Project title: Accelerating the lab to market transition of AI tools for cancer management

Grant Agreement: 952172

Call identifier: H2020-SC1-FA-DTS-2019-1

Topic: DT-TDS-05-2020 AI for Health Imaging

D10.1 Prospective assessment of the Repository sustainability

Lead partner:	Università di Pisa (UNIPI)
Author(s):	UNIPI: Emanuele Neri, Esther Ciarrocchi, Jorge Shortrede, Lorenzo Tumminello. MEDEX: Karine Seymour. HULAFE: Luis Martí-Bonmatí. QUIBIM: Ana Jiménez-Pastor, Fuensanta Bellvis-Bataller, Ana Blanco UV: Ricard Martinez
Reviewer:	EIBIR: Katharina Krischak MAT: Amelia Suarez
Work Package:	WP10
Due date:	Month 18
Actual delivery date:	28/02/2022
Type:	Report
Dissemination level:	PU

Table of contents

Abbreviations

AI	Artificial Intelligence
AI4HI	Artificial Intelligence for Health Imaging
API	Application Programming Interface
BBMRI-ERIC	Biobanking and Biomolecular Resources Research Infrastructure–European Research Infrastructure Consortium
DSA	Digital Services Act
EBCP	Europe's Beating Cancer Plan
EHDEN	European Health Data & Evidence Network
ELSI	Ethical, Legal and Societal issues
ESR	European Society of Radiology
FAIR	Findable, accessible, interoperable, and reusable
FTE	Full time equivalent
GDPR	General Data Protection Regulation
HIS	Hospital information system
ICT	Information and communications technology
IPR	Intellectual property rights
NGO	Non-Governmental organization
OMOP	Observational Medical Outcomes Partnership
PACS	Picture archiving and communication system
RIS	Radiology information systems
TCO	Total Cost of Ownership

Disclaimer

The opinions stated in this report reflect the opinions of the authors and not the opinion of the European Commission.

1. Introduction

The D10.1 is undertaken as part of WP10 “CHAIMELEON REPOSITORY SUSTAINABILITY AFTER PROJECT END”, framed under T10.1 “Actions to ensure CHAIMELEON repository sustainability”. This deliverable defines the overall strategy for the sustainability of the CHAIMELEON repository during and beyond the duration of the CHAIMELEON project. Moreover, this document will address the challenges that the consortium will have to overcome and the steps to follow for guaranteeing the sustainability of the CHAIMELEON repository beyond the end of the project.

The deliverable is divided in the following main sections that include: sustainability challenge, our vision beyond the project, CHAIMELEON governance, costs and the revenue of the CHAIMELEON repository.

The sustainability strategy will be reviewed and updated during the project in order to implement the best actions to guarantee the project’s sustainability after its end.

2. Objectives

The objective of this deliverable is to set out the following aspects:

- To define the sustainability plan for the repository and the strategies for ensuring the long-term existence of the CHAIMELEON imaging biobank.
- To analyse the context in which the CHAIMELEON repository will operate and the challenges that need to be dealt with. A section will be dedicated to the European initiatives in order to promote the synergies with other Repositories and Biobanks.
- To describe the governance structure of the repository in order to identify the key roles who will take care of the maintenance and management of its functionalities beyond the end of the project.
- To analyse the cost to operate the services of the repository.

3. The sustainability challenge

In recent years, biobanks have become essential tools for translating biomedical research into practice, leading to the precision medicine era with the aim of improving global healthcare treatment and services. Many nations have established specific governance systems to facilitate health data research and to address the complex ethical, legal and social challenges that they present, but this has not led to uniformity across the world. Despite significant progress in responding to the ethical, legal and social implications of biobanking, operational, sustainability and funding challenges continue to emerge [1].
Sustainability can be defined as “the endurance of systems and processes”. In terms of biobanking, sustainability can be considered within a frame of three main factors: financial, operational, and social [2]. The ability of a repository to secure stable sources of funding, ability to standardise procedures and protocols that guarantee the highest quality standards of operation, and compliance with legal, social, and privacy regulations, are all important factors that will determine its endurance [3].

The research power of biobank datasets and large biomedical databases are enhanced when they are combined with other equivalent datasets from other biobanks around the world. However, to achieve this it is necessary to implement uniformity in the method collection, extraction and codification of biological and medical imaging information. Furthermore, the ethical, legal and social (ELSI) implications must be considered appropriately [1]. Moreover, sustainability is challenged by both technical and privacy-related issues and a growing demand for quality, FAIRification, transparency, and accountability.
Biobank sustainability is a multi-faceted concept that many biobanks are facing to justify their continued existence [4]. Ensuring the long-term existence of a biobank requires high-quality and standardised data in order to guarantee the exploitation of them by different stakeholders. Indeed, imaging biobanks operate in a complex and dynamic environment, where scientific, ethical and legal values are intrinsically linked [5].
In recent years, virtual biobanks or biorepositories have emerged as a feasible option for the widespread sharing of research information and resources globally [6]. A virtual biorepository is an electronic database of biological samples and related information that is independent of where the actual specimens fiscally exist [7]. Virtual or imaging biobanks are not merely a collection of bioimages associated with other patient clinical data; rather, they involve advanced computer technologies where image data, metadata, and raw data can be used for imaging measurements and biomarker extrapolation. These new biobanks can contribute to developing innovative research fields such as radiomics and radiogenomics.

4. Our vision beyond the CHAIMELEON project

Biobanks are infrastructures that, in recent years, have been involved in transformative processes aimed at the construction of dedicated business plans and organisation from a business and commercial point of view. The aim should be to make these infrastructures more efficient and more sustainable. It is becoming evident that there is a need to improve the level of professionalism in management, in training and management of personnel, in the procurement and sustainability of structures, in a more corporate and industrial perspective on imaging biobanks. There is the possibility of establishing a precise plan for both structural and financial reorganisation, which will make it possible to make an imaging biobank more sustainable.

Long-term financial sustainability is a major concern for imaging biobanks and financial support represents a major hurdle for ensuring the long-term viability of them. In this respect, an important step is to develop a business model to identify strategies for self-sustainability and cost efficiency of the CHAIMELEON repository. The potential sources to contribute to the funding of the biobank may include the cost recovery related to access to datasets, commercialization of research results or derived products and funding through governmental institutions and agencies.

Furthermore, another important point is enhancing the availability and usability of the datasets of the repository which will lead to improved scientific networking and collaborative efforts. *In line with the main goals pursued by the CHAIMELEON Project, the repository will continue to be operative after the project ends, and the agreements with the data providers will also contemplate the availability of the datasets once the Project is concluded. The registered users (project participants as well as authorised external users) may retain access to the CHAIMELEON repository -with all information fully anonymized- to continue with the development, testing, and training of additional AI tools aimed at not only assisting clinicians,*

but also other players in the healthcare space, improving diagnosis, treatment, and follow-up of cancer and other severe diseases.

An important achievement of the CHAIMELEON project will be the creation of the European platform with DICOM datasets and other "-omics" information accessible, with some similarities to the TCIA, an US virtual biorepository. Researchers will have access to the datasets for many purposes, even to develop and test new algorithms. For this purpose, will be necessary to develop procedures to harmonize instruments for data collection, mining, and perform comparative analyses.

4.1 The EU strategy and our position

In February 2021, the 'Europe's Beating Cancer Plan' was launched by the European Commission. The plan addresses, amongst others, the need for accessible health data and combining them with new technologies to further the development of personalised medicine as a powerful tool to improve cancer treatment and lessen the burden of costs for potentially ineffective treatments.

CHAIMELEON addresses this need by setting up one of the most ambitious health imaging data repositories across Europe providing access to 40,000 cases of cancer, corresponding to approximately 20 million images. The CHAIMELEON repository will serve as a resource for development, testing, and training of AI tools aimed at assisting clinicians in cancer management and improving diagnosis, treatment, and follow-up. The project will thus contribute to a more precise and personalised management of cancer, making the most of data and digitalization in cancer care. In response to the European strategies for data and AI, CHAIMELEON will develop ethical standards for the use of health imaging data in the context of AI developments and aims to foster trust in AI solutions among healthcare professionals, patients, citizens, and stakeholders in both industry and academia. It will provide examples for maximising the value of imaging data in conjunction with other types of data and inform the discussion on a sustainable business model for health data repositories, data provision and accessibility.

Most significantly, CHAIMELEON will directly feed into one of the EBCP's flagship initiatives: the European Cancer Imaging Initiative since the aim of the initiative is to "develop an EU 'atlas' of cancer-related images, making anonymised images accessible to a wide range of stakeholders across the ecosystem of hospitals, researchers and innovators". In line with the European strategy for data, the EBCP seeks to improve the exploitation of real-world data using AI for cancer care aided by the establishment of the European Health Data Space, an EU-wide collaboration on a secure and patient-oriented use of health data for better healthcare, better research and better health policy making, to which the CHAIMELEON results will directly contribute.

To implement the ambitious EBCP, funding for research and innovation efforts is provided through various European programmes and initiatives, which are of interest for planning the sustainability of the CHAIMELEON repository: The Mission on Cancer, Horizon Europe, Digital Europe and EU4Health. CHAIMELEON will exploit these schemes to ensure continuation of the CHAIMELEON efforts and further use of its results.

For this purpose, CHAIMELEON has sought synergies and joined forces with four other European projects with similar aims and objectives and formed the AI for Health Imaging (AI4HI) group:

- CHAIMELEON: Accelerating the lab to market transition of AI tools for cancer management (led by Hospital Universitario y Politécnico La Fe, ES)
- EuCanImage: A European Cancer Image Platform Linked to Biological and Health Data for Next- Generation Artificial Intelligence and Precision Medicine in Oncology (led by University of Barcelona, ES)
- INCISIVE: A multimodal AI-based toolbox and an interoperable health imaging repository for the empowerment of imaging analysis related to the diagnosis, prediction, and follow-up of cancer (led by Maggioli SpA, IT)
- PRIMAGE: PRedictive In-silico Multiscale Analytics to support cancer personalized diaGnosis and prognosis, Empowered by imaging biomarkers (led by Hospital Universitario y Politécnico La Fe, ES)
- ProCancer-I: An AI Platform integrating imaging data and models, supporting precision care through prostate cancer's continuum (led by Foundation for Research and Technology Hellas, EL)

The group aims to combine their efforts and exploit the projects' outputs, agreements on standards and solutions in joint activities, initiatives, and funding applications. At present, a joint application is envisaged under the Digital Europe programme for the call "Federated European infrastructure for cancer images data" in the first half of 2022. The application will bring together the best of all five projects and lead to implementation on member state level to establish a pan-European digital infrastructure for access to cancer images and related patient data.

The main purpose of the AI4HI group is to establish a common strategy for the successful realisation of these similar biobanks, with respect to several aspects, such as dissemination, metadata models, and Ethical, Legal and Societal issues (ELSI). In particular, the network established the Steering Committee working group to decide on strategic actions, to discuss how to allow the sustainability of the platforms and their commercial exploitation. Another working group is dedicated to discuss the current status of metadata models for imaging biobanks, their limitations and possible solutions and improvements. At the end of December 2021, the working group submitted a white paper on these metadata models to the Journal European Radiology Experimental, and the manuscript is currently under evaluation.

4.2 European initiatives

EIBIR Imaging Biobank Catalogue

The European Institute for Biomedical Imaging Research (EIBIR) and the European Society of Radiology (ESR) have set up an imaging biobank catalogue with descriptions of imaging biobanks and image collections. As the amount of available imaging data from clinical practice and scientific research is growing rapidly and quantitative imaging biomarkers, radiomics and artificial intelligence rely on large imaging data sets for training and validation, the EIBIR Imaging Biobank Catalogue responds to the growing need of a collection of metadata on existing imaging biobanks and imaging collections. The catalogue contains the necessary relevant metadata on imaging collections and their annotated image data to make these findable for researchers and help them select the right datasets suited for their needs and immediately see the relevant access criteria and contact details. 2021 saw the soft launch of the EIBIR Imaging Biobank Catalogue with individual access being granted to database and

biobank managers contributing metadata about their imaging collections while the public launch is to follow soon.

The EIBIR Imaging Biobank Catalogue already has an established collaboration with the EU-funded Horizon 2020 projects EuCanShare and EuCanImage. It is envisaged to include the metadata about the CHAIMELEON imaging biobank in order to advertise it towards the research community.

The BBMRI-ERIC (Biobanking and Biomolecular Resources Research Infrastructure–European Research Infrastructure Consortium) is a European research infrastructure for biobanking. One of its main goals is to improve the storage, use and exchange of information regarding biobanks (i.e., metadata). A service offered by BBMRI-ERIC is the Directory [8], which is a catalogue containing information regarding all the European biobanks that are willing to share their data and samples with other research groups inside. The Directory is an opportunity for biobanks to gain visibility and to sustain themselves through data access fees, and it allows researchers to access a much broader amount of data and samples, thus improving the quality of their work. While originally it was focused mainly on biological biobanks containing tissues and samples, recently the Directory has expanded to host also metadata about imaging biobanks.

BBMRI-ERIC operates on a federated basis and it is composed of National Nodes. The process for the registration of a specific biobank inside the Directory relies on the existence of a BBMRI National node for the biobank coordinator, which in case of the CHAIMELEON project would be Spain. The country just recently became a BBMRI-ERIC Observer node [9]. Therefore, in three years it will be able to apply for a Full Member position which will allow Spain and the CHAIMELEON project to actively contribute to populating the Directory. In the meanwhile, we have collected through the Italian BBMRI-ERIC node the list of actions to be performed to include the metadata about the CHAIMELEON imaging biobank in the Directory, and they are summarised in Fig.1.

Step-1	Access to the BBMRI website, that **should be the website of the national node** (Italy, Austria, etc...).
Step-2	**Registration:** ask for credentials (within two weeks, a letter from the Director of the Node that will indicate the further steps in the path to membership).
Step-3	**Application for membership:** fill in a self-assessment questionnaire
Step-4	**Evaluation:** A commission evaluates the biobank in relationship with quality and ELSI (**Ethical-Legal-Societal Issues**) requirements.
Step-5	**Signature of the Partner Charter: in case of positive evaluation**, you become part of BBMRI.XX (depending on the national Node). **The identification number is given**.
Step-6	**Directory:** the biobank **will have a personal page** on the **national BBMRI website and BBMRI-ERIC website**. The biobank enters the **Directory (a kind of yellow pages)** of the biobanks where the Biobank collections are described in the MIABIS format.

Figure 1: Action list for the inclusion of the CHAIMELEON metadata in the BBMRI-ERIC Directory.

4.3 Stakeholder outreach

Dissemination and stakeholder outreach are key for the long-term take-up of the project outputs and will aid the sustainability of the CHAIMELEON repository. Tailored communication and dissemination activities for different stakeholder groups will help to raise awareness of the relevancy of the project and ensure high acceptability of its results, thereby

contributing to their long-term sustainability. The core stakeholders and target groups for the communication activities are outlined as follows:

- Information and communications technology (ICT) scientists, researchers, and technologists (in software engineering, big data, AI, machine learning, cloud computing, medical imaging, data safety, cybersecurity)
- Medical scientists, researchers, and clinicians (especially in the fields of oncology, precision/quantitative medicine, and biomarkers)
- Industry in medical and health sciences (vendors of PACS, HIS and RIS, medtech companies, pharma companies)
- Decision makers (at hospitals, head of departments, policy makers, public authorities (incl. data protection authorities), funding agencies (national and European))
- General public
- Multipliers/professional societies, NGOs, civil society

The dissemination objectives and strategies for these groups as well as appropriate dissemination means and channels have been defined as part of the project's First dissemination & communication plan (D11.2) and will be updated as the project progresses. Deliverable 11.2 describes in detail the overall strategy and activities undertaken during, and beyond the duration of the CHAIMELEON project. It has been set up as a living document and provides a framework for all the project's dissemination and communication activities.

The dissemination plan specifies that CHAIMELEON will develop proactive actions to disseminate the research outputs of the project. Raising awareness and facilitating uptake of the CHAIMELEON results will be achieved through distribution of promotional material (from print material to videos), scientific publications, presentations at (inter-)national conferences and general publications in the media. The project website will be used to sustain all project activities during the project and beyond its runtime. The CHAIMELEON partners will capitalise on existing partnerships, especially with major European professional societies, to identify early adopters and opinion leaders to maximise the impact of the project outcomes and results. The project partners actively seek collaboration with other projects to extend their networks to ensure continuation of the efforts within CHAIMELEON and sustainability of the repository. For example, the AI4H initiative is a direct outcome of these efforts. Moreover, CHAIMELEON is collaborating with other EU projects (e.g. H2020 SINFONIA project) and initiatives (e.g. EHDEN - European Health Data & Evidence Network). Information exchange with these projects will take place in joint workshops, through meetings and events, newsletters, and online dissemination activities, thus contributing to the sustainability of the project.

5. Governance

5.1 Governance structure

A centralised governance of the Repository is described in D2.2 “First complete verification of GDPR compliance of Repository prototype” led by UV. The following governance bodies and roles are proposed:

5.1.1. Corporate governance body or Directorate

This is defined as a supervisory body to which substantial decisions can be attributed in terms of defining the policies of the Repository and controlling its functioning in the main areas:

▪ Economic, administrative and personnel management.

▪ Approval of the General Terms and Conditions.

▪ Admission of new partners and/or users for specific projects.

▪ Engagement with and admission of new data provider nodes, approval of the integration of APIs, information sources, mobile applications, etc.

▪ Final resolution in procedures related to complaints and/or claims.

▪ Preparation of annual activity report and singular reports related to usage analytics or others

▪ Any others related to high-level decision-making processes in the Repository.

The Corporate Governance Body would be constituted by a Technical, a Scientific and a Legal /Ethics Director. They will appoint a Dataset Access Committee, in charge of the evaluation of the access requests for primary or secondary use of the datasets, with members from the Executive Team and the Advisory Panel.

5.1.2. Management bodies or Executive team

The body responsible for all functions necessary in accordance with the legislation in force or with the Repository's own needs for its ordinary functioning. It is the technical team responsible for the daily activities including

- Maintenance and update of the IT platform
- Support to users
- Prefiltering of data access requests
- Promotion and dissemination of the Repository
- Attraction of new data providers for the expansion of the Repository
- Liaison with the Boards and Advisory Group
- Fund raising

In most cases, these would be staff performing specific relevant functions:

▪ Data protection staff.

▪ Security officers and/or security committee.

▪ Data scientists

Without prejudice to what has been indicated in relation to the Advisory or Consultative Committee, it should be considered whether the Repository requires some professionalised and stable body, person or function for supervision and control as a Chief Data Officer.

Both the Board and the executive team will be carefully sized to ensure quality of operations while facilitating the Repository sustainability and avoiding excessive operational costs.

5.1.3. Panel of nodes

The representative body of the data provider nodes of the Repository.

5.1.4. Ethics Committee

In terms of ethics, the Repository may take into account different alternative or complementary operational criteria:

▪ to include a catalogue of committees, e.g. hospital and/or university committees, whose ethical functions are to be verified and formally recognised;

▪ to have its own body to which ethical verification functions are assigned;

▪ to make an agreement with an external existing committee for the evaluation of applications.

It should be borne in mind that only in university research and/or health research are such bodies available. Moreover, the composition, nature and tasks of these bodies would make it difficult to agree on their support if the volume of applications were significant. It is therefore proposed that an Ethics Committee should be set up. At least one risk should be identified in this area. Both from a strictly ethical point of view and from the point of view of the future EU Regulation on AI, different ethical variables can be identified depending on the nature of the data processing. This implies, in case of recognition of positive reports from other ethics committees, the need to consider the possibility of reserving review powers.

5.1.5. External advisory committee or Advisory Panel

The establishment of an advisory body will be considered. It could be assigned, among other tasks, the following:

▪ Guiding the tasks of the Repository, being heard and participating in the definition of its strategic lines.

▪ To promote interaction with all public and private sectors, including the so-called third sector.

▪ To propose programmes, actions or new developments.

▪ To report on annual reports, management reports and/or audits.

▪ To be consulted by the Repository's management bodies and to issue reports on its own initiative or when requested to do so.

▪ To promote specific areas:

→ Standardisation.

→ Interoperability.

→ Formation of regional, national and international consortia (federations of data spaces), whether public or public-private.

A plural body is proposed, open to society, with the presence of at least the following categories of representatives:

- Patients associations
- Business organisations.
- Professional associations related to the tasks of the Repository.
- Professional experts in data protection.
- Security professionals.
- Ethics experts.
- Humanities experts.
- Healthcare System representatives.
- Universities.

On the other hand, the possibility could be considered of integrating or defining, within the framework of the committee, some small, permanent, full-time unit that would contribute to the Repository's monitoring.

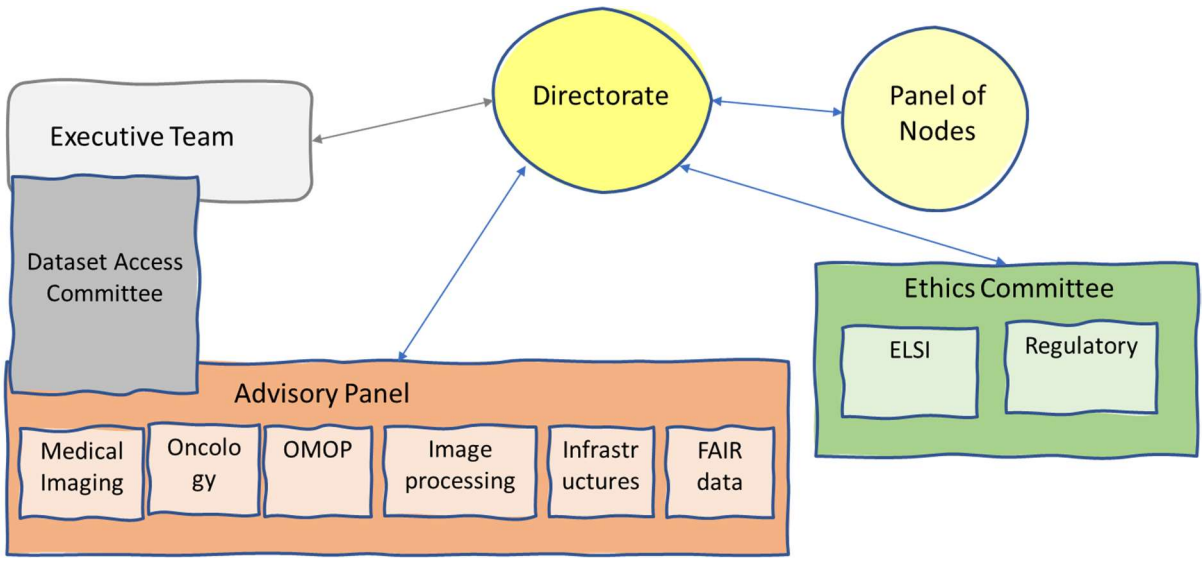

Figure 2: Proposed governance structure for a fully functional CHAIMELEON Repository

5.2 Access model

The CHAIMELEON Repository will use controlled access and will enable dataset use in the boundaries of the Repository only (no dataset download allowed, unless strictly required by a project and after having received explicit authorization). Only registered users will be able to

submit a Dataset Access request. In the registration process, all users will need to accept the conditions of use of the Repository which are being defined as part of WP2 “Ethical and legal aspects”. These conditions will include a commitment not to attempt to identify donor patients and to acknowledge the use of the Repository in any dissemination activity of their research results.

The Repository will implement user registration and authentication systems, under the leadership of UPV. A “Dataset Access request procedure” will be designed and an Access Committee will be set-up for the evaluation of requests, for the launch of CHAIMELEON as a fully functional Repository.

The access conditions to the CHAIMELEON Repository´s datasets will be a combination of:

- Controlled open access: accessible upon request and free of charge, with the Repository´s functionalities for usage traceability as an additional guarantee in the terms of art. 89 GDPR.

- Controlled restricted access: accessible upon request and with a fee for use. The pricing of the fee per use may be determined using a cost-recovery model or using a market-price model. The former is the most common model in research biobanks, nevertheless we are living in times of rapid evolution in the area of health data and market-price models cannot be disregarded at this early stage of the CHAIMELEON Repository development. Implications on public opinion and data providers willingness to contribute will also need to be carefully assessed.

The criteria for applying open or restricted conditions will be based on different technical and legal factors, including:

- The size of the collections: e.g. with restricted access for the largest collections.

- The level of processing of the datasets: e.g. with restricted access for harmonised datasets.

- The completeness of the datasets: e.g. with restricted access for datasets including annotated images and associated clinical data.

- Nationality of the requester: e.g. restricted access for non-EU-27 citizens, in consideration to the use of public EU funding for its development.

- Security and reliability in guaranteeing people's rights in the third country.

- Legal possibilities to establish a fee (Open data Directive and future Data Governance Act).

- Conditions defined in the accession agreement. For example, a public sector data provider that only supports open access for the use of the data it provides.

The design of the Access Request evaluation procedure will be conducted with the objective of keeping at reasonable levels both the complexity of the access request process for requesters and the burden of their evaluation for governance costs. Technical review by the CHAIMELEON Access Committee will not be required for projects that have already undergone a qualified peer review process. The Access Request evaluation procedure will also assess if requests for open (free of charge) access fulfil the eligibility criteria.

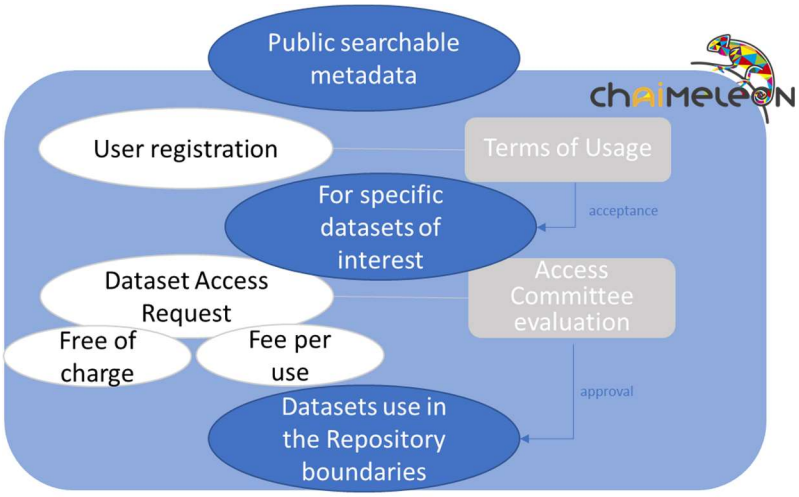

Figure 3: Schematics of data access workflow

5.3 Sharing agreements

An advanced draft "Data Sharing Agreement" has been included in Annex 3 of Deliverable "D2.2 First Complete verification of GDPR compliance of Repository prototype". This draft should be reviewed, and its final version signed before the Repository will be released.

CHAIMELEON Project is aimed at implementing a Repository for health images and committed to its long-term sustainability. Likewise, the project developments for data ingestion, deidentification, curation and harmonisation are aimed at providing access to high quality cancer image datasets, not only during the project duration, but also after project end. Thus, the nature of the project implies that data sharing agreements should cover the following scenarios:

1) - Agreements between the partners involved in the project.

a) The extension of the agreement scope, once it is precisely defined the governance and business model of the Repository for the after-project conclusion phase

b) The possible legal status of the partnership members.

2) Agreements with third parties who join the platform under equivalent conditions to those of the partners.

3) Agreements with third parties for specific purposes:

a) Requests for access to data for studies.

b) Requests for access to data for the purpose of developing technology.

c) Contributions with data sets.

c) Requests for processing without direct access to data.

The first Data Sharing Agreement covers these different possibilities, but it is likely that finally, two types of data sharing agreement will need to be defined:

a) Partnership agreement

b) Data requestor or data donor agreement for a specific purpose.

The DSA integrates the CHAIMELEON framework of Governance addressing ethical and regulatory aspects such as:

▪ Verification that the sets of data have been obtained legitimately in origin. It implies three possibilities:

▪ Patient consent or ethics committee approval for waived consent.

▪Irreversible anonymization of data.

▪ Verification that the acquisition of the dataset is based on a legal act adopted in accordance with current legislation.

The DSA defines the roles may be deployed in relation to the use of anonymized data:

• Data owner. This is that organisation that can assert ownership over the data.

• Data provider. This is the organisation that, whether or not the owner of the dataset, facilitates its use or exploitation.

• Data consumer. The organisation that performs analytics on the data provided by the data provider, either in its own facilities or through direct access to the resources provided by the data provider or data owner.

• User of results. Entity that uses the results obtained by analysing data from the Repository.

The DSA has adopted the Assessment List for Trustworthy AI (ALTAI). ALTAI was developed by the High-Level Expert Group on Artificial Intelligence set up by the European Commission and the Parties that carry out activities related to Artificial Intelligence as well as any other users of the platform shall submit an Artificial Intelligence Ethical Impact Assessment (EIA) using the ALTAI methodology.

The DSA is defined as a legally binding instrument whose purpose is to secure the commitment of partners and future partners to use the data processing platform in accordance with the stipulated conditions including security obligations and the prohibition to re-identify previously anonymised data.

6. Costs

This section describes the cost to operate the services of the platform, the services deployed at hospitals to streamline the process of data collection and curation, as well as the governance costs. The main backend for the central repository is based on cloud resources and it is designed in a way that minimises vendor lock-in. Therefore, costs related to generic public clouds are orientative.

6.1 Central repository

The software for the central repository of CHAIMELEON is implemented using the Code as a Service and DevOps approach which focuses on defining the virtual infrastructure, resources topology, dependencies and configuration actions in machine-actionable files that could reproduce the deployment of the whole platform with minimum user intervention. Moreover, the definition of the platform is independent of the resource provider, as the migration from one provider to another is straightforward.

Therefore, we will define the costs of the central repository based on the following principles:

1. The deployment on a public cloud offering, using Microsoft Azure as a benchmark.
2. The deployment on an on-premise site, using the Total Cost of Ownership of Microsoft Azure as estimation.
3. The cost of operating the platform.
4. The maintenance cost for the software.

Costs 1 and 2 are self-excluding. Costs 3 and 4 do not depend on the target platform and should be added to the deployment cost.

The Virtual Infrastructure of CHAIMELEON for the central repository implies the following resources.

- Front End: 1 XLarge node (8 VCPUs and 32 GB RAM), equivalent to a D8 v3 in Microsoft Azure.
- Processing Nodes: 2 XLarge nodes (8 VCPUs and 32 GB RAM) and 2 XLarge-NVidia100 nodes (8 VCPUs, 32 GB RAM, 1 NVIDIA V100), equivalent to a D8 v3 and a NC8as T4 v3 in Microsoft Azure.
- Storage Nodes: 3 XLarge nodes (8 VCPUs, 32 GB RAM), equivalent to a D8 v3 in Microsoft Azure.
- Storage: 100 TB SDD.

The following table depicts the cost for those requirements on a yearly basis. Larger plans could have a reduction on the public cloud costs. The values for the TCO are indicative.

Service type	Estimated Yearly cost	
	Azure	TCO
Regular Nodes	€14.412,70	€7.305,12
GPU Nodes	€9.454,33	€6.707,04
Storage Accounts	€17.725,83	€40.242,24
Electricity	€0,00	€509,65
Datacenter	€0,00	€27.709,09
Network	€0,00	€944,29
IT Staff	€0,00	€2.729,27
Total Cost	**€41.592,86**	**€54.254,40**

Table 1: Cost of the resources in Microsoft Azure and Total Cost of Ownership (TCO) equivalence.

The cost of operating the platform is additional to the cost of operating the resources. The cost of operating the resources is already considered in the previous deployment costs. Operating the platform will require:

- Active monitoring of the services.
- Security patches and updates to the resources and the Docker containers.
- Manage user tickets.
- Manage user enrollment.
- Update Docker Images.
- Backup-ing and rollout of new versions.
- Security incident management.

The effort required is estimated in 0.25 FTE of a senior engineer, which would be on the order of 14K€ yearly.

The cost of maintaining the software is much more difficult to evaluate, as changes in the versions of the software dependencies may have a great impact in the long term. The maintenance is envisaged to guarantee:

- The compulsory substitution of dependencies that require changes in the software due to security issues.
- Minor updates for compatibility with changes in third party equipment at the providers.

The costs for this last concept could increase if new features, new providers or new tools would be considered. As a minimum, we envisage the need of 0,13 FTE of a senior engineer, which will be in the order of 7K€ yearly.

Thus, the estimated annual costs for the central repository IT services correspond to 54.240€ as TCO for infrastructure plus 21.000€ for personnel.

6.2 Data preparation process

The data preparation process relies on the *Medexprim Suite*™ solution deployed within hospitals, used for the data collection, de-identification and curation of data.

Maintenance of the operational model within the same scope (no additional site or functionality) comprises provision for our hot-line, user support, bug fixes estimated to 0,2 FTE per site of a mixed profile (junior/senior engineer) in the order of 10k€ yearly per site, that is 100 k€ yearly for all 10 sites.

This does not include any new development or deployment of new sites.

6.3 Central web platform

The central web platform provides the tools to access the data stored within the central repository, to execute new analysis pipelines using the algorithms developed in the timeframe of the project and to mine all the information stored within the database. Therefore, the main costs of the central web platform are the following:

- Support & Maintenance: The effort required is estimated, as a minimum, in 0.20 FTE of a junior engineer, which would be on the order of 6K€ yearly. This is a continuous service.
- Clinical applications engineering to explain the users the different functionalities and use cases of the platform: The effort required is estimated, as a minimum, in 0.20 FTE of a senior engineer, which would be on the order of 10K€ yearly. This is a service on demand.
- Software development to fix bugs in the platform: The effort required is estimated, as a minimum, in 0.10 FTE of a junior engineer, which would be on the order of 5.5K€ yearly. Based on our previous experience in the installation and deployment of the platform, a ratio of 5% of bugs is estimated, so we calculate 0.10 FTE in order to fulfil all the bugs types and requirements. This is a service on demand.

Thus, the estimated annual costs for the central web platform maintenance and technical support service correspond to 21.500€.

6.4 Governance structure

An estimation of the annual governance costs for the Repository is presented in Table 2. It corresponds to a cost minimisation model for the first year of operation after project end. The operations will be targeted to ensure quality of operations of the Repository, in full legal compliance and its promotion to increase the user base and acquire new data providers. The Governance costs include the management, promotion and legal compliance services.

Governance Service	Description	Annual Cost
Directorate	1 Director (0,1 FTE)	€ 10.000
Executive Team	2 FTE	€ 70.000
Ethics Committee	2 members (0,1 FTE each)	€ 14.000
Advisory Panel	4 members, in kind except for services to Data Access Committee	€ 5.000
Promotional services	Dissemination actions	€ 10.000
Legal services	Contracts and agreements	€ 15.000
Total Cost		€ 124.000

Table 2: Cost of governance of the Repository

7. Revenues

7.1 Proposed business model

Understanding users´ needs

For designing an appropriate business model, it is crucial to have a good understanding of the needs of the main users, their incentives for adopting new solutions, their most burning needs and their willingness to pay.

Project partner QUIBIM is a high-tech enterprise specialist in AI-based tools for Radiomics. They have responded to a questionnaire elaborated by MATICAL, from their perspective as a potential user of cancer image datasets for AI tools development and testing. Their answers can be found in the Annex “Responses by QUIBIM to the Questionnaire for AI researchers”. They are very illustrative of the characteristics that datasets in CHAIMELEON should fulfil for meeting most urgent needs of the industry, and thus being commercially viable

The main conclusions from this interview are that valuable imaging data for AI development needs to meet requirements in terms of heterogeneity of the datasets, consistency of formats, size at large volumes, among others. While providing access to this type of datasets is in itself valuable, functionalities of the repository will determine the additional workload (and thus cost) needed for these datasets to be exploitable in the AI training, by the user enterprises. Such functionalities relate, for instance, to browse tools that facilitate understanding what type of data is contained in a collection (type of images and clinical conditions) or the level of heterogeneity by population characteristics in a dataset, as well as processing tools that improve harmonisation. Finally, another very important remark is that depending on the type of AI tool aimed, access to associated clinical data may not be crucial, and just the high-quality image datasets may be the main sought resource.

Understanding our unique proposition: CHAIMELEON Business Model Canvas

Figure 4: CHAIMELEON Business Model Canvas

Defining the key elements of CHAIMELEON business model:

The future commercial exploitation of CHAIMELEON Repository may be as (i) an isolated Repository providing access to datasets from different hospitals which agree to the data provision conditions, or (ii) as part of a larger infrastructure under which different existing imaging repositories are federated, with a common centralised governance. Both possibilities are currently contemplated, nevertheless, in the following only the scenario of the CHAIMELEON Repository as an independent entity will be considered in this initial plan for sustainability.

Setting-up the CHAIMELEON Repository is currently the goal of the H2020 CHAIMELEON Project. Advancing from the R&D prototype to an exploitable data infrastructure will require several important steps beyond the scope of this project, namely:

- To constitute a legal entity responsible for the Governance of the Repository, as a not-for-profit entity of public, private or hybrid nature. Its purpose is to facilitate the access of public and private organisations to health datasets for R&D purposes, under a fee per use model.

- To create the legal framework (data sharing agreements, data processor and data controller agreements, etc.) that legitimates this entity as facilitator of access to the datasets provided during the H2020 project execution, and to define the conditions for reimbursement to the data providers.

- To create the legal framework (service provision agreements, IPR terms, etc,) that rules the use of technologies by the H2020 technical project partners, as integrating components of the CHAIMELEON platform.

- To implement its governance structure, at a scale and complexity coherent with the Repository size of operations, along its developmental and consolidation stages as a pan European repository

- To disseminate and promote the Repository through the most appropriate channels and targeted to its main stakeholders, as described in section 4.3 *Stakeholders outreach.*

Fixed costs	
Technical costs	€ 196.754,00
Central IT resources	€ 54.254,00
Technical staff for maintenance and user support	€ 42.500,00
Technical staff for support of sites already deployed	€ 100.000,00
Administrative, financial and management costs	€ 75.000,00
Promotional costs	€ 20.000,00
ELSI costs	€ 29.000,00
Total fixed costs	**€ 320.754,00**
Variable costs	
Technical costs for new data provider	€ 40.000,00
Fee for data provision	€ 30.000,00
Total variable costs	**€ 70.000,00**
TOTAL COSTS	**€ 390.754,00**
INCOMES	
Public funding	€ 300.754,00
Private sponsorship	€ 60.000,00

Fee for data usage	€ 30.000,00
Total incomes	**€ 390.754,00**

Table 3: Preliminary assessment of CHAIMELEON Repository expenses and incomes

An initial forecast of annual expenses and incomes has been elaborated, for the first year of operation of the Repository, after project end. This forecast makes assumptions of costs minimisation aligned with the overall objective of promoting Repository sustainability. The operational costs include the fixed costs related to:

- the central infrastructure and the web platform (cost of IT resources and cost of technical personnel for maintenance and software updates, and user support)

- the cost of maintenance of the infrastructure for data ingestion and curation from the data provider sites, under the assumption that the connected hospitals will continue providing new datasets.

- the managerial costs related to the Directorate, executive team and Advisory Board.

- the promotional costs related to the executive team and professional hired services.

- the ELSI costs related to the Ethics Committee and professional legal services for elaboration of contracts and agreements with providers and users.

The operational costs also include variable costs related to the incorporation of new data providers. This forecast uses the assumption of one new hospital connected in year one after project end, with a fee for data provision of 30.000€ up front to cover the hospital costs for making data accessible.

Costs related to continuous improvement and novel development for upgrading the Repository functionality are not contemplated in this forecast, as it corresponds to a cost minimisation scenario.

The total costs for the Repository operation in year 1 after project end are estimated at 320k€. A hybrid approach for sustainability has been proposed accordingly, assuming a combination of public funding (e.g. European and/or National funds for healthcare research and/or digital transformation), private sponsorships (e.g. large enterprises of the medical imaging sector) and paid services. We assume that paid services will be at a very incipient level in year 1, corresponding to 10 clients at an average annual fee of 3.000€

In the next period, we will assess the validity of these preliminary estimations by benchmarking with access costs in other Imaging Biobanks, and by analysing the global market size for health imaging datasets for research.

7.2 Public funding

The public funding for sustainability of the CHAIMELEON Repository will be along the following main lines:

a) R&D funding for new developments aimed to incorporate new functionalities in the Repository (research prototype):

The Horizon Europe programme is a potential funding scheme for this type of action. Calls under the Health Challenge will be periodically scrutinised for the identification of relevant funding opportunities, in particular calls under the Mission Cancer initiative.

b) Implementation support funding for advancing the Repository to TRL9:

b.1) The Digital Europe Programme: This programme aimed to offer public co-funding for implementation actions aimed to accelerate digital transformation in all sectors. In particular the call “Federated European infrastructure for cancer images data call” expected to open in the second quarter of 2022 might be an excellent opportunity for co-funding next stages aimed to deploy access of new hospitals to the Repository.

b.2) National co-funding, expected by the Digital Europe programme: EU MS are invited to use European Regional Development Fund (ERDF) funds for co-funding initiatives supported by the Digital Europe Programme. The national or regional ERDF Managing Authorities are responsible for making these decisions.

b.3) Next Generation Europe (NGEU) funds: The NGEU funds are a package of programmes intended to support the recovery from the COVID-19, through a combination of grants and loans. Each EU Member State has established national targets for the related investments. Promotion of digitalisation and the adoption of AI solutions in transversal sectors including healthcare, is an important issue in the European Agenda. Therefore, co-funding to support connectivity of national institutional repositories and imaging biobanks to CHAIMELEON may be an eligible action under this funding scheme.

c) Consolidation and expansion phase:

c.1) Becoming a European Research Infrastructure: Being part of the European Research Infrastructure Consortia (ERICs) is a mechanism towards long-term sustainability, related to international recognition for excellence and access to specific funding schemes from national and European bodies. This possibility will be assessed at due time and in close cooperation with ERIC managers.

c.2) The Connecting Europe Facility (CEF): CEF is a funding scheme for the promotion of growth, jobs and competitiveness through targeted infrastructure investment at European level. This scheme is only eligible for the pan European expansion of mature technologies and solutions. It may be of relevance for CHAIMELEON Repository extensive deployment across Europe in the longer term.

8. Conclusions

In conclusion, we believe that the future of medical research is strictly related to that of biobanking, and the shared information which will rely on the number and diversity of available biomedical and imaging information, cost management and realisation, patient and citizen participation, and national and international institution governance. An effective biobank should offer high-quality and affordable medical imaging for planning research programs that will benefit everyone. Hence, the sustainability plan of the CHAIMELEON Repository aims to ensure its long-term existence beyond the end of the project. The main focus has been dedicated to analyse the context where the CHAIMELEON repository will operate and to identify the main strategies that will need to be implemented for the CHAIMELEON Repository self-sustainability, in relation to its governance, operational costs and funding strategies.

Annex

Responses by QUIBIM to the Questionnaire for AI researchers

Currently, how do you obtain images/data to train your algorithms?

The data we currently use at Quibim to train our models comes primarily from two main sources: (i) data provided by our clinical partners, either directly acquired to them -fully anonymized after signing the corresponding Data Transfer Agreement- or -processing it on their behalf when it contains personal data entering into the corresponding Data Processing Agreement-; (ii) open access databases. Some examples would be the following: Internet Brain Segmentation Repository (IBSR), ADNI (Alzheimer's Disease Neuroimaging Initiative), The Cancer Imaging Archive (TCIA), among others.

What are the key features that you seek in a dataset used to train your algorithms?

The key features that make a dataset robust to be used in AI development are:

- ***Heterogeneity**: the dataset needs to be as heterogeneous as possible by means of population characteristics. Therefore, when applicable, it would require patients from different ages, sexes, nationalities, etc. This, together with the variability in terms of images acquisition (i.e., different imaging centers, scanners, acquisition protocols, among others) will help on building robust and reproducible AI models.*

- ***Consistency**: It would be of high interest to find datasets of images with the same data formats (e.g., all DICOM, all NIfTI) and with structured non-imaging data (e.g., clinical data, molecular data, annotations).*

- ***Balance**: Sometimes this is something difficult to achieve when dealing with rare diseases or outcomes. However, while possible, it would be of interest to find balance across the different classes to predict in an AI-based classification model. In this sense, a high effort is required to collect data from the minority class.*

- ***Size**: AI models in general and deep learning models in specific, require large amount of data to properly represent the reality.*

- ***Dimensionality**: In the medical domain it is known that, to reach the final diagnosis, different tests are done: family history, symptoms, clinical data, imaging data, etc., always representing the real world (i.e., data that can be acquired on each site). Therefore, datasets combining all these sources of information are of highly interest.*

- ***Annotations:** When developing an AI model based on supervised learning (e.g., classification, segmentation) the corresponding annotation is required. Therefore, apart from the proper images their corresponding annotations are needed. It would be also of high interest to have multi-reader annotations in order to evaluate the variability across experts and to evaluate the outperformance (or not) of the AI model against the experts to prove the added value of its use in clinical routine.*

What are the key issues that you experience when you use datasets taken from the available sources? Do you spend time/resources to make these datasets processable by your algorithms?

Sometimes the available sources offer a large amount of data and it is necessary to invest a great deal of time in analyzing what type of data exists, what their characteristics are and whether they are compatible with the ideal dataset we need to train our model. In addition, once the data of interest has been located, downloading it occasionally presents memory limitations and makes it necessary to do it in batches.

Data curation and preparation is also an important task that takes most of the time when developing an AI model, this process includes:

- *Dataset organization by means of preparing the folders with the appropriate naming to facilitate data reading in an automatic way.*

- *Dataset split (train/validation/test) to guarantee variability across the different sets according to the interests (e.g., images from all the vendors in all the subsets, use different vendors in training and test, etc.)*

- *Exploratory Data Analysis: When dealing with clinical data it is of importance to understand the data we are dealing with (data types, missing values, correlations, collinearity, etc.).*

- *Image pre-processing: there are datasets in which it is necessary to apply ad-hoc pre-processing algorithms to have images that are as harmonised as possible.*

What would make your activity easier in terms of accessibility to image datasets? Would you appreciate annotated images? Would you appreciate harmonized/standardized images?

In many cases, the available datasets are linked to a particular anatomical region or disease. It would be very useful to have a common data repository that would allow filtering by these characteristics, making it possible to have all the information available for a patient suffering from, for example, prostate cancer, by applying a simple filter. At the same time, it would be very interesting to be able to apply a search option that allows selecting multiple data, for example, images acquired with a specific magnetic field, with the voxel size in a specific range, etc.

The availability of annotated datasets would be of high interest. If we want to develop a supervised learning model it would be use for both the training and test of the algorithm. In addition, even we are developing an unsupervised learning AI model, some annotations are required to perform its test. In addition, annotated images by multiple readers would be also of utility to assess for inter-reader variability and analyze the performance of the developed model against the variability found.

Finally, dealing with harmonized images would be of high interest. Ideally, this will reduce the need of having such amount of variability in the training datasets by means of different acquisition protocols, manufacturers etc. It will allow the AI models to focus on finding differences in the pathologies per se and forgetting about the multi-vendor variability.

Would access to image datasets linked to clinical data (as aimed in CHAIMELEON) save your company current costs for AI algorithm development, and thus you would be willing to pay a service fee for access?

The added value of the clinical data for the development of an AI algorithm would depend on the end purpose of that model. As much patient-related information is available at the time of model building the best, this would allow to different sources of information. However, if it is not really needed for the end goal of the model, we would not pay an extra fee. Let's comment some use cases:

- *Image segmentation: the main information required is the original images and the annotations (manual segmentations). With this would be enough to train and test the model. However, having some clinical data will allow segmenting the population to know its performance on different scenarios. Imaging we are developing an algorithm for the automatic segmentation of the lungs on CT scans. Having some demographic information such as the age will allow us to know if the models work properly for the age range of interest. At the same time, having information about if the subject is healthy or have any disease its of high interest to know if the models work properly in both healthy subjects and pathological patients.*

- *Image classification: here is where the clinical data has the most interest, specially if the classification is thought to detect a pathology. Historically, there are lots of works based on clinical data, therefore, it's of high interest to analyze the added value of the imaging features when combined with the clinical data. In this case, a company would find interest in paying an extra fee to get also the clinical data.*

9. References

1 - Chalmers, Don. “Has the biobank bubble burst? Withstanding the challenges for sustainable biobanking in the digital era.” *PubMed*, 12 July 2016, https://pubmed.ncbi.nlm.nih.gov/27405974/. Accessed 15 February 2022.

2 - Schacter, Brent. “A framework for biobank sustainability.” *PubMed*, 2014, https://pubmed.ncbi.nlm.nih.gov/24620771/. Accessed 14 February 2022.

3 - Yong, William H., and David Geffen. “Sustainability in biobanking.” *NCBI*, 1 January 2020, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6918833/. Accessed 14 February 2022.

4 - van der Stijl, Rogier. “Recommendations for a Dutch Sustainable Biobanking Environment.” *PubMed*, 27 May 2021, https://pubmed.ncbi.nlm.nih.gov/34042498/. Accessed 15 February 2022.

5 - Vaught, Jim. “ISSN 1947-5543 (Online) | Biopreservation and biobanking.” *The ISSN Portal*, 2021, https://portal.issn.org/resource/ISSN/1947-5543. Accessed 15 February 2022.

6 - Kumar, Awanish. “Virtual global biorepository: access for all to speed-up result-oriented research.” *PubMed*, 2020, https://pubmed.ncbi.nlm.nih.gov/32270405/. Accessed 15 February 2022.

7 - De Souza, Yvonne G. “Biobanking past, present and future: responsibilities and benefits.” *PubMed*, 28 January 2013, https://pubmed.ncbi.nlm.nih.gov/23135167/. Accessed 15 February 2022.

8 - BBMRI-ERIC Directory. *BBMRI-ERIC Directory*, https://directory.bbmri-eric.eu/#/. Accessed 15 February 2022.

9 - Ministerio de Ciencia e Innovación de España. “BOE-A-2021-9488 Resolución de 31 de mayo de 2021, del Instituto de Salud Carlos III, OA, MP, por la que se publica el Convenio con el Ministerio de Ciencia e Innovación, para la participación de España en el Consorcio de Infraestructuras de ...” *BOE.es*, 31 May 2021, https://www.boe.es/diario_boe/txt.php?id=BOE-A-2021-9488. Accessed 15 February 2022.