945358-BIGPICTURE
Central Repository for Digital Pathology

WP2 - Infrastructure and database hosting

D2.02 Report on the pilot system establishment

Lead contributor	2 - LIU
	Joel Hedlund - joel.hedlund@liu.se
Other contributors	5 - CSC 6 - UU 15 - ULIEGE 37 - JANSSEN 38 - BAYER

Document History

Version	Date	Description
V0.1	23 Sep 2021	Preview for consortium orientation
V0.2	30 Sep 2021	"First draft" for PMO review
V0.3	18 Oct 2021	Draft for managing board approval
V1.0	22 Oct 2021	Final version

Abstract

This report describes the outcome of the establishment of the repository infrastructure pilot system that was carried out in BIGPICTURE Task 2.3. The intent is that this report will provide users with a baseline understanding of the status and functionality of the pilot system, such that users can know what to expect, and to enable productive use of the system in Task 2.5 Pilot operation as well as effective and well prioritized iterative improvement in Task 2.6 Iterative development.

This report contains a user-facing overview description of the pilot system, as well as a description of the establishment process; challenges, mitigations and resolutions, and outstanding issues. It also makes recommendations for follow-on work in Tasks 2.5 and 2.6.

These tasks mostly concern collaborative development and adaptation of existing open source software and services, so to provide some background understanding of the methodology employed, this report also provides a brief overview description of methods and results from Task 2.2 Establish development platform, including such topics as version control and issue tracking, even though such topics are not anticipated to be of general high interest outside of WP2.

In Task 2.3 we established a pilot system based on initial requirements gathered in Task 2.1, to allow pilot data taking and inference in Task 2.5. This included software components and necessary hardware for system integration and testing, as well as planning and implementing production practices and documentation. Here, we established the central repository, and one satellite node based on Federated EGA technologies[1], with federated interoperable access enabled by existing identity and trust federations using the ELIXIR AAI (Authorization and Authentication Infrastructure)[2] and the REMS (Resource Entitlement Management System). The central repository was constructed in national academic compute centers in Finland and Sweden, enabling data mirroring cross-border where possible for increased availability and disaster resilience. The federated node, a satellite to the central repository, was set up as a demonstrator of technical ease of deployability, and of technical readiness of the federated system, as a preparation for adding more federated nodes in the future. The central repository is the primary resource for project data. The design provides a private execution environment, where users can log in using ELIXIR AAI and access data to entitlements in REMS. This design will be used as the basis for development of a private execution model, which will be refined in Task 2.6, and is intended to enable morphological data mining using algorithms developed in WP4 over the course of the project. Basic discoverability features were implemented based on straightforward search and filtering of identified public metadata features, and export of compatible metrics for advertising available datasets through the BBMRI Directory.

The pilot system will be enhanced during the 15 months period of piloting. There will be new interfaces implemented in collaboration with WP3 and WP4, implementing various discovery queries, improving the data access granting allocation based on the planned policy models and related metadata, implementing a system to run the private inference execution also to utilise those output for the discovery metadata. Together with WP4 we will implement a platform to run Cytomine, which will also provide annotation metadata for data. The data repository capacity will be increased on demand. Security assessment and service operations improving is a continuous work. For example reporting and monitoring will be done.

[1] https://ega-archive.org/federated

[2] https://www.elixir-europe.org/services/aai

Methods

The repository infrastructure is now established in a way that enables data mirroring between data centers at the BIGPICTURE partners Uppsala University and CSC. The services are based on sensitive data technologies principally developed by ELIXIR and that are in current production use at CSC and Uppsala University, which also coordinate the Finnish and Swedish nodes of ELIXIR, respectively. These technologies were originally developed to address needs in genetics. A large part of the development effort in Work Package (WP) 2 therefore concerns adapting these technologies to needs arising in BIGPICTURE and digital pathology, without affecting operations of production sensitive data services for genetics, while at the same time striving to keep development efforts as closely connected as possible, so that improvements in either effort can benefit both.

The establishment comprised development activities in the following areas, which are described in further detail in separate sections below:

1. Development platform: Establishing a code development collaboration platform, connected to Task 2.2 - Establish development platform (M1-M3).
2. Deployment environments: Establishing pilot system deployment environments.
3. Data mirroring: Developing data mirroring technologies.
4. AAI: Integration with ELIXIR Authentication and Authorization Infrastructure (AAI).
5. Entitlement management: Deployment of dedicated Resource Entitlement Management System.
6. Metadata: Extending the metadata model to fit initial needs identified by BIGPICTURE.
7. Sensitive data storage: Deploying sensitive data storage based on Federated EGA technologies.
8. Discovery: Deploying Beacon for search and discovery.
9. Private inference execution model: Developing a model for executing untrusted algorithms against sensitive data to produce performance metrics, without disclosing confidential information from either.
10. Discovery export: Export of compatible metadata for advertising in external directories.
11. System security

The outcome of the efforts in these areas are described in corresponding subsections to the Results section.

1. Development platform

The repository infrastructure development platform is used to coordinate efforts among the software developer teams in WP2, but for openness and transparency, it is also visible to others, who can use it for example to see status and follow development progress. For that reason, an overview description is provided in this report.

The selection of development processes and tools in Task 2.2 - Establish development platform (M1-M3) was carried out taking into account the choices made in existing collaborations in Nordic and European ELIXIR projects, where the core services that the BIGPICTURE repository is based on are currently being developed as open source software. Priority was given to finding a solution that could 1) facilitate code development, 2) ensure alignment with upstream efforts, and 3) allow independent development of original and specialised systems, while 4) at the same time enabling all efforts to benefit bidirectionally from improvements made in either initiative.

There are multiple genomic information related projects going on supporting the development of the solutions for sensitive data. GitHub has been used historically and currently in these efforts, and has proven to be a highly effective tool for coordinating the tasks and managing the source code repositories with version control and issue tracking etc.

In the course of Task 2.2 - Establish development platform (M1-M3), no indications were found as to why the needs in the BIGPICTURE repository infrastructure development would not be optimally served by

continuing the use of the same platform that has adequately served development efforts of the core services. GitHub was therefore chosen as the development platform also for the BIGPICTURE repository infrastructure development efforts, in order to support and strengthen the collaboration across the teams.

2. Deployment environments

The system deployment consists of microservices implementing the data management services which were deployed on Kubernetes[3] container clusters both at CSC and in Uppsala University IT systems. The details are in the results section below.

The private inference execution platform is being piloted in winter 2021. This is a novel service requiring piloting and after that we know which kind of hardware and services are needed.

3. Data mirroring

Data mirroring technologies are initially using the normal data submission pipeline that is built to handle all data submissions to the BIGPICTURE repository, meaning that no special technology will initially be developed for the sole purpose of data mirroring. The intention is that this will make the overall system design simpler and therefore easier to make secure.

4. AAI

The key AAI functionality is to let users authenticate using their home organisation identity provider, which is easily done for academic organisations that can leverage existing identity and trust federations such as eduGAIN. In this way, BIGPICTURE does not need to set passwords and does not need to implement its own identity management system.

ELIXIR AAI is a service provided by the European infrastructure for life-science (ELIXIR). BIGPICTURE uses this service to handle authentication and authorizations of users accessing services and resources such as data. A service looking to integrate with ELIXIR AAI must first register itself as a relying party with the ELIXIR AAI service. This can in a first stage be utilised as a test environment, without need to fulfil all the requirements. This allows testing and development with a user account assigned to the ELIXIR AAI environment.

There is ongoing work outside of BIGPICTURE to create a new AAI service under the name of Life-Science AAI, likely as a successor to ELIXIR AAI, that aims to support a much wider range of life-science infrastructures also outside of ELIXIR. However, as Life-Science AAI is not yet mature enough for production use in BIGPICTURE, we are using ELIXIR AAI in the interim. We intend to migrate BIGPICTURE service integration to Life-Science AAI at such a point where this new service enters production mode, and has reached sufficient stability and has a broad enough feature set to satisfy BIGPICTURE needs.

5. Entitlement management

Resource Entitlement Management Service (REMS) is based on the existing development work done for earlier projects[4]. In BIGPICTURE the technical work involved integration to the data ingestion.

[3] https://kubernetes.io/

[4] https://www.csc.fi/en/rems-kayttovaltuuksien-hallintajarjestelma

CSC's work for REMS has been supported by the Ministry of Education and Culture of Finland and by Academy of

There are admin users who can create the acceptance workflows and process manually in the web portal. The organisation who submits the data defines what is considered necessary information in requests to access the dataset, however WP3 and WP5 are working on the harmonisation of the data access policies, in Task 3.12 - Honest broker requirements gathering (M1-M9), Task 3.13 - Honest broker pilot system development (M10-M15), Task 3.14 - Honest broker mechanism pilot testing and operation (M16-M24), and Task 5.2 - Investigating and developing the legal and contractual framework for a sustainable BIGPICTURE repository (M1-72). These topics are also continuously addressed cross-sectionally by the Honest Data Broker Task Force and the General Data Protection Regulation (GDPR) Task force.

Then acceptance to access a dataset is stored in ELIXIR AAI, which information the data management middleware is able to use when the user of the data authenticates to ELIXIR AAI in order to read data.

Later in the project (2022) we plan to introduce data access ontologies to assist in the data access granting process. This will be done in collaboration with WP5 in Task 2.6 - Iterative development (M10-M71) and Task 5.2 - Investigating and developing the legal and contractual framework for a sustainable BIGPICTURE repository (M1-72). Also the data submission may send attributes, which allows automatically to add resource allocation related information to REMS.

6. Metadata

The Metadata Task Force (WP3) is specifying the metadata model with mandatory and optional fields, some of which may vary between the dataset types. Also the sensitivity level may vary. A high priority in this work has been to ensure a high degree of alignment with how the European Genome-phenome Archive (EGA) repository has been defining the information model.

The middleware itself is using metadata for data access management (dataset identifiers), producing discovery information (search index).

Metadata will evolve and thus the middleware and especially the discovery database need to evolve, taking into account that later revisions may be very expensive, as they may require data contributors to revise submitted data to conform to the updated standard. Together with data submission tasks, we are investigating what could be optimal interfaces including protocols and file formats for sending the metadata. Adding annotations for an image will be investigated especially with WP4, and Task 4.5 - Develop a technical implementation of the open digital slide, annotation, and meta-data format (M1-M36). This topic is also being addressed cross-sectionally in the Metadata Task Force.

BIGPICTURE will strive to make extensive use of Digital Imaging and Communications in Medicine (DICOM) as an interchange format, however we are aware that the DICOM standard for pathology is not yet finalized, and thus expect to have to adapt our implementation as the standard matures. We are also aware that not all metadata that BIGPICTURE needs to handle is representable in DICOM (notable exceptions include as access rights, DataCite DOI information, cohort statistics, or - crudely expressed - "anything above the patient level") and will thus need to be handled using other formats. Some of these pertain to existing services that already have established formats and in some cases even executable validators. One example is the DataCite service, which we will use to mint DOIs (Digital Object Identifiers) for datasets, and which accepts metadata for example in schema.org LD-JSON or XML. In such cases we aim to use these formats and validators as-is, with as little interference as possible. For other less well-delineated types of metadata, we aim to make use of XML formats, in order to keep in alignment with the core services (such as Federated EGA) which have existing XML schemas which could be extended or adapted to fit.

Finland grants 271642 and 263164 to construct Biomedinfra, the Finnish consortium for ELIXIR, BBMRI and EATRIS ESFRI.

A draft BIGPICTURE metadata standard release schedule for the near-term has been discussed in the Metadata Task Force:

- **M9 v0.1.0**: Initial de-facto baseline version for iterative refinement in Task 2.6 - Iterative development (M10-M71).
- **M18 v1.0.0**: First useful version for real clinical submissions, envisioned to be synchronous with Milestone 8 - Delivered clinical dataset via the submission portal to the BIGPICTURE repository (M18).
- **M24 v2.0.0**: First "battle-hardened" version, incorporating lessons learned from real-world use, envisioned to be synchronous with D2.05 - Report on pilot operations (M24). After this the repository enters Task 2.8 - Large scale archive operations (M25-M72).
- **M30 v2.1.0 (if needed)**: The timing of Milestone 9 - First non-clinical dataset (M30) could allow for a minor update at this point in time, for necessary changes (if any) majorly affecting only non-clinical datasets.
- **M?? v3.0.0 (if needed)**: Based on findings during or around Task 2.8 - Large scale archive operations (M25-M72) zero or more major revisions could become advisable, however the benefits of any major update would have to be weighed against the costs of updating already-submitted datasets to conform to the new version of the standard.

7. Sensitive data storage

The sensitive data storage was deployed based on Federated EGA technologies. The Swedish data center uses an S3 based hardware system operated by SafeSpring and procured through SafeSpring, through a framework agreement made available to Swedish Higher Education Institutions (HEI) by the Swedish University Network (SUNET). In the Finnish data center the submitted data is stored on an existing Network File System (NFS) storage service. Also an object storage system exists[5] and use of this system will begin in Q1/2022. The large storage capacity investments will begin at a later stage and as necessary, probably in the end of 2022 or 2023, so that the hardware lifetime can be matched to the project lifetime. Hardware renewal has not been budgeted in this project.

8. Discovery

Discovery metadata parameters are those metadata parameters that can be used for searching the repository for datasets of interest, and for public advertisement of available datasets. Discovery metadata parameters can be classified as either public, registered, or controlled, which signifies that the corresponding parameter is approved for showing to 1) all public users, or 2) only to registered users, or 3) only to users who have already obtained ethical and legal controlled access to the dataset in question.

GA4GH Beacon[6] was chosen as the discovery service to use in BIGPICTURE, as it supports distributed queries against also registered and controlled discovery metadata parameters, and can report matches without disclosing any confidential information from the matching datasets.

Deploying Beacon for the BIGPICTURE pilot system was not straightforward, since it is designed with only genomic data in mind. In other words, the data model (schema) and the query logic is tailored for that. Support for queries against pathology data will be developed as part of Task 2.6 - Iterative development (M10-M71). For the pilot system, "number of images in the dataset" was chosen as the first public discovery metadata parameter. Support for more advanced queries will be enabled iteratively, as the BIGPICTURE metadata standard matures.

[5] https://research.csc.fi/-/allas

[6] https://beacon-project.io/

Therefore, in collaboration with the Metadata Task Force (WP3 especially), we have worked to define and classify what metadata fields are to be used for discovery queries, with data from pathologists, and have defined a database schema which is external to the metadata database itself. The technical implementation of the metadata schema and setting up the framework to query against the discovery database was done by CSC.

During the course of Task 2.6 - Iterative development (M10-M71) we aim to implement more queries based on specifications by the Metadata Task Force at such a time as further as definitions and classifications become available, at which point we will iteratively update the information model accordingly and use the authorised metadata access if needed. Discovery database information can be classified for public, registered users and data access granting requiring data. For the pilots, however, the latter information was not yet implemented.

9. Private inference execution model

One of the purposes of the pilot system is to provide a secure sandboxed environment to enable the development of a BIGPICTURE private inference execution model. The intended purpose of this model is to allow arbitrary AI algorithms to be used to make inference against arbitrary datasets, in a way such that no sensitive details of neither data nor algorithms become accessible to either rights-holder or third parties. Inside the secure sandboxed environment that is provided by the pilot system, it is possible to use realistic and also potentially sensitive data in order to facilitate this development.

This development will result in a model for robustly defining inputs and outputs, and resources available for the computation, such as hardware profile and memory and temporary storage availability and limits, network connectivity (if any; unlikely) and presence and version of code libraries, drivers and other supporting software. This private inference execution sandbox is intended to enable discovery metadata production, algorithm validation, and other usage patterns like morphological data-mining.

10. Discovery export

WP2 has initiated the discussion with BBMRI-ERIC, on how BIGPICTURE can export compatible metadata for advertising available datasets in the BBMRI directory. bio.tools database can point to BIGPICTURE repository as well, but it makes sense to add there when the data is opened for the research outside the consortium. Thus, it is too early to implement this.

11. System security

Security has many aspects. Traditionally, aspects such as confidentiality, integrity and availability are considered. In human health data the sensitive human data is in key position and thus data protection impact assessment is planned together with WP5. WP2 has also made preparations to carry out outsourced penetration tests with a company. The security in software development and in system deployment and operations of the system has been practiced during the genomic data development projects.

Results

The pilot system is now established, and includes software components and necessary hardware for system integration and testing, as well as planning and implementing production practices and documentation. Here, we established the central repository, and one satellite node based on Federated EGA technologies, with federated interoperable access enabled by existing identity and trust federations using ELIXIR AAI and REMS. The central repository was built in national academic compute centers in

Finland and Sweden, enabling data mirroring cross-border where possible for increased availability and disaster resilience. The federated node, a satellite to the central repository, was set up as a demonstrator of technical ease of deployability, and of technical readiness of the federated system, as a preparation for adding more federated nodes in the future. The central repository is the primary resource for project data. The design provides a private execution environment, where users can log in using ELIXIR AAI and access data according to entitlements in REMS. This design will be used as a basis for development of a private execution model, which will be refined in Task 2.6, and is intended to enable morphological data mining using algorithms developed in WP4 over the course of the project. Basic discoverability features were implemented based on straightforward search and filtering of the one metadata parameter that has hitherto been identified by the Metadata Task Force as suitable for use in discovery and classified as public, namely "number of images in the dataset".

To support the development of the pilot system in WP2, and of BIGPICTURE metadata standards, converters, and validators in WP3, WP2 has invited BIGPICTURE partners to submit pilot data to a common inbox for the project. This inbox was implemented as a shared folder on the AIDA Nextcloud system provided by LiU/AIDA Data Hub. All interested parties in BIGPICTURE were invited to apply for access to this data, with the intention that all should be able to contribute and monitor progress of the development of converters, validators and standards. Pilot data should be sufficiently variate to accurately represent the variability that is expected to occur in the actual clinical and non-clinical data that will be collected by WP3 during the course of the project, yet it must still be strictly anonymized and non-confidential, as its intended use is in aiding the development of systems and processes that do not yet exist, and where no guarantees for information security can thus be provided. At the time of pilot system launch, about 0,5 TB of pilot data had been gathered, in various forms and formats. Generated mock data was also used in order to complement this pilot data, for example for technical performance- and interface tests.

To demonstrate technical ease of deployability, the federated node was not deployed in any of the central repository data centers, but on separate hardware resources provided to Uppsala University by the SNIC Science Cloud (SSC). SSC is a national academic cloud host operated by the Swedish National Infrastructure for Computing (SNIC), and offering services to Swedish HEIs. SSC is organizationally and geographically separate from SUNET and SafeSpring (which provide hardware resources for the Swedish central repository data center). To demonstrate functionality and federated search capability, the federated node was populated with non-confidential mock data, and a federated query was successfully carried out, to meet Milestone 6 - Successful technical demonstration of deployment of the first federated node (M10). As SSC does not guarantee suitability for use with GDPR §9 special categories of personal data, the federated node was decommissioned following this demonstration, so as to reduce risk of misinterpretation and confusion, and to reduce risk of accidental sharing of confidential information. SSC will be retained as a potential venue for similar demonstrations in the future, if such a need should arise.

Overview

In short, WP2 is responsible for:

- Implementing APIs for receiving and accessing images, AI models and metadata.
- Repository to store data. Storage and related middleware (software) and operational aspects.
- Providing a platform for private inference execution.
- Provide a server to run REMS, which WP3 manages.
- Provide a server to run Beacon and create the metadata index based on the query needs.

1. Development platform

The collaborative code development of the open source core services that the BIGPICTURE repository is based on is coordinated on GitHub, in the following repositories:

- Beacon - https://github.com/CSCfi/beacon-python
- Beacon network - https://github.com/CSCfi/beacon-network
- Beacon network UI - https://github.com/CSCfi/beacon-network-ui
- Federated EGA - Sensitive Data Archive (SDA) repositories on the Nordic e-Infrastructure Collaboration (NeIC) GitHub organization. https://github.com/orgs/neicnordic/
- Metadata submitter - https://github.com/CSCfi/metadata-submitter
- Metadata submitter frontend - https://github.com/CSCfi/metadata-submitter-frontend
- REMS - https://github.com/CSCfi/rems

A side benefit arising from continued use of the established development collaboration platforms is that procedures for version control, issue tracking, and code review, as well as unit- and integration tests are already in place and well-understood, to ensure that each change is tested and good coding practices are followed.

BIGPICTURE strives to carry out as much of its open source code development work as possible in-place, so as not to create unnecessary and costly code divergence. However, some highly BIGPICTURE specific work, that is intended to be made publicly available as open source software as soon as the BIGPICTURE general assembly can approve such publication, is currently made available only to BIGPICTURE partners on the BIGPICTURE GitHub organization:

- BIGPICTURE metadata standard: https://github.com/imi-bigpicture/metadata-schema

2. Deployment environments

While the data management middleware is designed to work well in different kinds of deployment environments, the Swedish and Finnish deployments are quite similar in nature.

The Swedish data center is implemented as an OpenStack[7] project in an Infrastructure-as-a-Service (IaaS) cloud operated by SafeSpring, according to a framework agreement made available to Uppsala University by the SUNET. Here, Terraform[8] and Ansible[9] are used to to set up Kubernetes to run in High Availability mode, on top of which the sensitive data pipeline is then deployed. The data center also collects all trace logs and provides an index into the recent logs. The older indexes are backed up to object storage (S3). Access to the different environments is secured through Hashicorp Vault where encryption keys for Kubernetes credentials are kept.

The Finnish data center at CSC utilizes physical servers in a piloting environment running a Kubernetes cluster, and plans to integrate a Hashicorp Vault setup just as the Swedish data center does for the production phase. The piloting environment follows a Infrastructure as Code (IaC) making use of Terraform and QEMU[10] for deployment of the Kubernetes cluster. Five servers were procured to serve the needs of the pilot system. In the future, more investments will be carried out in order to meet needs presented by AI model inference execution.

The current assessment is that these servers in combination well exceed the needed capability to carry out private inference execution against available test data at this time, and to meet Milestone 7 - Sufficient compute to support inference against available test data (M10), and will provide a good starting point for further development of the private inference execution model in Task 2.6 - Iterative development (M10-M71).

[7] https://www.openstack.org/
[8] https://www.terraform.io/
[9] https://www.ansible.com/
[10] https://www.qemu.org/

3. Data mirroring

The data mirroring technologies will initially utilize the normal data submission pipeline. The implementation in the pilot system is that the Swedish data center has dedicated a specific data transfer inbox account for mirroring data submissions that were originally sent to the Finnish data center, and vice versa at the Finnish data center for submissions that were originally sent to the Swedish data center. The initial approach is semi-automated, in the sense that most processes are automated, whereas some are kept manual, so as to allow manual inspection and/or intervention at critical steps, to ensure the system as a whole is indeed behaving as intended. One such example is manual synchronization over email to make sure that mirroring transfers have successfully concluded. At such a time where enough confidence is established in the correctness of the system as a whole, the manual inspection/intervention steps will be foregone in favor of a fully automated and thereby more scalable process.

The functionality of the data mirroring was verified using available pilot data and metadata in available formats.

The BIGPICTURE metadata standards, file formats, and validators are being iteratively developed, and these verifications will be repeated with iteratively more representative pilot data as such becomes available. These topics are pursued in collaboration with WP3 and WP4 in Task 2.6 - Iterative development (M10-M71), Task 3.3 - Development of ontologies and standardized metadata sets (M1-M48), Task 3.4 - Defining and implementing datasets and metadata sets (M6-M12), and Task 4.5 - Develop a technical implementation of the open digital slide, annotation, and meta-data format (M1-M36). These topics are also being addressed cross-sectionally in the Metadata Task Force.

We expect that a first verification of a successful mirroring event with real clinical data fully conformant to the BIGPICTURE metadata standard will take place at the latest in connection with Milestone 8 - Delivered clinical dataset via the submission portal to the BIGPICTURE repository (M18).

4. AAI

The services, like REMS (see the next section), integrate to the AAI system, which forwards user attributes to the receiving service. In the other direction, REMS will submit authorisation information, which the data management middleware will consume. A user who wants to access granted data authenticates with the data access client against the AAI service.

Findings from Task 3.12 - Honest broker requirements gathering (M1-M9) are being further developed in Task 3.13 - Honest broker pilot system development (M10-M15) and Task 3.14 - Honest broker mechanism pilot testing and operation (M16-M24). These topics are being continuously and cross-sectionally addressed by the Honest Data Broker Task Force and the GDPR Task force. A fully aligned understanding is thus expected to be in place by the time the repository goes into production mode in Task 2.8 - Large scale archive operations (M25-M72).

Processes and policies then need to be defined, for example to establish what can be considered an appropriate level of assurance for identity provider (IdP) provided identities in BIGPICTURE. To provide a frame of reference, "just having an account at a public cloud service" is probably insufficient for identity assurance needs arising in BIGPICTURE, while banking level assurances like "showing passport and physically signing paper agreements onsite at the data center in Finland" is probably a stricter interpretation than can be reasonably motivated in practice. In the same time frame as outlined above, WP5 will define more precise requirements for the processes regarding the authentication and authorisation, such as evaluating the need for the multi-factor authentication.

Initially, the pilot system will not make use of the full feature set of ELIXIR AAI, and access to the pilot system will instead be offered to a small number of pilot users. As the GDPR roles in BIGPICTURE are

becoming better understood and more developed, the system will open up to incrementally richer feature sets and larger numbers of users.

5. Entitlement management

Deployment of a dedicated REMS instance went as planned, though it will be iteratively enhanced during the project in Task 2.6 - Iterative development (M10-M71). The first instance was set up by CSC, as CSC is the developer of the web portal and it is using a SME company in Finland for the development as well.

REMS was integrated to ELIXIR AAI as planned so that the users are identified. Authenticated users are able to look at the catalog of datasets and apply for access to them. The amount of visible metadata is limited on purpose. It is important to note however that neither REMS nor its catalog of datasets are intended to serve as the main discovery service in BIGPICTURE (cf Discovery below). Users are thus expected to have found datasets of interest through other means, for example through searching using Beacon, or a conventional Internet search engine, from where they can click through to the REMS page for the dataset of interest, where they can then apply for full access to the dataset.

At the moment of writing, support for the data use ontology to support data access committee work is being implemented. The work with the Metadata Task Force and data submission process is ongoing to define the information needed for the data access committees to do their work. The application forms and processes may become more standardised by data contributors in WP3.

6. Metadata

Metadata is submitted to the pilot system in XML format. Each submission constitutes a dataset.

In the coming months, the BIGPICTURE metadata standards, file formats, and validators are expected to mature, through the course of being iteratively developed, and progressively more representative pilot data is thus expected to become available. We envision much of the metadata will be representable in DICOM, however we fully expect to have to use other formats for metadata that is not. Some of the metadata that is not representable in DICOM already has established mature formats, which in such cases will be used as-is, with as little interference as possible. These topics are pursued in collaboration with WP3 and WP4 in Task 2.6 - Iterative development (M10-M71), Task 3.3 - Development of ontologies and standardized metadata sets (M1-M48), Task 3.4 - Defining and implementing datasets and metadata sets (M6-M12), and Task 4.5 - Develop a technical implementation of the open digital slide, annotation, and meta-data format (M1-M36). These topics are also being addressed cross-sectionally in the Metadata Task Force. We expect that the first submission of real data fully conformant to a version v1.0.0 of the BIGPICTURE metadata standard will take place at the latest in connection with Milestone 8 - Delivered clinical dataset via the submission portal to the BIGPICTURE repository (M18).

7. Sensitive data storage

All data is encrypted with GA4GH encryption standard (Crypt4GH[11]) before submission to the archive. This is a public key encryption standard very similar to PGP. The submitter encrypts with the public key of the archive and then once files have arrived we check that it is possible to decrypt the files with the repository operator's private key and to make some integrity checks (calculate sha256 and md5 checksums). Then the encryption header of the file is separated and stored in the database while the encrypted content is put into the backend storage. The technology for the backend storage is different in the respective countries. The Swedish data center uses an S3 compatible backend while the Finnish data center uses an NFS filesystem, however the Finnish data center plans to transition to an object storage

[11] https://pubmed.ncbi.nlm.nih.gov/33543751/

backend also in 2022.

8. Discovery

The discovery service (Beacon) database is created from the submitted datasets. The key information of the dataset is the identifier to find that dataset in the data access granting portal (REMS). These two systems are not directly integrated together. The first supported query was against "the number of images in the dataset". Work is ongoing in the Metadata Task Force to define more discovery metadata parameters and to classify them as either public, registered, or controlled, which will enable progressively more advanced queries.

9. Private inference execution model

The initial model is based on the Cytomine job submission pipeline. We have adapted Cytomine to run on Kubernetes, both for the purpose of easing deployment, and to make use of a well known securable environment. The default software router has been completely replaced with a Kubernetes-based system that can be isolated from the rest of the system to limit the risk of unauthorized data access.

The private inference execution sandbox is in place, but it has not been extensively tested in the pilot system, as a lot of work still remains in order to build the mechanism to launch the AI algorithms in the backend. This work will be carried out in collaboration with WP4 in Task 2.6 - Iterative development (M10-M71) and Task 4.6 - Implement a container-based infrastructure to integrate, connect and run external AI-algorithms (M12-M60). We estimate that a first version may become available in Q1/2022 and be iteratively improved from there.

10. Discovery export

Ability to export compatible metrics for advertising available datasets through the BBMRI Directory has been included as a boundary condition in the Metadata Task Force work with developing the BIGPICTURE metadata standard. The current assessment is that the BIGPICTURE metadata standard will be sufficiently rich so as to allow export of compatible metadata for advertising in the BBMRI directory at a time where datasets approved for external access has been uploaded to the BIGPICTURE repository, which will happen at the earliest in connection with Milestone 8 - Delivered clinical dataset via the submission portal to the BIGPICTURE repository (M18).

The current assessment is that BIGPICTURE will have the capability to export compatible metrics for advertising available datasets through the BBMRI Directory. Dialogue with BBMRI has been initiated to ascertain whether actually doing so is desired or even appropriate. Put shortly, the BBMRI Directory is a tool that collects and makes available information about biobanks throughout Europe that are willing to share their data and/or samples, and to collaborate with other research groups, whereas BIGPICTURE is not an official biobank, nor are many of its data contributors.

There do however exist other external discovery services that could be suitable candidates to fill possible BIGPICTURE needs for external advertisement, such as for example bio.tools, or even conventional Internet search engines. WP2 will continue discussions with BBMRI Directory, as well as investigate what other opportunities could be appropriate to pursue, in the course of Task 2.6 - Iterative development (M10-M71).

11. System security

Data Protection Impact Assessment (DPIA) work was started together with WP5. Technical and

organisational measures to protect data privacy will be formulated within the project for the data submitting organisations. The secure software development and operational practices have been formed also in the preceding genomics projects.

WP2 has made preparations to have its system security penetration tested by external international cybersecurity expertise in the form of an outsourced company. In this activity, reports will be delivered to the system owner on a level of detail that is useful to the system owner, which could be extremely dangerous to disclose to other parties, in unintuitive ways. For example, even an enumeration of what attack vectors were unsuccessful and therefore adequately protected against, could be very useful for an external party contemplating what attack vector they should try first, or what could be reasonable attack vectors against other, softer targets. We intend to make a summary report without sensitive details available to members of the consortium. This activity is planned to take place in Task 2.4 - Security review of pilot systems (M11).

Discussion

Findings from Task 2.1 Infrastructure requirement gathering (M1-M3) indicated that BIGPICTURE partners could benefit from having a more detailed description of the pilot system and the services that it is based on. The scope of this report was therefore broadened to include these topics.

Conclusion

The BIGPICTURE pilot system has now been established and has adequately demonstrated that the planned functionalities are in place, and in some cases even slightly ahead of the project schedule.

The repository infrastructure code development is coordinated in an open and transparent manner on an established development collaboration platform with functionalities that are well-known to the development teams, such as version control, issue tracking, code review, unit- and integration tests, etc. We expect that this development platform will serve the project well.

In order to maximise the potential for knowledge gain from the planned system security penetration testing in Task 2.4 - Security review of pilot systems (M11), we have also strived to make headway also in implementation of features that were not necessarily intended for inclusion at this early stage of the project. One such example is our adaptation of Cytomine to run on Kubernetes, both for the purpose of easing deployment, and to make use of a well known securable environment. The intention behind such efforts is of course to ensure that early versions of features of foreseeable core project importance can be penetration tested early, so that it can become easier to demonstrate to stakeholders that it can be reasonable to entrust this system with precious and confidential data.

One finding from Task 2.3 - Pilot system development (M4-M9) translates into recommendations for future project activities, some of which may be partially addressable in WP2 for example in Task 2.6 - Iterative development (M10-M71), but where others likely require action in other Work Packages and Task Forces. The observation and immediate action is that the GDPR roles in BIGPICTURE service use and provisioning so far are insufficiently understood, and as a result, access to the pilot system will be offered to a small number of pilot users, and will initially have a very basic feature set, until a more scalable/sustainable solution can be found for BIGPICTURE as a whole. Our recommendation is that this more scalable/sustainable solution for BIGPICTURE as a whole should be developed carefully, for example by WP6, WP1, WP2, and the GDPR Task Force, taking the following factors into account (in decreasing order of priority):

1. Business cases enabled.
2. Information security.
3. Liability.

We expect that having a well-developed understanding of the BIGPICTURE sustainability model will become increasingly more important for robust project decision making, the further the project

progresses. In Task 2.6 - Iterative development (M10-M71), we will therefore seek to collaborate with WP6 to ensure that this understanding is developed, with the expectation that this will improve our ability to prioritize those repository infrastructure developments that best support the envisioned business models.