a computer script to all partners that can be run locally in 10s of labs each contributing a few thousand samples and the results are then integrated.

Another example is the DataShield method [DATASHIELD2014] which is more focussed on phenotype/epidemiological data and enables interactive meta-analysis. However, phenotype data is typically not yet standardized requiring much more interaction between the data provider and the central research team to 'harmonize' the data such that it can be analysed in unison. DataShield enables data providers to give permission to execute a particular analysis procedure and to only return the analysis results, without disclosing the underlying data. This can be done in a more interactive fashion between data provider and researcher.

Because of size of the data and its nature, the paradigm of moving computations to data can substantially improve the computational applications. This has been promoted over the last 10 years and has become practically available with the advent of cloud technologies that can also be deployed within the perimeter of a BMS data provider. For example, private clouds for processing of biobank data has been developed and its use has been demonstrated by the BiobankCloud project. Similar efforts have been reported in other BMSes.

Pilot designs for cross-esfri implementation

This section describes new pilot designs that implement and integrate existing and emerging solutions towards BMS infrastructure needs (i.e. answer sub-question 3):

Data discovery without disclosing identity (beacons)

Several pilots have been started to enable discovery of sensitive data without disclosing the identity of the data donors, i.e., prevent the search to result single identifiable records. Examples of this emerging data access model include the public database of the Dutch pathology labs and CafeVariome projects.

One of the most well known initiatives for sensitive data discovery is the GA4GH Beacon Project. It allows genomic centres in the world to make their data sets discoverable through a standard data query interface by supporting simple questions such as "Does your data resource have any genomes with this allele at that position?". More than 70 Beacons worldwide have been lit – several of them across the Europe funded by ELIXIR. These Beacons will provide data using three access tiers model whereby the normal Beacon queries are public and aggregate data are served through registered tier. The registered data access is implemented by integrating Beacons to the ELIXIR AAI services. There is a clear opportunity to link the beacon concept to CORBEL use case 3.4 (integrating clinical, genomics and imaging data for biomarker discovery), in particular enhancing the beacon concept for the secure and privacy-sensitive discovery of tumor mutations in disparate data collections. In ECRIN there is also development of a metadata repository system, holding details of clinical research data objects of all types (i.e. a variety of documents and papers as well as datasets), identifying the studies that generated them and the arrangements under which the objects can be accessed, and that will allow user and machine based search.

Secure compute cloud extension across multiple providers

Another class of pilots tries to combine the concern that data cannot leave the premises with the need to share data between multiple BMS providers, the need to bring additional compute power to the data, and the need to overcome the limitations of the federated data approach. In this emerging data access model systems from multiple BMS providers and/or compute infrastructure are linked together into one secure system such that data is still only persisted within the (fire)walls of the BMS data provider while researchers can access multiple data sets in unison during analysis.

For example, the Tryggve project (see appendix 1) has connected two OpenStack clouds in the same network to enable cross-border extension of secure clouds. This enables accessing of sensitive data on secure computing environments regardless of their location, and allows either moving of the data or moving of the software tools to make combined analyses. The result is a virtual computer environment that connects the different sites during analysis that can be dissolved when the analysis is complete, combining the best aspects of the centralized model with the federated model. This approach, demonstrated between ELIXIR Nodes of Sweden and Finland, also provides an example of utilising secure cloud backend for extending local system, while preserving the security of the local system [TRYGGVE16]. The secure cloud IaaS used is the Finnish secure ePouta cloud.

Another example is the piloting of scalable cloud-based compute infrastructure for large-scale automated image processing within the BBMRI-NL project in collaboration with Euro-BioImaging. Currently, large-scale automated image processing is only possible for institutes that have access to a compute cluster; moreover, such clusters are not scalable. Use of cloud services such as creating a virtual openstack cluster that spans multiple sites or the use of federated cloud provided by EGI (www.egi.eu) is in early test phases. In addition, BBMRI-NL is working on 'e-lan' which is a lightpath based dedicated network between the Dutch research centers which can provide the necessary fast and trusted network foundation to underpin the cloud extension model and first pilots of a 'virtual cluster' that ran on remote storage (for sensitive data access) gave promising results.

Similarly, the cloud federation can be implemented on the level of web applications. In BBMRI-NL and CORBEL WP3, we are currently working on couplings with data warehouse solutions like tranSMART, XNAT and MOLGENIS. The vision is that imaging biomarkers stored in XNAT can be transferred seamlessly to TranSMART/MOLGENIS, where they can be analysed in conjunction with clinical and genetic data. This depends on development of a single-sign-on system for the various web services and compute platforms mentioned above would simplify their usage in such data integration scenarios. The implementation of these can potentially benefit from FAIR data backbone technologies developed as part of the interoperability task in CORBEL WP6.

BMS IT infrastructure in a box

The challenges of dealing with sensitive data put large demands on BMS data providers when setting up IT infrastructure, which even when using existing solution is still a huge configuration challenge. Therefore, there is an emerging model where all the necessary software is bundled and preconfigured so its configuration can be easily shared. This pilot model is an interesting complement to the secure compute cloud model described above.

An example of this model is BIBBOX (http://bibbox.org) which currently under development in BBMRI-ERIC common service IT (focusing on open source biobank software) and in the H2020 project B3Africa (focusing on LIMS and bioinformatic solutions, http://www.b3africa.org). BIBBOX provides a virtual machine (configurable by vagrant/puppet) to host (open source) software tools bundled together as docker containers. Through the combination of data and software in an “application container”, which are even transferable between VMs. BIBBOX provides easy to use functionality for installation, monitoring and backup of software tools. Furthermore integration support for user management (LDAP, SSO) and ID management will be provided. BIBBOX was started in biobanking / bioinformatics, however is not limited to this research domains. Developer information can be found at the BIBBOX GitHub and BIBBOX docker hub https://github.com/bibbox/bibbox-documentation.

ECRIN has also shown interest in this model, focussed on a centralized hosting environment for clinical trial systems (open source and commercial) that can provide a high quality PaaS service to researchers, that is designed to meet the specific requirements of clinical trials data management (for instance by maintaining a transparent, fully documented validation regime of systems and system changes, or flexible, fully documented and reporting back-up, restore and disaster recovery systems). Such a system must also incorporate robust access control between and within different studies.

Federated request workflows

When data needs to be requested from many BMS providers this puts a large burden on both researcher and the BMS providers as many requests must be processed for one study. Therefore there is emerging the model of ‘federated request workflows’ by creating interoperability between the various request workflows and by centralizing the management and processing of these requests.

A well known example is the EGA that is moving to a more federated model (see appendix 2). This is includes federated authentication and authorisation - currently EGA maintains its own single centralised password based authentication and DAC controlled authorisation. It will be necessary to enable third party platforms to authenticate EGA users and verify authorisations, e.g. enabling access to data cached in a cloud; Identity management - currently EGA account identities are not linked to any other user identity, e.g. an ELIXIR identity, REMS; Secure streaming of subsets of data - currently EGA data is delivered as very large encrypted data files, no matter what part of a dataset a user wishes to view or analyse (e.g. genotype data for a particular locus for a many samples, or a single chromosome for a few samples). We need efficient methods to securely stream more granular slices of datasets from either a local cache or from the central EGA; and Data encryption at rest - currently EGA datasets are encrypted and transferred to a user’s machine. However, to carry out any analysis of the data the user must first decrypt the files. We require standardised containers and interfaces around the human genotype and phenotype file formats so that data analysis tools can read/write encrypted files.

Similarly, within BBMRI simple interfaces are being developed to ease the interaction between researchers and many biobanks. One of such pilots is the BBMRI-ERIC negotiator that enables conversation between a group of biobanks and a researcher. Another pilot is implemented within BBMRI-NL named the 'Dutch national tissue portal' where researchers can post one request to all 50+ pathology labs to request relevant samples; this portal then processes the request and also supports the complete sample logistics.

Conclusion

This report survey of existing models and pilot designs for access to sensitive data. Existing models included use of anonymous/de-sensitized data extractions, data request workflows and access committees, closed analysis environments where data is brought together centrally for analysis and federated analysis where analysis travels to the data and data is kept locally. In addition new pilot designs are being proposed such as data discovery without disclosing identity (such as beacons), secure cloud extension across providers, BMS infrastructure in a box and federation of request workflows.

We recommend that in the coming years in CORBEL we will evaluate these, share security issues, discuss technical implementation challenges and develop shared materials and best practices, for example:

- Evaluation of beacon query extensions (with ELIXIR implementation study, 12/2017)
- Federated request workflows and piloting of local EGA in CORBEL (D6.5 month 48)
- OpenStack and EGI fedcloud extension (with EGI and EXCELERATE WP4)
- Web cloud extension (TranSMART, MOLGENIS, XNAT) in CORBEL (D6.5 month 48)
- Evaluation of BIBBOX for other BMS infrastructures (with BBMRI-ERIC)

References

[CW16] Compute workshop. Graz. https://drive.google.com/drive/folders/0ByT9CIvxqM5JLVJOclcyUHRZdzg

[CANHAM16] Steve Canham and Christian Ohmann (2016) A metadata schema for data objects in clinical research. Trials 17:557. DOI: 10.1186/s13063-016-1686-5

[DataSHIELD14] Gaye A et al (2014). DataSHIELD: taking the analysis to the data, not the data to the analysis. *International Journal of Epidemiology.*

[DIRECTORY] Holub Petr, Swertz Morris, Reihs Robert, van Enckevort David, Müller Heimo, and Litton Jan-Eric. Biopreservation and Biobanking. December 2016, 14(6): 559-562. doi:10.1089/bio.2016.0088.

[ELSI16] ELIXIR ELSI policy https://drive.google.com/file/d/0BxqILhwJcm1qME00QWRKUmtEVXM/view

[FAIR16] Wilkinson et al (2016) The FAIR Guiding Principles for scientific data management and stewardship. http://www.nature.com/articles/sdata201618

[GA4GHa] SECURITY TECHNOLOGY INFRASTRUCTURE. Standards and implementation practices for protecting the privacy and security of shared genomic and clinical data. https://genomicsandhealth.org/security-infrastructure-read-online

[Holub16] BBMRI-ERIC Security & Privacy architecture document. https://documents.egi.eu/public/ShowDocument?docid=2677

[EGA2015] Lappalainen et al., *Nature Genetics* **47**, 692–695 (2015)

[Murat15] Sariyar Murat, Schluender Irene, Smee Carol, and Suhr Stephanie. Sharing and Reuse of Sensitive Data and Samples: Supporting Researchers in Identifying Ethical and Legal Requirements Biopreservation and Biobanking 13(4): 263-270. http://online.liebertpub.com/doi/abs/10.1089/bio.2015.0014

[REMS1] REMS application leaflet https://confluence.csc.fi/download/attachments/36605742/REMS%20leaflet.pdf

[REMS2] REMS publication Download paper

[OECD07] OECD Best practice guidelines for biological resource centers. http://www.oecd.org/sti/biotech/oecdbestpracticeguidelinesforbiologicalresourcecentres.htm

[MIABIS16] Merino-Martinez Roxana, Norlin Loreana, van Enckevort David, Anton Gabriele, Schuffenhauer Simone, Silander Kaisa, Mook Linda, Holub Petr, Bild Raffael, Swertz Morris, and Litton Jan-Eric. Biopreservation and Biobanking. August 2016, 14(4): 298-306. doi:10.1089/bio.2015.0070.

[THLDA] https://www.thl.fi/en/web/thlfi-en/topics/information-packages/thl-biobank/researchers/sample-and-data-access

[TRYGGVE16] Successful demonstration of cross-border use of secure cloud, Tryggve summary report, https://wiki.neic.no/w/ext/img_auth.php/2/28/Crossborder-secure-cloud-tryggve-summary.pdf

Delivery and schedule

The delivery is delayed: Yes, see Grant Agreement Amendment AMD-654248-22
Reason: Change of personnel (twice) in critical roles.

Adjustments made

None

Appendices

Appendix 0: Terminology

We use the following terminology based on [ELSI16] and [GA4GH16] and [Holub16]:

- **Anonymous (or Anonymised) Data** is data that does not relate to an identified or identifiable natural person or to data that was personal data at the time it was collected but which, using best practices, has been rendered anonymous in such a manner that the data subject is no longer identifiable.
- **Application service providers** entities that provide software and other application services, such as web-based or mobile applications, for manipulating and analyzing data. See appendix 2.
- **Consent** means any freely given, specific, informed and unambiguous indication of their wishes by which the Data Subject, either by a statement or by a clear affirmative action (such as a signed document), signifies agreement to Personal Data relating to them being processed.
- **Data Access Committee (DAC)** means a designated group of individuals who are made responsible for reviewing applications and granting permission for access to access-controlled datasets. Decisions to grant access are made based on whether the request conforms to the conditions under which data is made available by the Service.
- **Data Provider** means the individual researcher or investigator or body of researchers or investigators that makes data available or submits data for access and use in the context of an ELIXIR Service.
- **Data service providers** means entities that provide data storage, protection, management, access, query, and transmission services and optionally ensure that data transmitted or uploaded to other destinations are qualified according to the specifications for both data and metadata constraints and semantics, as appropriate.
- **Data Subject or Individual** refers to an identified or identifiable natural person (individual) whose data are accessed (e.g. patients, donors or study participants).
- **Data Transfer Agreement (DTA)** means an agreement or contract made between a Data Provider and a Service Provider (i.e. when data is submitted to an ELIXIR Service – “data in”) or a Service Provider and a Service user (i.e. when an ELIXIR Service makes data available to researchers – “data out”) that governs the conditions under which the data is transferred and defines the rights of the contracting parties regarding future data usage. The DTA can take the form of general terms of service or terms of use.
- **Data User** means the individual researcher or investigator or group of researchers or investigators that accesses and/or uses data made available as part of an ELIXIR Service.
- **Genetic Data** means data relating to the genetic characteristics of an organism that have been inherited or acquired and which may provide unique information about the physiology or the health of that organism or individual.
- **Infrastructure service providers** means entities that provide technology resources and technical support for storing, managing, transmitting, and computing electronic data. See appendix 1.
- **Open/public access** means access is not restricted and the data is publicly available.

- **Personal Data** means any information relating to an identifiable natural person (Data Subject); an identifiable natural person is someone who can be identified with reasonable efforts, in particular by reference to an identifier such as a name, an identification number, location data, online identifier or one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that person. Genetic Data may be considered non-Personal as long as it does not fulfill the criteria of Personal Data.
- **Processing** means any operation that is performed with Personal Data, whether or not by automated means, such as collection, recording, organization, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available (including use or making available for research purposes), alignment or combination, restriction, erasure or destruction.
- **Pseudonymised Data** (also known as 'coded' or 'linked' data) is data that can only be connected to the Data Subject by using additional, separately kept information (a 'key') that would allow certain authorised individuals (e.g. the clinical team who collected the data) to link them back to the identifiable Data Subject.
- **Restricted access** means that access is constrained. There are various methods for restricted access such as rule based access (based properties of the user such as from what organisation she is) and committee based access.
- **Sensitive Data** means Personal Data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, genetic data, data concerning health or data concerning a natural person's sex life or sexual orientation.

Appendix 1: survey of "cloud" infrastructure providers

This appendix provides a short overview of existing e-infrastructure providers that provide facilities for sensitive data, or have shown interest to do so.

EGI, EGI Fedcloud and EGI DataHub

EGI is foundation with 24 countries + CERN + EMBL-EBI which provides the following solutions:

- Federated cloud - virtualized resources, cloud storage, support
- Grid for high throughput computing data analysis - computing, data mgt, storage
- Federated operations - manage operations while retaining local control (helpdesk, service registries)

Scale: 630K (HTC/grid) and 7k cores (cloud), 260PB of disk and 240PB of cloud

EGI uses 'ONEDATA' software which enables distributed and decentralized repositories.
Pluggable for various data types so existing data services can talk to it. Concept of 'projects' called 'data spaces'. High throughput clients for large data centers, metadata mgt, data migration/replication. Open data platform interactions: OpenAire (OAI-PMH), Web Gui, HTTP, REST, POSIX (using Fuse) and CDMI (community portal).In addition EGI provides a federated AAI (SAML, OIC, social ids) and AAI proxy service linking IdPs 'outside' EGI ecosystem.

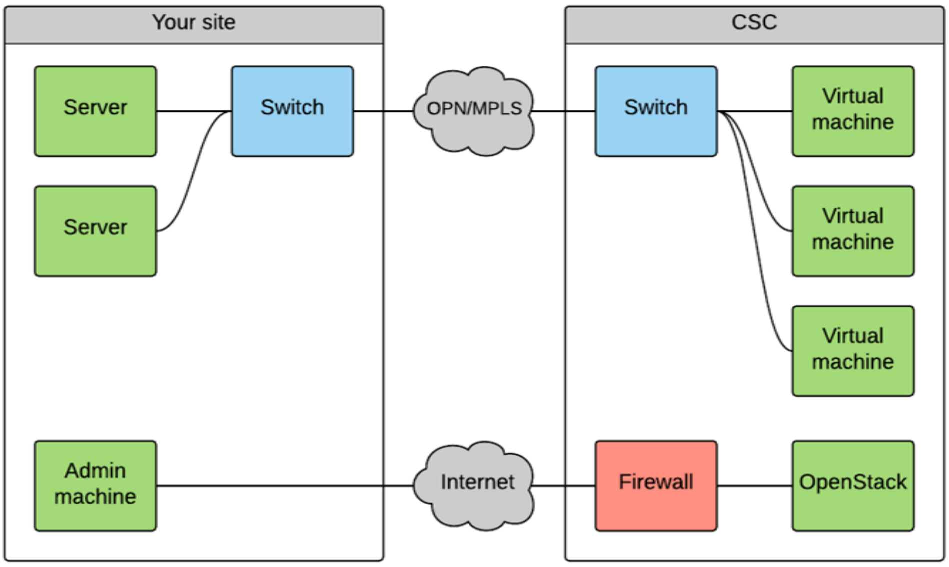
At the moment EGI doesn't support use cases on sensitive data (use cases on EGI either work with open data, or the data is anonymised before being moved onto EGI). However EGI is open for discussing with RIs how e-infrastructures could better support the handling of confidential data. EGI just started discussing this in the context of the new flagship project jointly with EUDAT and INDIGO-DataCloud for the EINFRA12(a) H2020 call.

Finnish secure ePouta cloud

The ePouta (https://research.csc.fi/epouta) is a cloud computing environment (Infrastructure as a Service, IaaS) developed for handling sensitive data. It is CSC's service and CSC is also ELIXIR Finland node. As funded by the Finnish Ministry of Education and Culture the service is free of charge for academic Finnish research projects. A pricelist is available for other researchers.

ePouta is a secure compute cloud that meets high information security regulations. It is suitable for all fields of science, but due to its high level of security it is particularly suited to meet the requirements of bioscience and human data research. The cloud was audited in 2015 for ISO/IEC 27001 the national governmental standard VAHTI 2/2010 raised information security level.

ePouta is based on OpenStack cloud middleware. OpenStack Volumes use centralised storage is based on CEPH and also using NFS services from NetApp is possible. There the virtual NFS server instances belong to the customer VLAN as well. Customer projects' networks are separated in L2 (VLAN) level. Thus, the cloud customer projects are in virtual private cloud.



The customer network connection is either via Optical Private Network (OPN) or Multiprotocol Label Switching Virtual Private Network (MPLS VPN) connection. The cloud middleware management Web UI/API access is firewalled so the customers IT-admins register their IP address via CSC's cloud and network support.

EMBL-EBI Embassy cloud

Embassy Cloud (http://www.embassycloud.org/) 'tenants' have direct access to the EMBL-EBI data, services and compute. This is a practical and cost-effective alternative to replicating services and downloading vast, public datasets locally. Tenants can access their workspace from anywhere in the world, reducing the need for capital investments in hardware and related operational costs.

A workspace consist of a dedicated, secure, private, virtual data centre hosted within our VMware installation; An allocation of CPU, RAM and storage resources for you to manage according to your project's needs; Internal and external network configuration of your space, specified by you, with simple firewall and VPN functions; Your host organisation is solely in charge of the systems administration within your Embassy Cloud space; You can access the EMBL-EBI catalogue of VM images or create and upload your own; You may be granted access to specified internal EMBL-EBI resources, or have a selection of EMBL-EBI databases mirrored into your space (e.g. ChEMBL, Ensembl, 1000 Genomes Project archive, Uniprot).

This Infrastructure as a Service is hosted in our Tier 3+ secure data centre in Hemel Hempstead, UK, and is logically outside the institute's local area network (LAN). Traffic from an Embassy Cloud workspace to EMBL-EBI's public data resources and services is retained within our own network infrastructure, but neither EMBL-EBI nor any other cloud client can access it.

AARC, AARC2

AARC (Authentication and authorisation for research and collaboration, https://aarc-project.eu/) is an European Commission funded 2-year project (5/2015-4/2017) which gathers together research communities and identity federation operators around Europe to develop the identity infrastructure necessary for authenticating researchers and helping the relying services to decide their access rights. AARC builds on the success of federated identity management and the research and education identity federations operated by national research networks.

AARC project itself does not operate services but develops and disseminates architectures and frameworks that can then be taken over by e-infrastructures (such as, EGI, EUDAT, GEANT or PRACE) and research infrastructures. Among other things, the project has developed and published an AAI blueprint architecture describing a reference model for an AAI (authentication and authorisation infrastructure) and its functionality. The project has also developed policy frameworks for incident response, level of assurance and data protection in distributed infrastructures. Pilots have been carried out to test the frameworks in practice.

AARC2 project (5/2017-4/2019) is going to be the follow-up project for AARC, widening the project to cover more pilots with research infrastructures and to disseminate the results of the AARC project. One of the pilots in AARC2 focuses on Life science AAI, a common AAI to serve the authentication and authorisation needs of life science research infrastructures and their users. The pilot will be carried out by the ELIXIR hub, BBMRI-ERIC, INSTRUCT and INFRAFRONTIER. CORBEL Work Package 5 will be the key stakeholder for the pilot, ensuring the pilot will serve the needs of all life science research infrastructures.

http://www.geant.org/Projects/GEANT_Project_GN4-1/deliverables/D9-2_Market-Analysis-for-Virtual-Organisation-Platform-as-a-Service.pdf

Tryggve project - IT services for sensitive data

The ELIXIR Nodes in Finland, Denmark, Norway and Sweden have joint forces in developing IT services for human data. With the support of NeIC - Nordic e-Infrastructure Collaboration - the work has been organised in a project, Tryggve. The project aims to provide researchers a trusted set of e-infrastructure capacities, software tools and common processes to transfer, store, share and process sensitive biomedical data in cross-border collaboration. Tryggve aims to produce solutions that are applicable for the life science communities beyond the initially participating countries. The project supports research projects through a use case programme.

The target of the project is to produce secure services that enable accessing sensitive data on secure computing environments regardless of their location, and allows either moving of the data or moving of the software tools to make combined analyses. The secure computing environments mentioned are located in each of the participating countries: TSD2.0 (University of Oslo), Mosler (National Bioinformatics Infrastructure Sweden), ePouta secure cloud (CSC, Finland), and Computerome secure cloud (Technical University of Denmark). Each of the systems provide users with a secure virtual computing environment, but the implementation and offered services of the systems vary somewhat

Some key outcomes of the Trygve project so far include advances in secure data transfer [1], connecting secure clouds across countries [2], assessment of AAI options [3,4], and studying the use of portable software environments in the form of Docker Containers [5]. Several other outcomes of the project are documented in the reports available at project web site (direct link): https://wiki.neic.no/wiki/Tryggve_Reports_and_Materials

The outcomes of the Tryggve project are supporting the service development of the participating ELIXIR Nodes, and it is planned that they can become parts of their service offering. Further, the Tryggve partners are currently preparing for second project stage, Tryggve2, which is planned to further develop the outcomes to operative services, to work in close connection with data owners, such as biobanks, and to include major research and infrastructure use cases as integral part of the project. Thus there exists a clear plan for sustaining successful results of the Tryggve project.

Project web site is https://wiki.neic.no/tryggve

References:

[1] SFTP Beamer, https://github.com/neicnordic/sftpbeamer
[2] Successful cross-border use of secure cloud, https://github.com/NBISweden/Knox-ePouta
[3] Tryggve Processes for Authentication, https://wiki.neic.no/w/ext/img_auth.php/5/50/150831-D9-Authentication.pdf
[4] Tryggve Processes for Authorization, https://wiki.neic.no/w/ext/img_auth.php/d/dd/Tryggve-D9b-ProcessesForAuthorization.pdf
[5] Software Provisioning Inside a Secure Environment as Docker Containers, https://wiki.neic.no/w/ext/img_auth.php/0/05/Galaxy-Docker-Report.pdf

SURF research cloud

SURF is the Dutch national collaboration of academic compute centers. It provides the national network and supports local compute facilities in their local services. One of these is the SURF HPC cloud (https://www.surf.nl/en/services-and-products/hpc-cloud/index.html), available as a Infrastructure as a Service to researchers. The infrastructure ranges from a single work station to a complete cluster (including GPU and high memory) and can be expanded to suit your needs.

A national Research Cloud, also as a pilot of the European Open Science Cloud, is in development. The Research Cloud is a vision which aims to provide a national service hub for research which enables the collaborative working of National and European Institutions, Universities, UMC's & Industry in a federated and secure manner to accelerate research.

It will include a cloud management layer combined with collaborative management technology that facilitates the hosting of numerous application architectures that may be offered as discrete or shared services. It will also offer connectivity, seamless integration and cloud bursting opportunity where appropriate with:

- Federative High Performance Computing (HPC) infrastructures (National and International)
- Public cloud providers such as Amazon AWS, Google Cloud and MS Azure
- Online, Nearline and Offline storage solutions that enable collaborative working with trusted research environments in a secure and responsible manner
- Private Cloud & Local infrastructures
- The European Open Science Cloud

A key element of the management platform will be role based access that will enable management views that are relevant to the entity and subgroups concerned. The integrated accounting and reporting, means that institutional management will be able to easily obtain insight into their infrastructure and resource usage that is connected to the hub.

In addition the Research Cloud aims to integrate complete Research Data Management (RDM) workflows from - concept to analysis to publication & archiving, data processing pipelines, Trusted Third Party (TTP) authorisation workflows and best practice tooling for Researchers.

CERN / European Helix Nebula Science Cloud

EHNSC is a project for procurement of commercial cloud services. E.g. 3k VMs over 45 days, Azure 4.8k, Deutche Borse cloud exchange, IBM softlayer, T-systems (also included data intensive workflows, 500TB storage). It uses a hybrid cloud model - mixing commercial cloud with locally installed services. Lessons learned: to effectively use this you need reliable and performant network. There advantageous to use Geant, also in light of ingress costs and ability for providers to access whole network of research organizations around Europe. Major challenges: Disrupting the way IT resources are provisioned; In-house, public e-infra and commercial cloud are not integrated; Org/financial models may not be appropriate; Procuring cloud is also matter of skills and educations; Legal impediments. See recommendations in www.picse.eu/roadmap.

Now started pre-commercial procurement to a single tender (HNSciCloud): Hybrid cloud platform, IaaS level services (range of vms, os, message queues, network, storage (pb), cpu, backup); Connected to Geant (high end network, federated AAI); Service payment models (accounting, pay per usage). Currently ELIXIR, BBMRI-ERIC and Euro-BioImaging on board and CORBEL cluster. At pilot phase opportunity for other RIs to join. Subsidised use of commercial cloud services is expected model to be supported. Not used for long term archiving of storage, but more to relatively quickly handle temporary peak loads. Long term archiving done with inra from publicly sector infra. Sometimes you may need infra close to an instrument ("online"), etc. So local IT infra funded by the public sector is not anticipated to vanish. Once "online" data preprocessing is done, data can go to an "offline" HPC site for further analysis. Only this offline data is considered for (partial) offloading to a commercial cloud provider.

Appendix 2: survey of application service providers

This section surveys existing applications for data access used and/or developed within the context of the BMS infrastructures to facilitate access to sensitive data. This survey provides emerging models and pilot designs of solution elements.

European Genome-phenome Archive (EGA)

The initial EGA model of centralised submission and distribution data is not applicable to studies where the local ethics agreements stipulate that the identifiable data must remain within a jurisdiction, e.g. national healthcare or regional biobanks. In these situations, any researcher wishing to access or study these datasets would be required to move their computational pipeline to the data. For example, as part of the FNIH funded Type 2 Diabetes Knowledge Portal (http://www.type2diabetesgenetics.org/) we are building a European federated portal at EMBL-EBI so that relevant datasets submitted to the EGA can be queried and analysed remotely via the main knowledge portal of the project.

Scalable high performance computing is increasingly being provided by public and private cloud providers. A researcher that has been granted access to EGA datasets might need to use a third-party cloud infrastructure to analyse the data. In this scenario services such as local and federated data access, identity management, secure delivery from/to the cloud environment, and secure storage of the data at rest are required. This is a new service that allows authorized third-party services to programmatically check compliance with the current user data access authorizations from the ELIXIR coordinated repositories such as the EGA database each time user accesses a file in the cloud or other remote service. A first planned project using EGA data within the private, secure, cloud at CSC in Finland will provide our reference implementation.

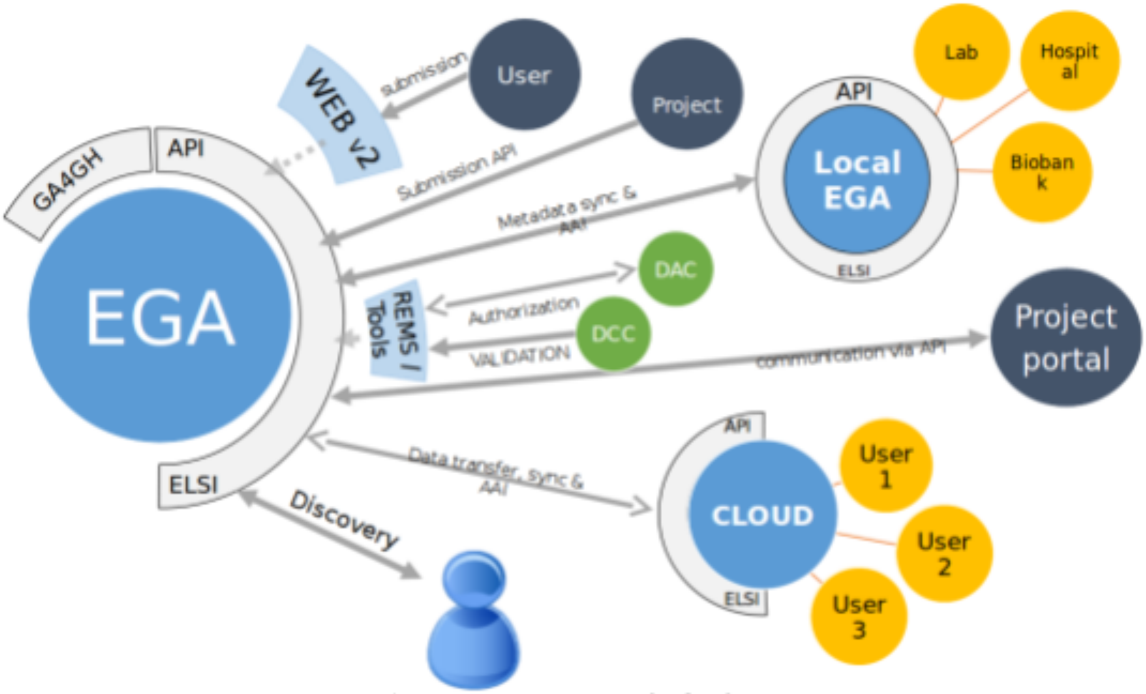


Figure 2: A proposed model of EGA federation

High level requirements for EGA federation

- Federated authentication and authorisation - currently EGA maintains its own single centralised password based authentication and DAC controlled authorisation. It will be necessary to enable third party platforms to authenticate EGA users and verify authorisations, e.g. enabling access to data cached in a cloud.
- Identity management - currently EGA account identities are not linked to any other user identity, e.g. an ELIXIR identity, REMS.
- Secure streaming of subsets of data - currently EGA data is delivered as very large encrypted data files, no matter what part of a dataset a user wishes to view or analyse (e.g. genotype data for a particular locus for a many samples, or a single chromosome for a few samples). We need efficient methods to securely stream more granular slices of datasets from either a local cache or from the central EGA.
- Data encryption at rest - currently EGA datasets are encrypted and transferred to a user's machine. However, to carry out any analysis of the data the user must first decrypt the files. We require standardised containers and interfaces around the human genotype and phenotype file formats so that data analysis tools can read/write encrypted files.

Global Alliance for Genomics and Health (GA4GH) initiatives and EGA

1. Streaming API task team
    a. The Global Alliance Data Working Group Directory and streaming API Task Team's goal is to bridge the gap between existing local file based access and processing of sequencing reads and variants, and remote web based API access. The team is developing a protocol for bulk streaming of read sequencing data, with an initial focus on using existing next-generation sequencing file formats (SAM/BAM/CRAM) and standard internet protocols for transfer, with a future path to others. Initially,

the team created a sequencing reads streaming specification that was implemented by seven different clients and five different server groups (http://samtools.github.io/hts-specs/htsget.html). Beyond this initial specification and implementations, a comprehensive security model around data streaming. Within this team, the EGA has begun a pilot implementation for secure streaming human identifiable data (Figure 3).

2. File formats task team
    a. This team's primary role is for the maintenance and development of the main file format specifications used by the genomics community, e.g. for raw sequencing reads the SAM/BAM/CRAM formats, and for genotypes the VCF/BCF formats. Critically this group has representatives from the primary library implementations of these specifications, and require all changes to be agreed. Recently, the group has been created proposals for a standardised container for encrypting these files at rest whilst maintaining key functionality such as random access.
3. Software security task team
    a. The GA4GH Security Working Group (SWG) leads the thinking on the technology aspects of data security, user access control, and audit functions, working to develop or adopt standards for data security, privacy protection, and user/owner access control. The role of the software security task team is to take the high-level policy documents and work with development groups to create working code that conforms to these security principles.

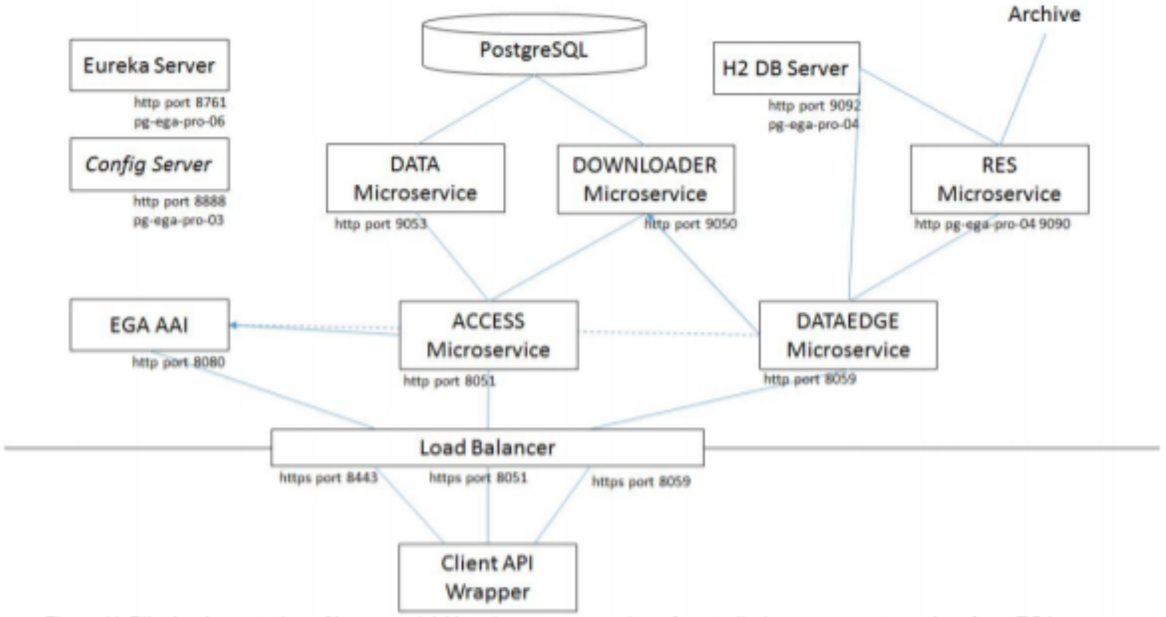


Figure X: Pilot implementation of integrated AAI and secure streaming of controlled access genotype data from EGA

Future Work

In summary, the EGA use-case is well-placed to act as a close engagement point between the GA4GH initiatives around secure delivery of human genotype data and corresponding aims in CORBEL (Task 6.3). In GA4GH, we are involved in multiple task teams (listed above) and will provide implementations of the GA4GH security framework. We will work with CORBEL partners such as the BBMRI MOLGENIS biobanking platforms and imaging resources such as Euro-BioImaging to provide specific working implementations, whilst ensuring that our work is generalisable and not use case specific.

tranSMART

tranSMART (http://transmartfoundation.org/) is a data integration, sharing, and analysis platform for clinical and translational research. It allows users to search, view, and analyze data through a web interface, thereby allowing easy access to explore such data from multiple domains at study level.

As a knowledge management platform, it enables scientists to develop and refine research hypotheses by investigating correlations between genetic, phenotypic and clinical data and assessing their analytical results in the context of other data sets. tranSMART's capacity to store and integrate data from multiple domains is leveraged by the many tools it connects to, such as R, Galaxy, and Spotfire (but Spotfire is not open source; subject to additional license).

tranSMART was initially developed by Johnson & Johnson, and released as an open source project in 2012. It has since successfully been deployed in a number of large pharma companies, such as Johnson & Johnson, Sanofi and Pfizer, and is further developed in numerous publicly funded and public-private consortia, such as IMI eTRIKS, IMI EMIF, IMI Translocation and CTMM-TraIT. Further development and community building is coordinated through the tranSMART Foundation.

MOLGENIS

MOLGENIS is a open software suite for scientific data (http://www.molgenis.org). It consist of a flexible core platform that can be configure to support any science, i.e., configure data structure, analysis scripts and user interfaces. In addition, there are emerging tool suites built on top of this platform mainly in the area of biobanking and genomics, i.e., best practice data models, analysis produres and data analysis and integration 'apps'

MOLGENIS core is a flexible software platform (PaaS-like) that enable end-users to fully configure its data structure, analysis capabilities and user interfaces. The data model can be configured by upload of data in any data model using a simple spreadsheet format as well as specific formats such as VCF. In addition users can upload analysis procedures using R and python scripts. Finally, users can completely change the look and feel by uploading custom styling or even complete web applications using a plug-in mechanism.

On top of this platform several tool suites are emerging. MOLGENIS is used to create biobank catalogues on the level of collections, data items, and samples. Examples include the BBMRI-ERIC biobank directory, the BBMRI-NL catalogue and the PALGA pathology public database. Another large application are of MOLGENIS is complex data integration introducing tools like BiobankConnect, SORTA, and FAIR data point tools. In addition, MOLGENIS is being used for genomics data analysis including applications in the clinic, introducing capabilities to annotate and visualize genomics data using genome browsers and annotation tools such as GAVIN. Otherwise, MOLGENIS is used for many data management and sharing tasks such as sharing DNA variant classifications between Dutch genomics labs, sharing Energy research data in the north of the Netherlands, and various patient, mutation and data registries.

MOLGENIS development has been mainly sponsored by academic labs and consortia such as BBMRI-NL, BBMRI-ERIC, RD-connect, UMCG, LifeLines, UMCU, LUMC, Panacea, BioSHaRE, EnergySense.

XNAT image sharing

XNAT (https://www.xnat.org/) is an open source imaging informatics platform developed by the Neuroinformatics Research Group at Washington University. XNAT was originally developed in the Buckner Lab at Washington University, now at Harvard University. It facilitates common management, productivity, and quality assurance tasks for imaging and associated data. Thanks to its extensibility, XNAT can be used to support a wide range of imaging-based projects.

XNAT enables data access via a website (manual upload and download), via the DICOM protocol and via an application programming interface (API), which makes it flexible. Furthermore, XNAT stores not only the images, but also image-derived information, such as annotations and processed versions of the images. It is therefore of interest for the more advanced, technically oriented researchers, and for large studies which require automated image analysis.

To enable storage of medical imaging data in a central archive, we have created a national imaging platform: http://www.ctmm-trait.nl/trait-tools/xnat [SK1]. This platform is built on top of the open source XNAT software (www.xnat.org) [SK2]. XNAT facilitate secure storage and management of medical imaging data and image-derived data like segmentations and quantitative radiomics features. Images and derived data can be accessed (provided the user is authorized to do so) both via a graphical interface and via a programming interface, facilitating automated batch processing. The TraIT XNAT service is used successfully by several large multi-center studies. To transfer the data from the local PACS to the central imaging platform, while applying proper de-identification/pseudonymisation, we use the Clinical Trial Processor (CTP) software, which is endorsed by the Radiological Society of North America (RSNA) and also recommended by TraIT. The TraIT XNAT service was established thanks to funding of the Dutch CTMM TraIT project and the FP7 BioMedBridges project. The service is currently maintained by funding of BBMRI-NL and Lygature TraIT. A fair-use policy has been adopted: http://www.ctmm-trait.nl/work-packages/work-package-2-biomedical-imaging/trait-bmia-features-and-services/pricing-model.

In the CORBEL project (WP3), we have developed pipelines for automatic processing of imaging data stored in the XNAT archive. These software pipelines can run on a compute cluster and connect directly to XNAT via the REST-based application programming interface (API) to retrieve the data, and to upload the results. Also, a standardized data type for storage of the resulting image-derived data (processed images, segmentations, imaging biomarkers) has been developed. We are currently configuring this data type on various XNAT instances, among which the TraIT XNAT service.

[SK1] XNAT.bmia.nl: Klein S, Vast E, van Soest J, Dekker A, Koek M, Niessen W: XNAT imaging platform for BioMedBridges and CTMM TraIT. Journal of Clinical Bioinformatics. 2015, 5(Suppl 1): S18 http://jclinbioinformatics.biomedcentral.com/articles/10.1186/2043-9113-5-S1-S18

[SK2] XNAT: Marcus DS, Olsen T, Ramaratnam M, Buckner RL: The Extensible Neuroimaging Archive Toolkit (XNAT): An informatics platform for managing, exploring, and sharing neuroimaging data. Neuroinformatics. 2007, 5 (1): 11-34. http://link.springer.com/article/10.1385%2FNI%3A5%3A1%3A11

ELIXIR Authentication and Authorization Infrastructure (AAI)

The ELIXIR Authorization and Authentication Infrastructure (AAI) provides a single sign-on to all ELIXIR services (1). The AAI services allow researchers to use their academic or social media identities while linking these to a unique identifier used for communicating the identity and authorization attributes to the ELIXIR services. Access can be managed by assigning individual to a group (for example to support access to ELIXIR intranet) or link ELIXIR identity to a particular dataset allowing user to access data.

(1) https://www.elixir-europe.org/services/compute/aai

Resource Entitlement Management System is an open source application used for brokering access to sensitive data (REMS1). It is part of the ELIXIR AAI services. In a typical use case researcher(s) apply access to the data using an electronic application form within REMS. A submitted application is then reviewed by a Data Access Committee that oversee data access. Approved access rights are available to authorized services storing the data. Researcher(s) with appropriate authorization log into the service to either download data to their local system or analyse data using provided resources (REMS2). The ELIXIR EXCELERATE WP9 integrates REMS as part of the EGA workflows for managing data access rights.

DataSHIELD federated analysis

DataSHIELD (http://www.datashield.ac.uk/) [DATASHIELD14] is an infrastructure and series of R packages that enables the remote and non-disclosive analysis of sensitive research data. Users are not required to have prior knowledge of R. Analysis requests are sent from a central analysis machine to several data-holding machine storing the harmonised data to be co-analysed. The data sets are analysed simultaneously but in parallel, linked by non-disclosive summary statistics. Analysis is taken to the data – not the data to the analysis.

DataSHIELD is implemented entirely via free, open source software: at heart, a modified R statistical environment linked to an Opal database deployed behind the firewall at each data-holding organisation. Analysis is initiated in a standard R environment at the analysis machine, with communication between the analysis and data-holding machines controlled via secure web services. The same infrastructure and approach may also be used with just one data source – this is then referred to as “single site DataSHIELD” providing a freeware-based approach to creating a secure data enclave.

Perun

Perun is an application which consists of several interconnected components. Core part of the Perun is used for managing users, virtual organizations (projects), groups, facilities and resources. Additional components use information managed by core part of the Perun and adds additional functionality. Detailed view on the Perun is available here.

Nordic human data service offering

The Tryggve project operates an open call for use cases on cross-border use of sensitive data in research. The call is intended for research teams aiming to utilise biomedical sensitive data from several countries, and who are in need of secure IT systems and services. Use cases can be proposed at any time, and the approved ones will get support and access to secure infrastructure free of charge, supported by NeIC (Nordic e-Infrastructure Collaboration) and ELIXIR Nodes of Finland, Denmark, Norway and Sweden. Currently the use case support is limited to research conducted mainly in these four countries.

A Tryggve use case is entitled to access to the secure storage and computing environments affiliated with the project. There are Secure remote desktop systems (TSD[1] at University of Oslo, and Mosler[2] in Sweden), as well as secure cloud IaaS (ePouta[3] cloud at CSC, Finland and Computerome[4] secure cloud in Denmark). In addition to accessing the systems, experts from the project team offer support and tools for secure data transfer, software installations, accessing external archives, and even support in meeting legal and ethical requirements.

Future work aims at extending the use case programme to wider so-called infrastructure use cases, in which Tryggve experts team up with data owning organizations to jointly develop tools and processes for bringing data accessible to research. The aim would be to simplify the process of accessing data in a secure and powerful data processing environment.

[1] http://www.uio.no/english/services/it/research/storage/sensitive-data/index.html

[2] http://nbis.se/infrastructure/mosler.html

[3] https://research.csc.fi/epouta

[4] http://computerome.dk/