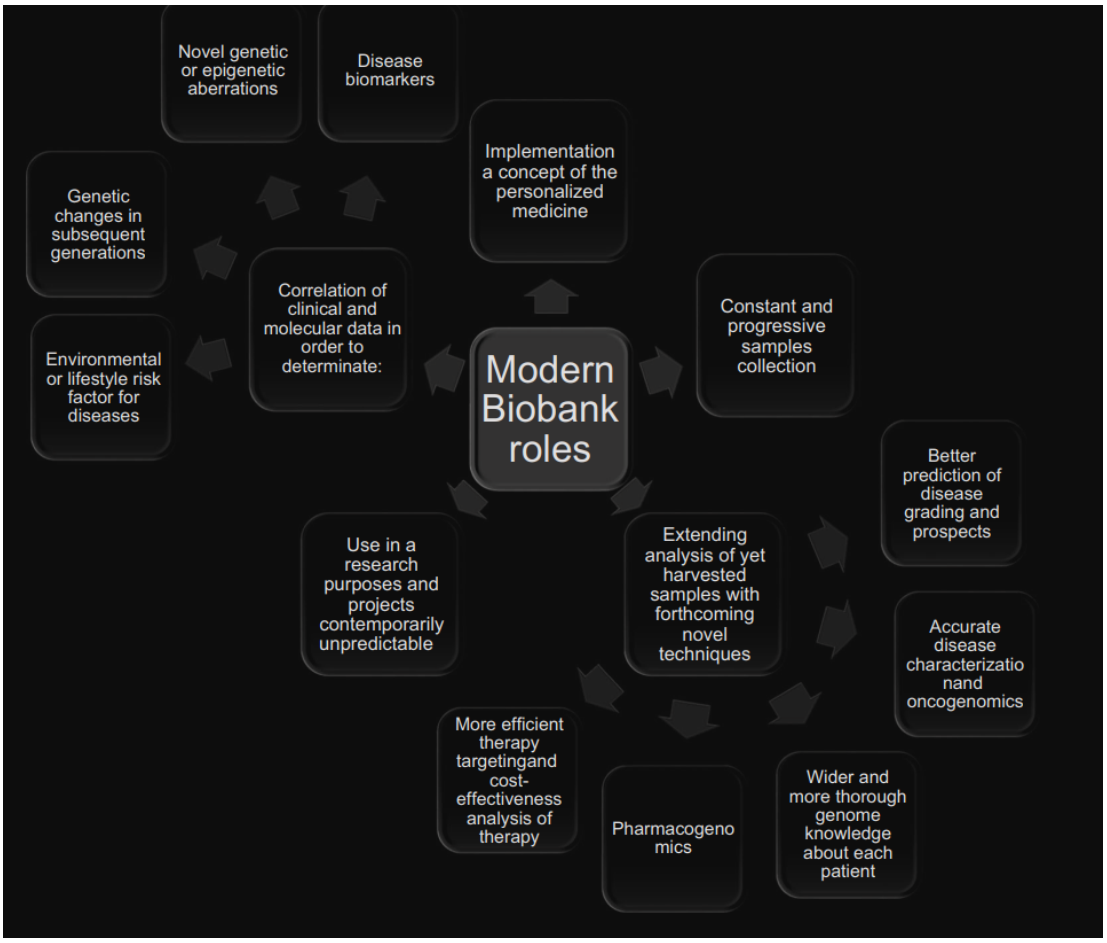
Lecture review:

Biobanking definition:

- a proper definition of biobanks is large collections of biospecimens linked to relevant personal and health information (health records, family history, lifestyle, genetic information) that are held predominantly for use in health and medical research.

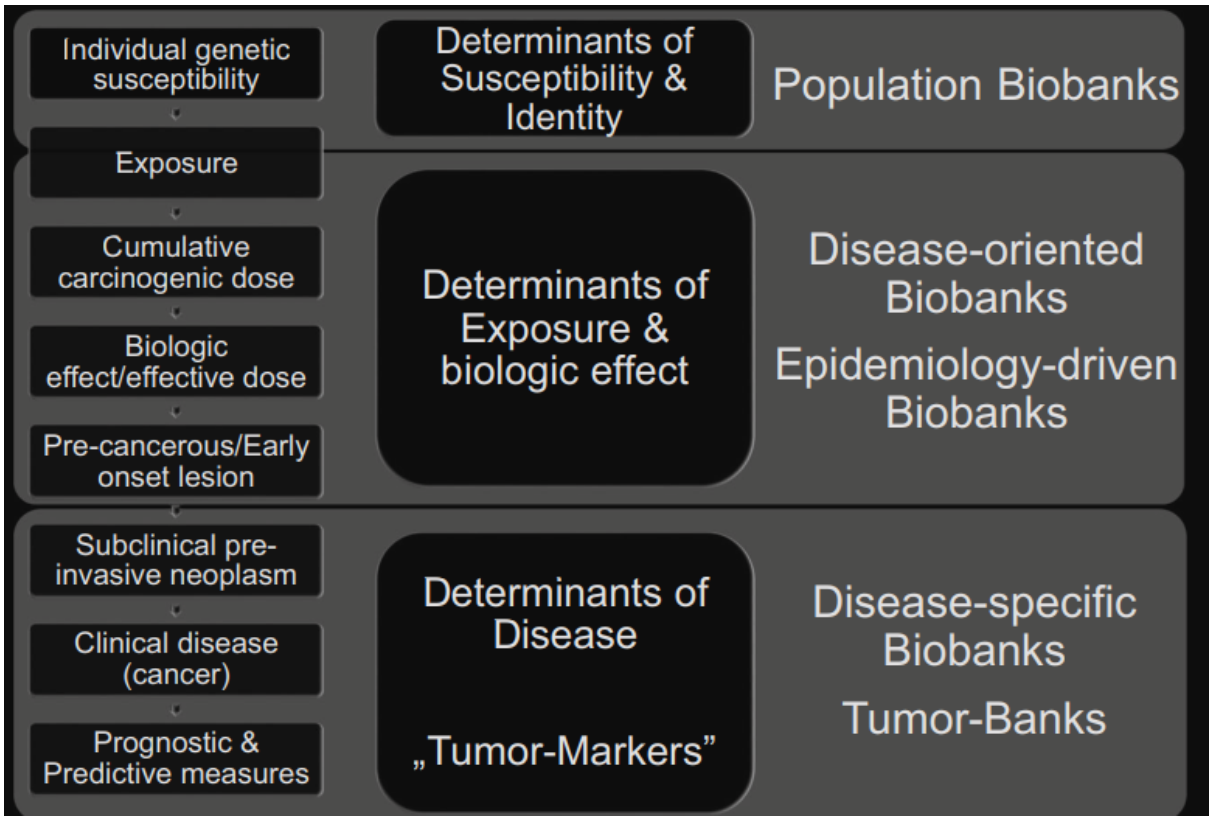
- **International Organization for Standardization:**
  biobanking : process of acquisitioning and storing, together with some or all of the activities related to collection, preparation, preservation, testing, analysing and distributing defined biological material as well as related information and data



In addition, some have illustrated biobank categories based on the associated opportunities of biomarker discovery [19]:

(i) Population biobanks (biomarkers of individual genetic susceptibility and identity)
(ii) Disease-oriented and epidemiology-driven biobanks (biomarkers of exposure and biological effect)
(iii) Disease-specific biobanks, such as tumour banks [19].

A second method of classification considers *the type of samples collected*, such as biobanks collecting frozen tissues, formalin-fixed paraffin-embedded (FFPE) tissues, cells, whole blood and derivatives, urine, buccal cells and saliva, bone marrow aspirate, semen, hair, nails and nucleic acids (DNA, RNA, cDNA/mRNA, microRNA) [3, 15].



what is a virtual biobank:

Virtual biobanks are a type of biobank that ***do not store physical samples,*** but rather provide a single point of access to a range of biospecimens using networks of ethical sources.

1.Picture archiving and communication systems(PACS)

[source][source2]

PACS provide storage and convenient access to medical images such as ultrasounds, MRIs, CTs, and x-rays.

Vendor Neutral Archive : Similar to PACS, **VNAs** are archives for DICOM-based images and content. However, they allow organisations to integrate the viewing and storage of different health IT systems regardless of vendor restrictions.

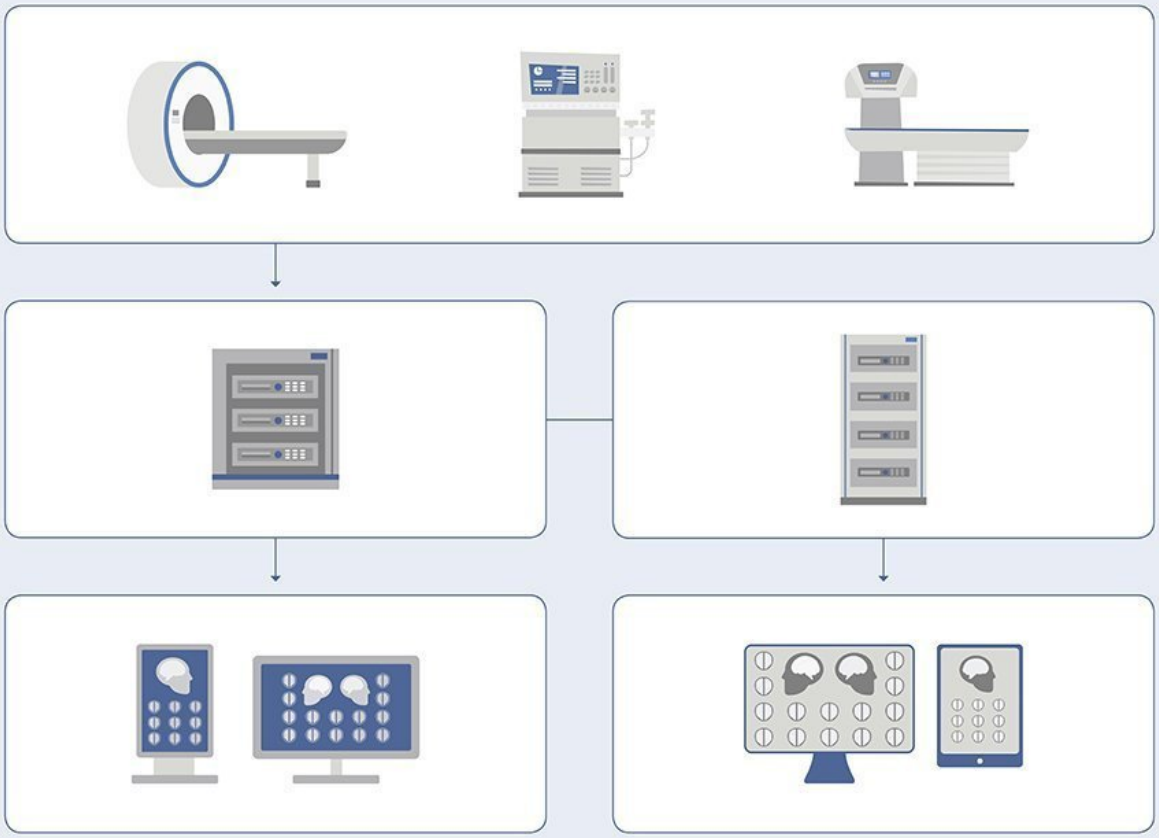
It distances itself from all suppliers of medical imaging equipment. The simplest vendor-neutral archive definition is that it is an application that stores medical images in a standard format with a standard interface. Therefore, images stored in VNA can be accessed from any workstation, regardless of vendor.

[source]
DICOM® — Digital Imaging and Communications in Medicine — is *the* international standard for medical images and related information. It defines the formats for medical images that can be exchanged with the data and quality necessary for clinical use.

DICOM is implemented in almost every radiology, cardiology imaging, and radiotherapy device (X-ray, CT, MRI, ultrasound, etc.), and increasingly in devices in other medical domains such as ophthalmology and dentistry. With hundreds of thousands of medical imaging devices in use, DICOM® is one of the most widely deployed healthcare messaging Standards in the world. There are literally billions of DICOM® images currently in use for clinical care.

The technology behind a medical imaging system

Healthcare organizations are gradually moving toward a population health management model and better patient enablement, and medical imaging informatics is at the forefront. Here's a breakdown of the major technologies that store and share images and allow them to be viewed.

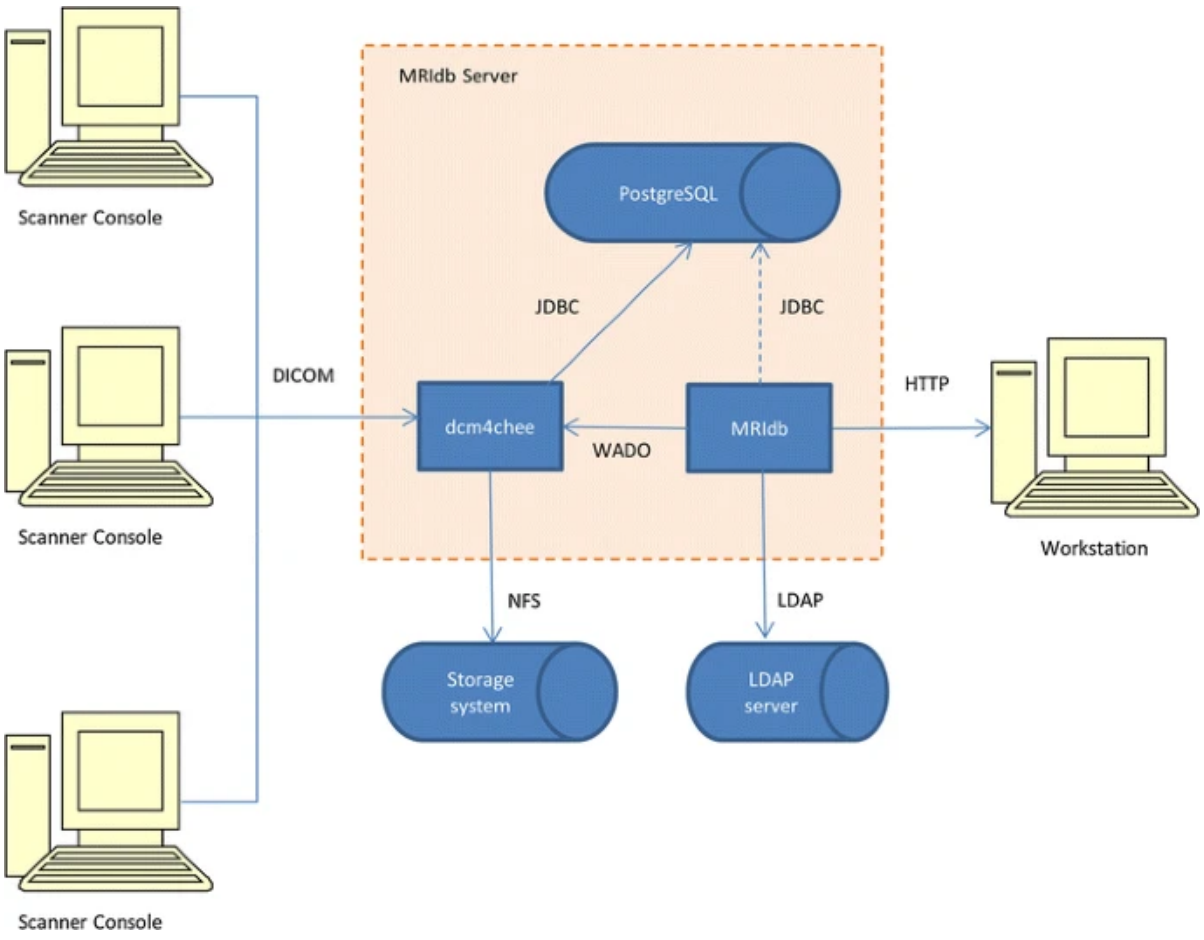


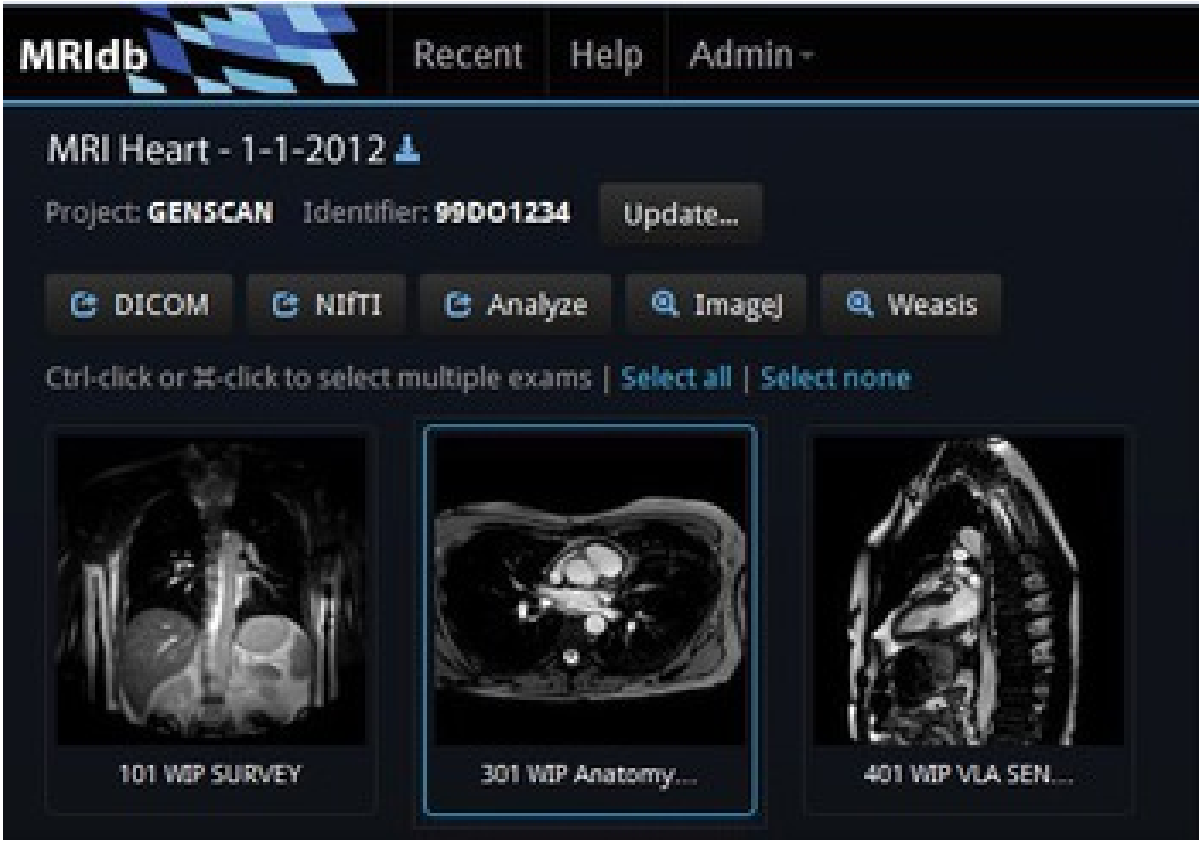
Companies use PACS technology to view images that have been created and stored on dedicated scanners (an MRI image can only be stored on an MRI scanner); they are then shared and viewed through a standard called DICOM.

VNAs allow images to be stored and shared beyond the dedicated scanners and PACS. They also allow images to be viewed on different screen types and devices (universal viewers) that display all image types and formats from various healthcare modalities.

[source][source2]

> MRIdb combines a DICOM image server (PACS, based on DCM4CHEE) with a web interface for administration and image viewing. It provides study management facilities with role-based access control, the ability to assign scans to studies, and audit logging. Images may be viewed in the browser using Weasis or exported locally in several formats.





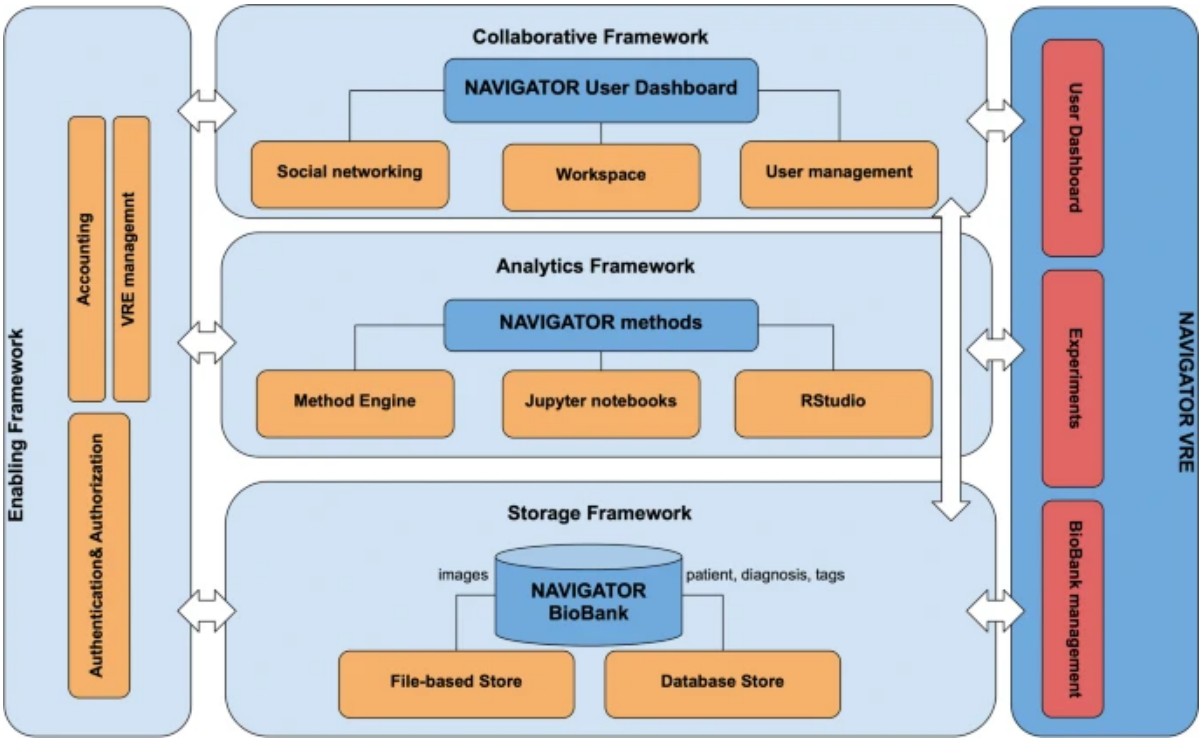
2.successful projects

[source]

2.1 NAVIGATOR: an Italian regional imaging biobank to promote precision medicine for oncologic patients

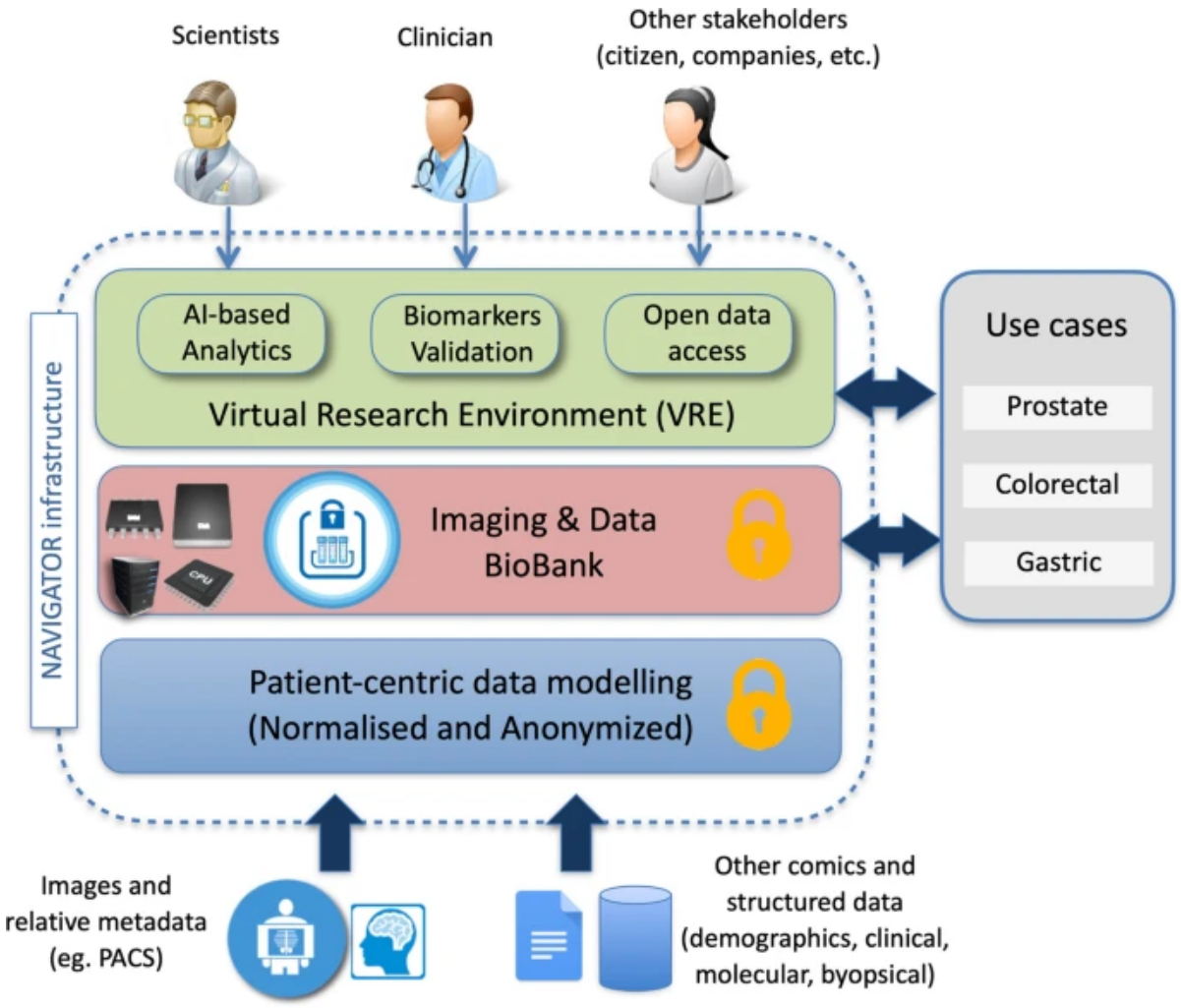
NAVIGATOR is an Italian regional project boosting precision medicine in oncology with the aim of making it more **predictive, preventive, and personalised** by advancing translational research based on quantitative imaging and integrative omics analyses. The project's **goal is to develop an open imaging biobank for the collection and preservation of a large amount of standardised imaging multimodal datasets,** including computed tomography, magnetic resonance imaging, and positron emission tomography data, together with the corresponding patient-related and

omics-related relevant information extracted from regional healthcare services using an adapted privacy-preserving model. The project is **based on an open-source imaging biobank and an open-science oriented virtual research environment (VRE).**



VRE

It is instantiated to provide the three sets of functionalities described in the architecture section (red boxes). The web applications will be developed by the project to address the specific requirements of the NAVIGATOR community, ensuring different user typologies are granted access to the application and data with dedicated access rights. The Collaborative Framework offers tools to exchange messages and experiences as well as a workspace (personal file system) where data can be uploaded and shared via fine-grained access control with other users. The analytics framework offers Jupyter notebooks and RStudio environments that users can adopt out of the box to operate over data extracted from the biobank (or otherwise uploaded by the user in the workspace); in scope, the method engine allows scientists to integrate their custom data analytics methods (Java, Python, R, C++, etc.) and web services to share them with other users (individuals or groups) as a service.



2.2 Biobank in a Box

[source]

Problem:

•**Data Heterogeneity**: Different research projects often use different data formats, standards, and terminologies, which makes it difficult to combine and analyse data from multiple sources. Combining data from multiple sources requires significant effort to clean, transform and integrate the data, and to resolve conflicts or inconsistencies.

•**Lack of Metadata**: Metadata, i.e. information about the data, is crucial for understanding the context, provenance and quality of the data, but it is often missing or incomplete, making it difficult to interpret the data and reuse it in new contexts.

•**Lack of Standardisation**: The lack of standardisation in data representation, data management and data analysis can make it difficult to compare and reuse data across different research projects.

•**Data Quality**: Ensuring the accuracy, reliability and completeness of data is essential for its reuse, but this can be challenging, if data is collected from multiple sources using different methods.

•**Data Provenance**: Data provenance refers to the origin and history of data, including information about how it was collected, processed and stored. Understanding the provenance of data is important for evaluating its quality and relevance for a particular research question. Proper documentation of data provenance can help researchers to understand the context in which the data was generated, identify potential biases or limitations and assess the suitability of the data for a particular research question.

•**Data Integrity**: Data integrity is used to describe the quality of the data through the life cycle. It comprises the maintenance of the data as well as the assurance of its accuracy over its entire life cycle. Moreover, it ensures the transparency and trustworthiness of scientific data.

•**Lack of Infrastructure**: There is a need for infrastructure that can support data interoperability and reuse, such as data portals and repositories, and for tools and services that can facilitate the process of finding, accessing and using data.

•**Data Privacy and Security**: Protecting the privacy of research participants and ensuring the security of data is critical, but it can be difficult to balance these concerns with the need to make data available for reuse.

•**Legal and Ethical Issues**: There may be legal and ethical considerations that limit the reuse of certain types of data, such as sensitive personal data .

FAIR principles

The FAIR principles are a set of guidelines for making data more Findable, Accessible, Interoperable and Reusable. In other words, they are intended to help researchers and organisations make their data more easily discoverable and usable by others, which can lead to more efficient and effective research. In simple terms, the FAIR principles can be described as follows:

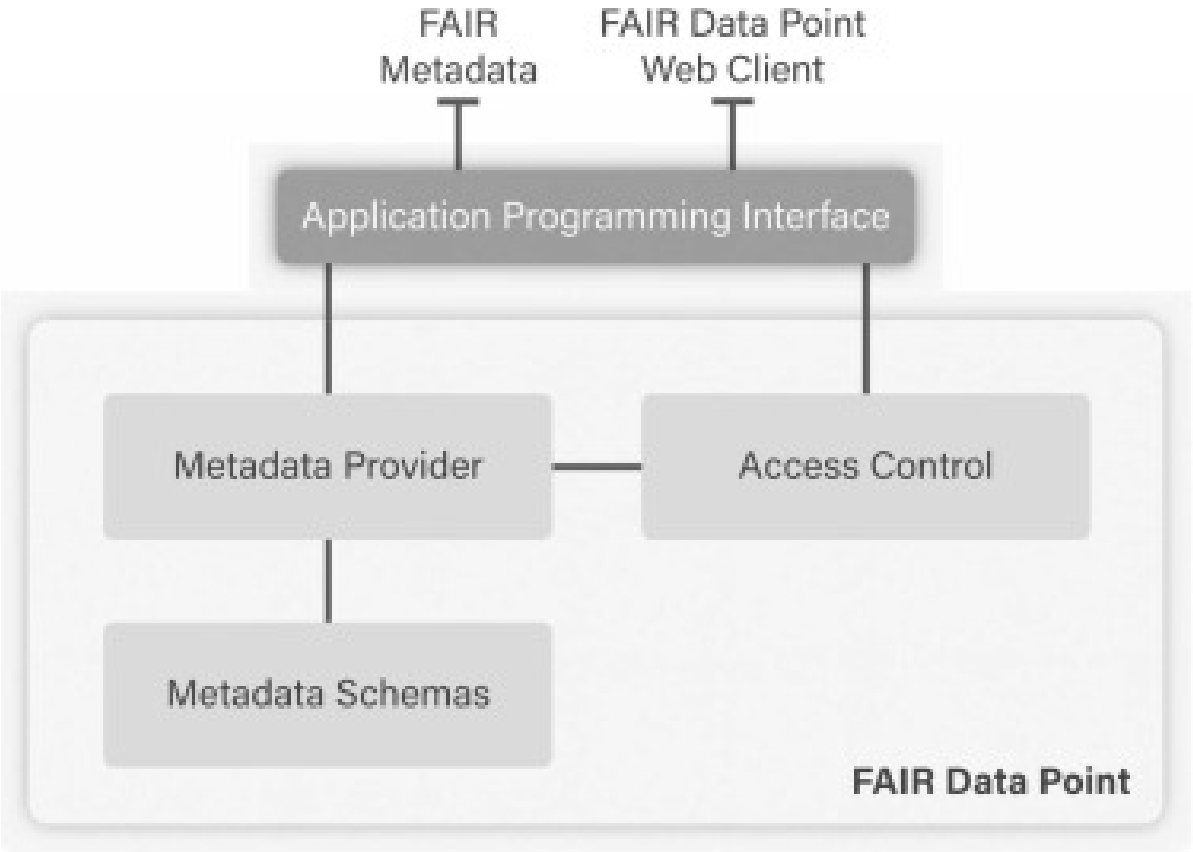
- **Findable**: Data should be easy to find, using unique and persistent identifiers and clear and detailed metadata. For example, if data are reused years later in other experiments, the identifiers must remain the same. The metadata describing the data should be rich and explicitly include the identifier of the data it describes. Data and metadata should be registered or indexed in a researchable resource.

- **Accessible**: Data should be available for access and use, using open and standardised protocols and formats. Data and metadata should be retrievable by their identifier. It is important to keep up-to-date accurate metadata and not to delete it, even if the corresponding data is no longer available.

- **Interoperable**: Data should be easy to integrate with other data, using standard data formats and ontologies. This allows datasets to be combined, merged and analysed together. Additional documents can be created, describing, e.g., exactly what was recorded and the units of each variable.

- **Reusable**: Data should be designed for reuse, using clear and detailed documentation and metadata and following best practices for data management. Metadata should describe the data using various accurate and meaningful attributes, such as file name, creation date and time, author, and when the file was last updated. Metadata categories should be chosen carefully, in a way that can be most useful for other researchers. Metadata should also contain a data usage

licence, which should be specific, clear and accessible. To increase the chance of the data being reused, the licence should be machine-readable, making it independent of spoken language. Metadata should also contain information regarding the data provenance.



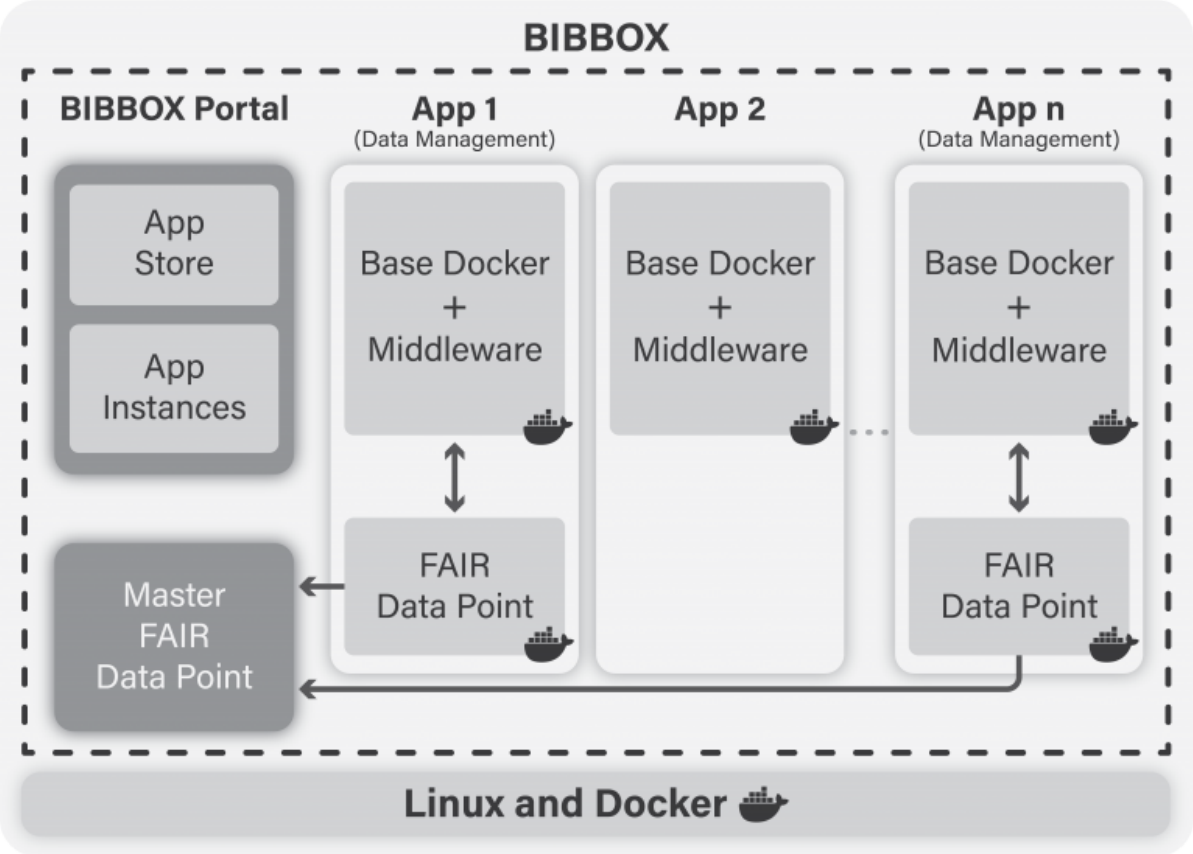
[github]

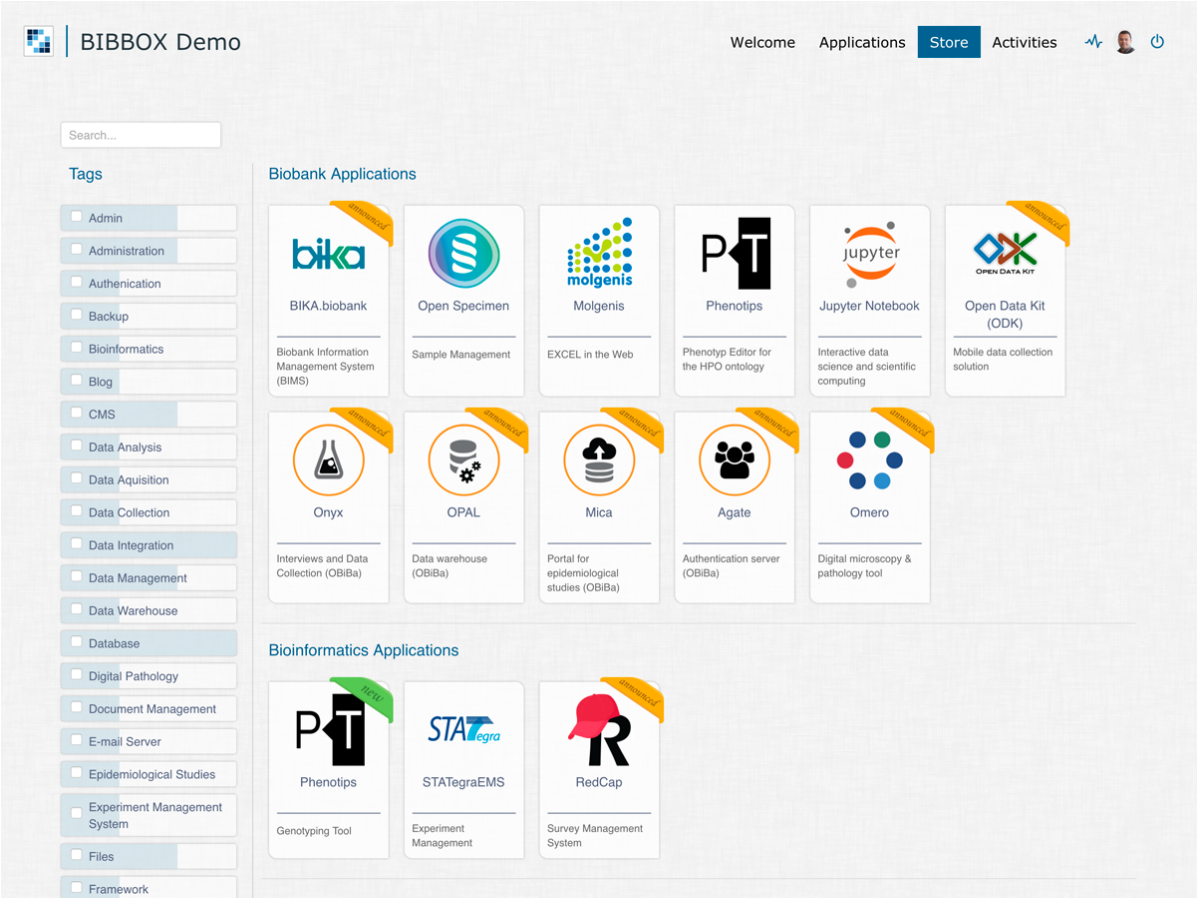
What is the BIBBOX Framework ?

The BIBBOX Framework is a Basic Infrastructure Building Box (BIBBOX). As it originally targeted biobanks, it might sometimes also be known as a "Biobank in a Box". The BIBBOX provides the possibility to **create and install Apps, and also serve them directly towards the end-user**. In the current

state, we are building Apps to support pathologists, bioinformaticians and biobanks in their daily work as well as in data management. The current framework mainly serves as a workflow demo SAAS-System.

Intended to offer an **easy installation, deployment and integration of open-source software solutions (Apps)** in the fields of **bioinformatics and biobanking.**

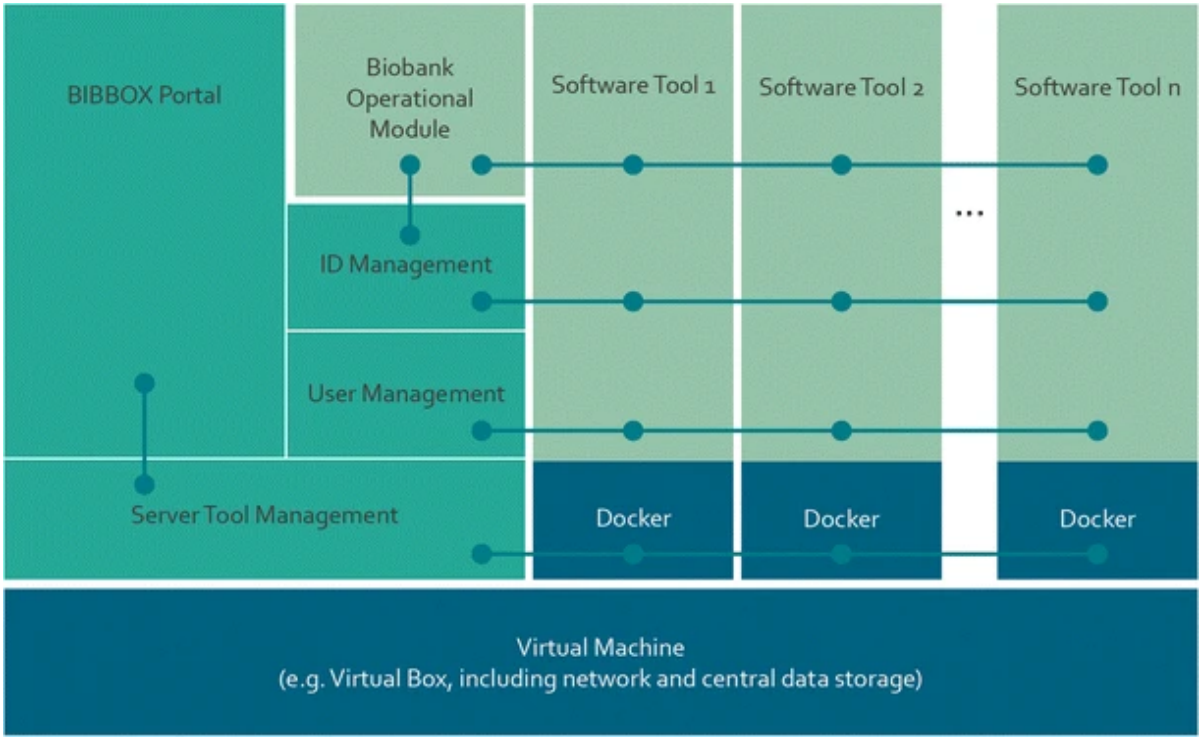




[source]

We have evaluated the available tools, described them in a catalogue (BiobankApps) and made a selection of tools available to biobanks in a reference toolbox (BIBBOX) that are use-case driven.

BIBBOX System Architecture:



The biobank operational module (BOM) covers the core functionality to operate a biobank, e.g. collection / study management, sample acquisition and sample metadata management, sample processing, sample storage, sample and data retrieval/distribution as well as data integration and cataloguing.
[Source2 for BOM]

[source]
A laboratory information management system (LIMS) is central to the informatics infrastructure that underlies biobanking activities.
Baobab LIMS: opensource

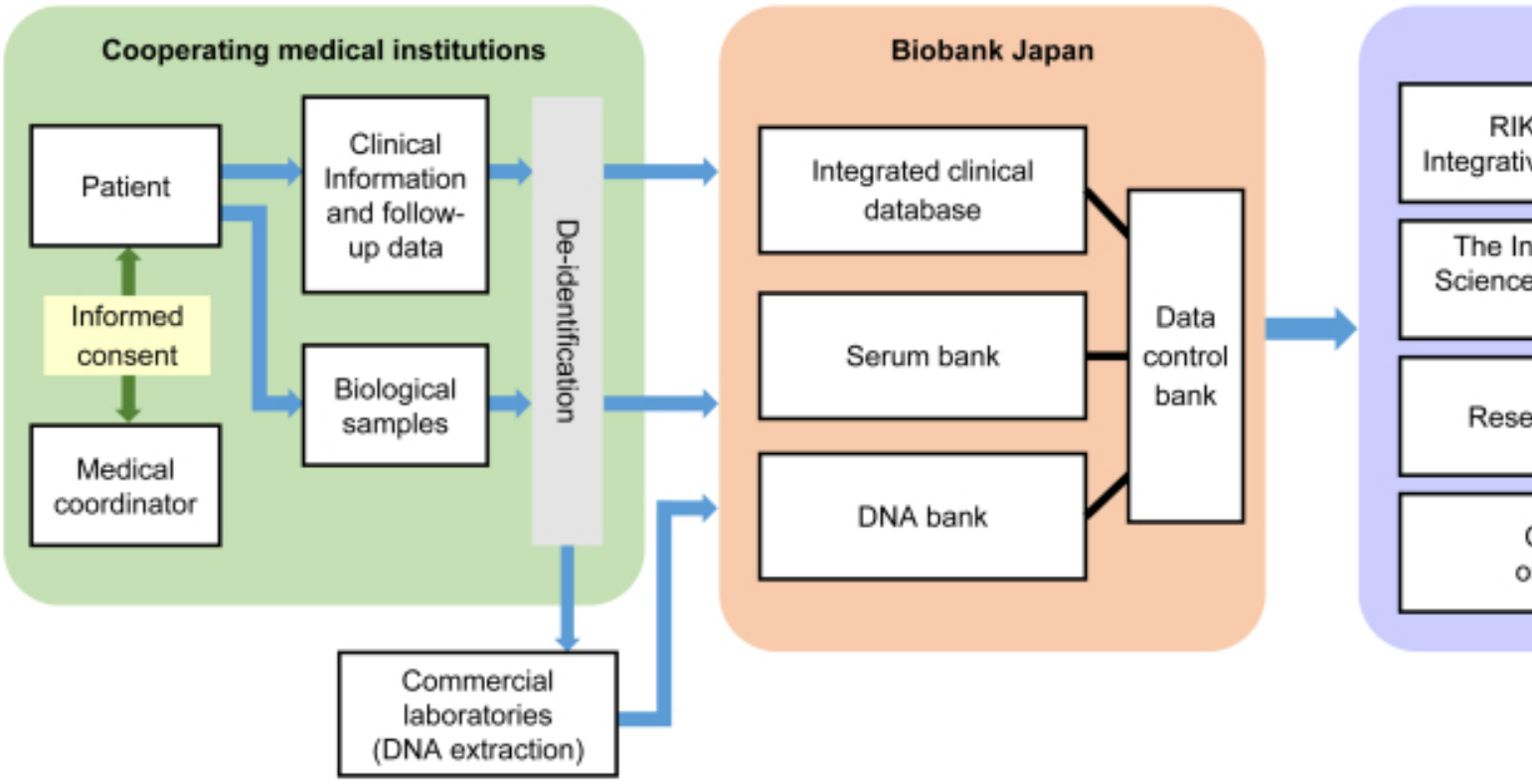
2.3 BBJ

The BioBank Japan (BBJ) Project was launched in 2003 with the aim of providing evidence for the implementation of personalised medicine by constructing a large, patient-based biobank (BBJ).

The BBJ is the first patient-based biobank in the world.

[source]

The flow of the collection of the sample:



Patient-based biobanks allow for the identification of susceptibility genes for common diseases because large numbers of cases are registered.

In contrast, population-based biobanks allow for estimation of environmental exposures and gene–environment interactions, but

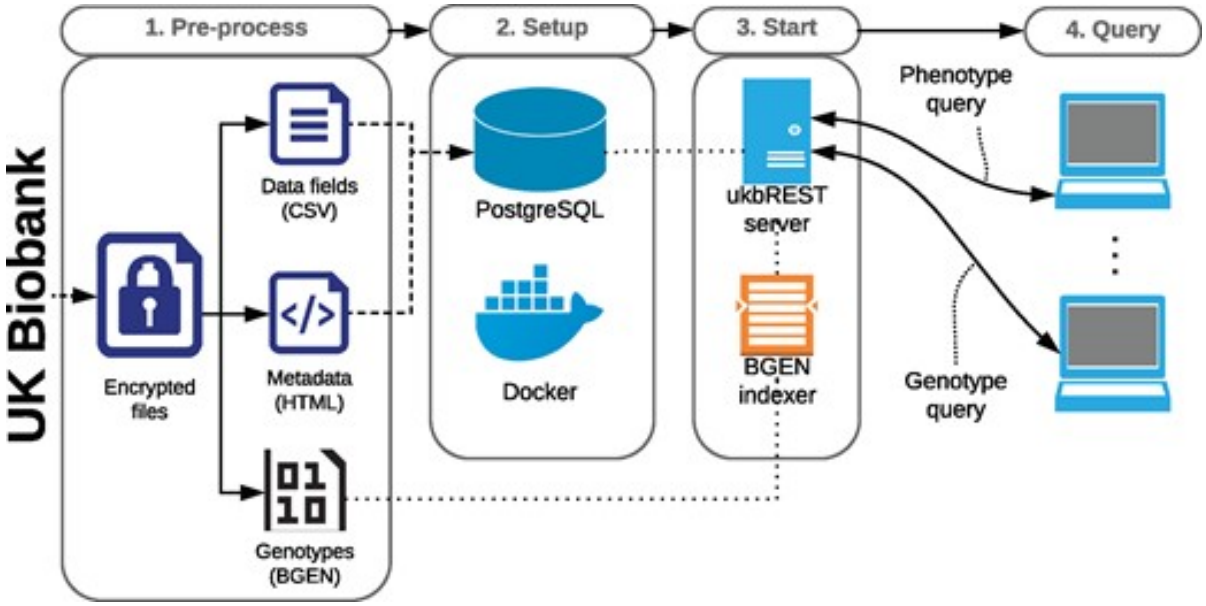
long-term follow-up is needed because a large number of disease onset cases is needed to achieve sufficient statistical power. Therefore, patient-based biobanks and population-based biobanks should work in cooperation with the goal of implementing personalised, precision medicine in the future.

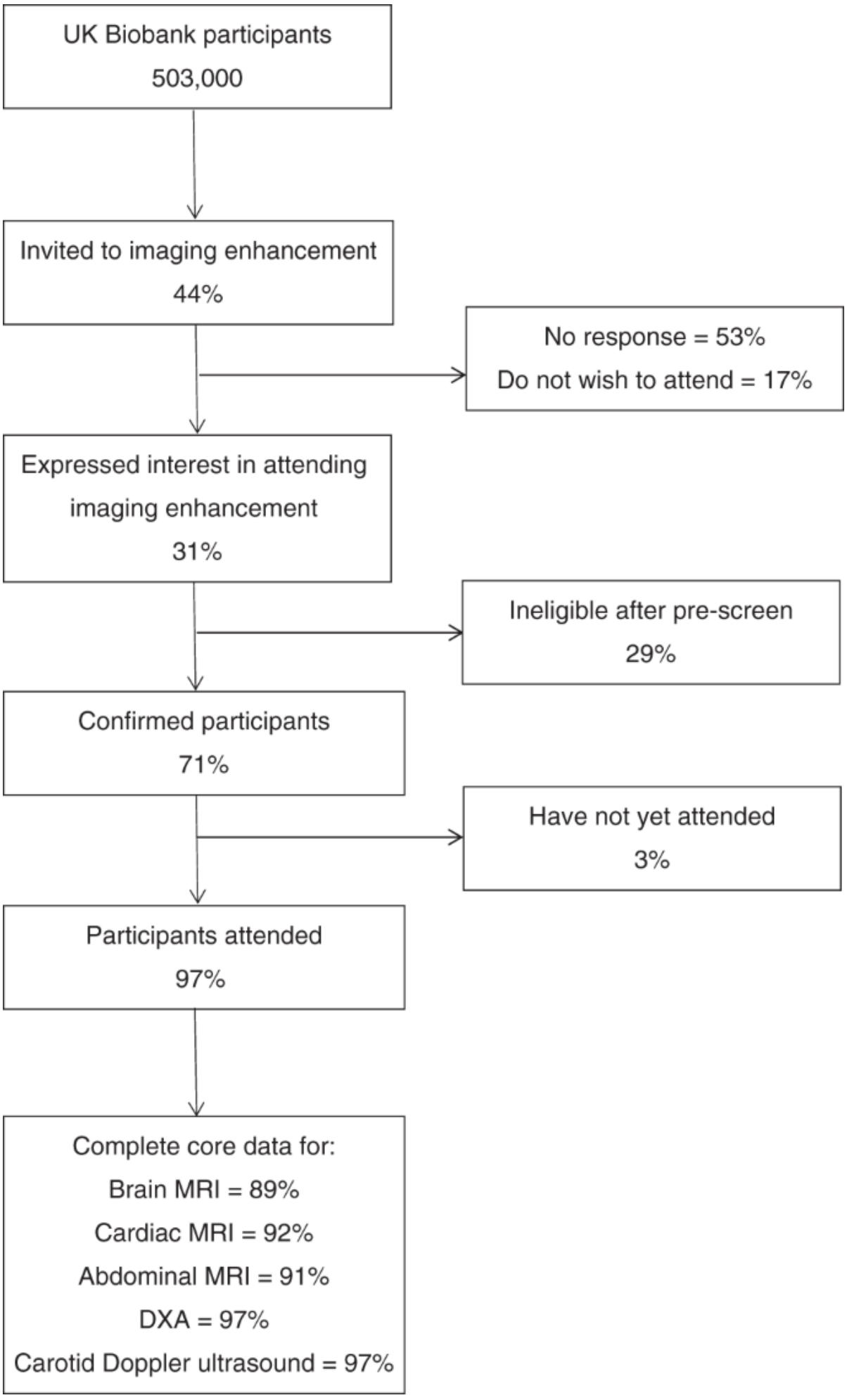
2.4 UK biobank

[Data description github]

[source][source2]

Flow chart of participation in the UK Biobank multi-modal imaging study.


1. Pre-process
2. Setup
3. Start
4. Query
UK Biobank
Data fields
(CSV)
Encrypted
files
Metadata
(HTML)
Genotypes
(BGEN)
PostgreSQL
Docker
ukbREST
server
BGEN
indexer
Phenotype
query
Genotype
query


UK Biobank participants
503,000
Invited to imaging enhancement
44%
No response = 53%
Do not wish to attend = 17%
Expressed interest in attending
imaging enhancement
31%
Ineligible after pre-screen
29%
Confirmed participants
71%
Have not yet attended
3%
Participants attended
97%
Complete core data for:
Brain MRI = 89%
Cardiac MRI = 92%
Abdominal MRI = 91%
DXA = 97%
Carotid Doppler ultrasound = 97%

3.ethics

- **Respecting research participant autonomy**:
  This means that i**ndividuals have the right to make their own decisions about whether or not to participate in biobanking research**, and their decisions should be respected.

- **Informed consent**: This refers to the process by which individuals are **informed about the risks, benefits, and procedures** involved in biobanking research, and voluntarily agree to participate.

- The notion of informed consent was shaped in the aftermath of World War II, and as a response to **the abuse of human beings** (inhabitants of concentration camps, prisoners of war, soldiers, civilians, etc.) in medical research and experimentation before and during the war. **Informed consent is a legal procedure to ensure that a patient knows, and understands, the risks involved in a given therapy or treatment.**

  [source]
  informed consent in the case of **biobanks is quite different from informed consent in a given experiment or medical procedure**. In the case of biobanks, what informed consent really means is **that participants entrust the biobank organisation—and its governing bodies—with the safeguarding and proper handling of their medical and physical and genetic information.**

  biobanks must provide **donors** with the choice to **“opt out”**, revoke their initial consent and, subsequently, **require the biobank to delete, or destroy, all information and samples**

Overview of the circuit of data and samples within a biobank. Blue colours represent information and materials circulating from the donor/patient through different steps while the other colours represent various factors, which influence the status of the biobank. The biobanks are different, having different characteristics, depending on the biological material they store (i.e., tissues, cells, blood, biological fluids, DNA, RNA, etc.), the types of information they collect, and also how the data is processed and organised.

- **Ethical considerations for return of results**: This refers to the ethical considerations involved in deciding whether or not to return individual research results to participants, and if so, how to do so in a way that is respectful and informative.

  [source]
  "People need to feel that they are part of something larger and that their donation feeds into a mutual, respectful relationship. This cannot be done simply by talking in abstract terms about the potentially significant medical ben 56 efits that might result from biobank research at some unspecified point in the future. Certainly, medical advances are relevant, but our research shows that participants in many countries expect individual feed-back from check-ups and also expect the possibility of gaining information about research advances that result from the biobanks in which they are participating, as long as their tissue or DNA is part of the biobank. They rarely expect money in return, but want to be appreciated as donors and be treated well."

- **Data privacy and donor confidentiality**: This refers to the need to protect the privacy and confidentiality of participants' personal and medical information.
- **Data storage and consent**: This refers to the need to ensure that participants' data is stored securely and that their consent for its use is obtained and respected.

- **Diversity in biobanking to promote equity**: This refers to the need to ensure that biobanks are representative of diverse populations, in order to promote equity in research and healthcare.

ISBER published a biobank handbook called Best Practices for Repositories that includes topics such as specimen collection, processing, and retrieval, training, ethical issues,

4.biobanking software solutions

IT systems :

IT systems have to serve databases in a real-time, easy access and user-friendly manner. Additionally, privacy protection and anonymization components cannot be neglected . The enormous number of digital, sensitive data requires optimised both hardware and software. There have been developed international standards for data obtaining and processing to enable compatible data sharing. International organisations such as NCI, ISBER, caHUB, OECD, EC-JRC are providers of recommendation useful in harmonising every branch of BBs' functionality

[source]

Table 2 List of crucial guidelines concerning biobanking issues

Institution	Document	Content	Year	Source (date of access: 27.07.2017)
National Cancer Institute	NCI Best Practices for Biospecimen Resources	A. Scope, applicability, and implementation B. Technical and operational best practices C. Ethical, legal, and policy best practices	2-0-1-6	https://biospecimens.cancer.gov/bestpractices/2016-NCIBestPractices.pdf
International Society for Biological and Environmental Repositories	Best Practices for Repositories: Collection, Storage, Retrieval and Distribution of Biological Materials for Research	Repository planning considerations, facilities, storage equipment and environments, quality management, safety, training, records management, cost management, biological material tracking, packaging and shipping, specimen collection, processing and retrieval, legal and ethical issues for biospecimens, specimen access, utilization and destruction	2-0-1-2	http://c.ymcdn.com/sites/www.isber.org/resource/resmgr/Files/ISBER_Best_Practices_3rd_Edi.pdf
Molecular Medicine Ireland (MMI)	MMI Guidelines for Standardised Biobanking	Part I: Pre-clinical standard operating procedures Part II: Clinical standard operating procedures Part III: Laboratory standard operating procedures	2-0-1-0	http://www.molecularmedicineireland.ie/wp-content/uploads/2015/09/MMIGuidelinesforStandardisedBiobanking_FINAL160610.pdf
Organization for Economic Cooperation & Development	OECD Guidelines on Human Biobanks and Genetic Research Databases	Part I. Guidelines on Human Biobanks and Genetic Research Databases Part II. Annotations	2-0-0-9	https://www.oecd.org/sti/biotech/44054609.pdf

4.personalized healthcare in biobanks

Personalised medicine (P4) defines a new approach to a patient and the disease. The concept of this personalization comprises 4 features :

- **Predictive** - ability to **conduct fast, precise and wide analysis** of **risk for particular diseases** requiring easy access and affordable methods.
  However, biobanks play a crucial role in discovering new predictive factors like genetic aberrations . In turn, correlating the discoveries with clinical data may facilitate predicting and support the next step of P4 - prevention.

- **Preventive** - comprises the idea of **avoiding disease progression by an early application of accurate** and personalised treatment. It may not seem to be a novel concept because there are yet implemented effective preventive solutions like **vaccination**, but unlike vaccines, which are recommended for the majority of the population, personalised medicine focuses on individuals.
  Biobanks-aided advancement can bring us to a higher level. In future, **it will be possible to elevate the prognostic value of early symptoms and combine them with genome data** which will finally lead physicians to quick and accurate diagnosis and enable them to administer the right treatment on time.

- **Personalised** - **genotypic and phenotypic** differences in the human population have a significant influence on **treatment efficacy.** The more individualised it is, the more efficient results are obtained. Recently whole genome and whole exome sequencing are widely available and more affordable. Deep knowledge about **genetic** and **environmental** circumstances of the patient **increases the accuracy of diagnosis and treatment**. Biobanks are centres of both types of data.

- **Participatory** - increasing **awareness** of both patients and medical professionals and their mutual communication are the basis of P4 medicine.

> Conversely, in this point the importance of IT companies increases, since they mediate the patient-doctor relationship by the development of intuitive, **accessible** and **privacy-safe-oriented systems.** Moreover, bioinformatics and new IT solutions are crucial for processing and organising huge amounts of data collected from a patient.

Each component creates a possibility for more efficient and suitable treatment choice. These units pose a chance to create a core of each part of a personalised medicine approach consistent with evidence-based medicine (EBM).
Oncologic diseases are an especial benefits of personalised medicine solutions. In the context of P4 medicine, biobanks may significantly develop the process of prevention, diagnosis and finally the treatment dedicated to the individuals. Numerous studies prove the significant role of biobanks in mentioned steps

9. Api
[source]

Results

We present Isabl, a customizable plug-and-play platform for the processing of multimodal patient-centric data. Isabl's architecture consists of a relational database (Isabl DB), a command line client (Isabl CLI), a RESTful API (Isabl API) and a frontend web application (Isabl Web). Isabl supports automated deployment of user-validated pipelines across the entire data capital. A full audit trail is maintained to secure data provenance, governance and ensuring reproducibility of findings.

Conclusions

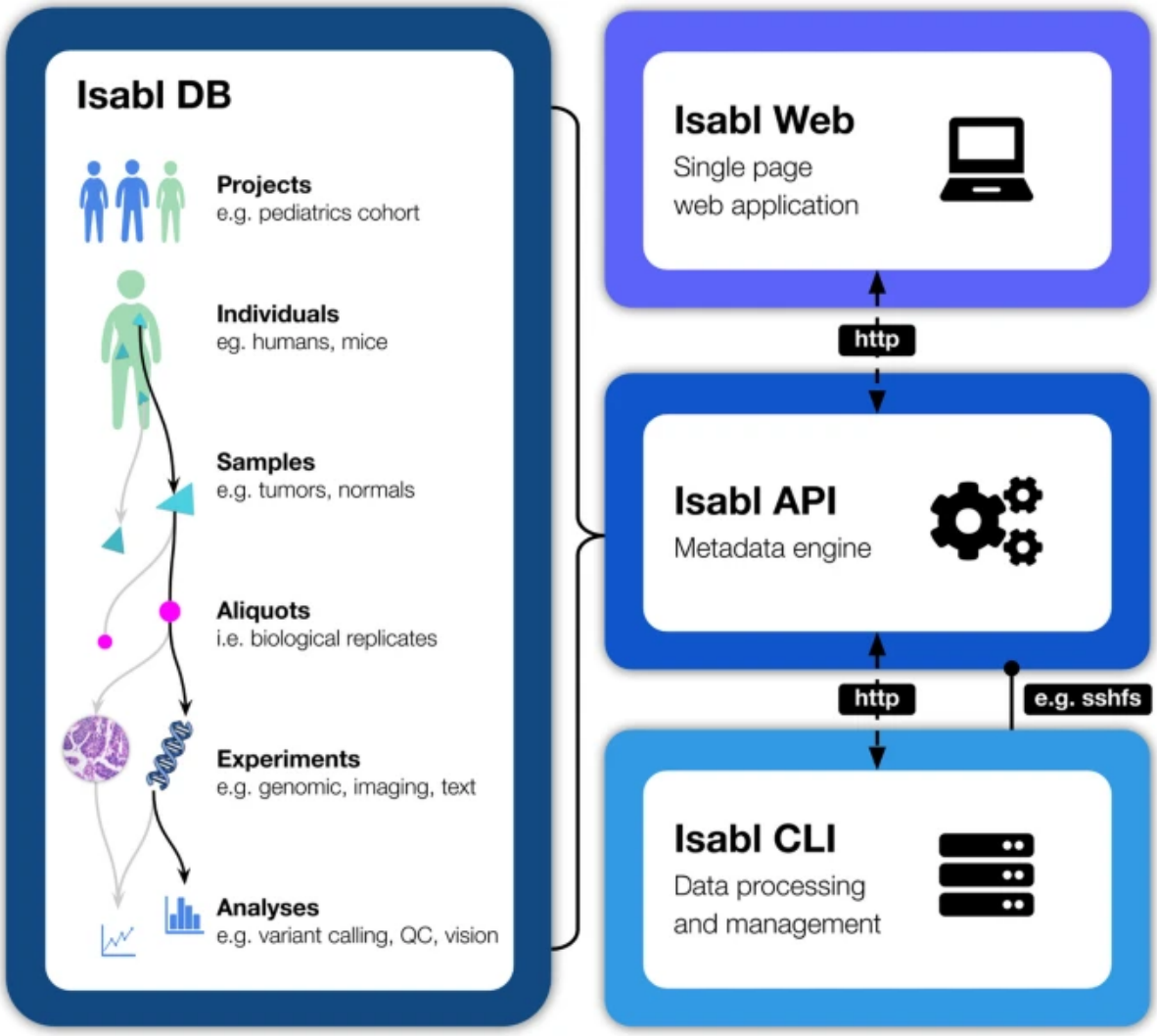
As a digital biobank, Isabl supports continuous data utilization and automated meta analyses at scale, and serves as a catalyst for research innovation, new discoveries, and clinical translation.

a

b

Schema	Description	Examples
Individuals	A unique subject	*Patients, Mice*
Centers	Center where the *Individual* is coming from	*Center A, Hospital B*
Species	Individuals species	*Human, Mouse*
Samples	Biological material collected from a given location and time	*Blood, Tissue*
Diseases	Disease linked to the sample	*Breast Cancer, Lung Cancer*
Experiments	Raw instrument data	*FASTQS, BAMS, PNG, CSV, TEXT*
Techniques	Technique used to generate the *Experiment* data	*WGS, Targeted, WTA, Imaging*
Platforms	Generating platform utilized	*Illumina NovaSeq 6000*
Analyses	An immutable *Application* result	*QC, Variant Calling, Image Processing*
Applications	A versioned, assembly aware type of analyses	*BWA, Mutect, Pindel*
Assemblies	Versions of reference genomes	*GRCh37, GRCm38*
Project	A collection of *Experiments* that will be analysed together	*Project 100*
Groups	Group linked to the *Project*	*Clinical Department*
Submissions	A batch submission of metadata	*Batch 1*

Isabl's relational model maps workflows for data provenance (e.g. Individuals, Samples, Experiments), processing (e.g. Applications, Analyses), and governance (e.g. Projects, Users). **a** An individual-centric model facilitates the tracking of analyses conducted on experimental data obtained from related samples. Analyses are results of analytical workflows, or applications. Experiments are analyzed together and grouped in projects. Additionally, schemas to track metadata for diseases, experimental techniques, data generation platforms, and analyses cohorts are also provided. Lines with one circle represent foreing keys, whilst lines with two circles represent many to many relationships. **b** A brief description of these schemas with examples



Schematic representation of Isabl's microservice architecture. Isabl DB provides a patient centric relational model for the integration of multimodal data types (i.e., genomic, imaging) and their corresponding relationships (individual, sample, aliquot, experiment, analyses). Isabl Web facilitates visualization of results and metadata in a single page application. Isabl API powers the linkage to other institutional information systems and is agnostic to data storage technologies and computing environments, ensuring metadata is accessible even when the data is no longer available (FAIR A2). Isabl CLI is a Command Line Client used to process and manage digital assets across computing paradigms (i.e. cloud, cluster). Arrow connectors indicate database relationships between Isabl schemas, dashed lines indicate metadata transfer through the internet, solid line indicates a data link between the data lake and the web server (e.g. sshfs, s3fs, https)