Spain National Tripartite Event

Open Science in Data Intensive Scientific Communities: Genomics

Salvador Capella-Gutierrez

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

FEDER
Fondo Europeo de Desarrollo Regional
UNIÓN EUROPEA
"Una manera de hacer Europa"

GOBIERNO DE ESPAÑA
MINISTERIO DE CIENCIA E INNOVACIÓN

Instituto de Salud Carlos III

INB Spanish National Bioinformatics Institute

elixir SPAIN

Context

European Commission | European Health Data Space
European Cancer Mission

eosc

1+MG

B1MG

elixir

Global Alliance for Genomics & Health
Collaborate. Innovate. Accelerate.

International Cancer Genome Consortium

ARGO Accelerating Research in Genomic Oncology
International Cancer Genome Consortium

iMPaCT

Instituto de Salud Carlos III

INB Spanish National Bioinformatics Institute

BSC Barcelona Supercomputing Center
Centro Nacional de Supercomputación

elixir SPAIN

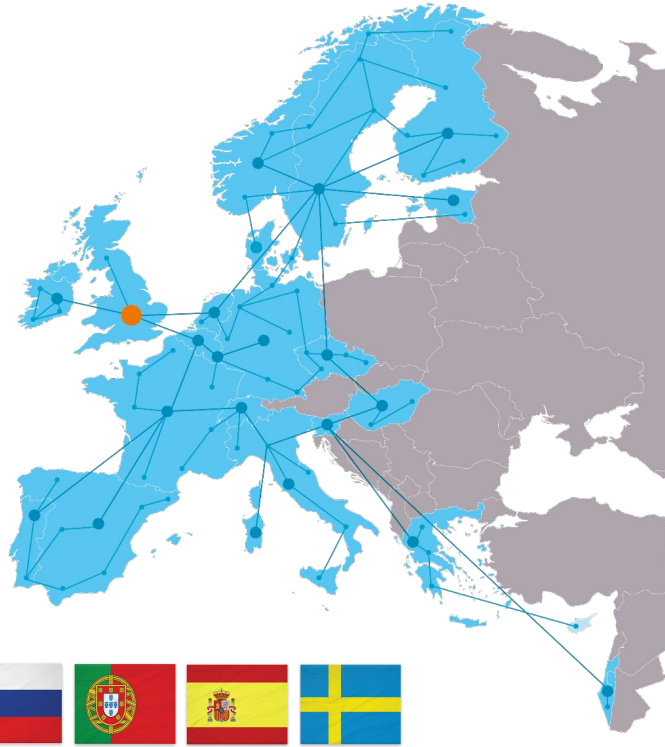
ELIXIR – a distributed infrastructure for life sciences

The goal of ELIXIR is to **coordinate life science resources from across Europe so they form a single infrastructure**.

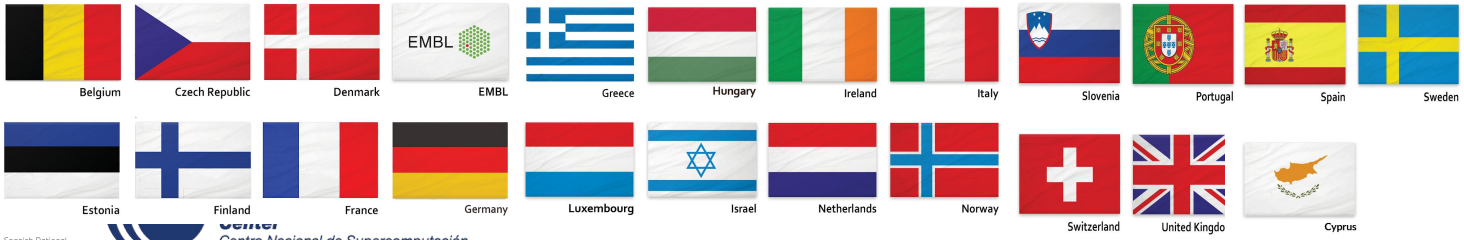
This makes it easier for scientists to:

- Find and share data
- Exchange expertise
- Agree on best practices in scientific research
- Find resources (e.g. databases, software tools, training)

Each node is committed to financially sustain the infrastructure.



ELIXIR Members



Instituto de Salud Carlos III

Spanish National Bioinformatics Institute

Centro Nacional de Supercomputación

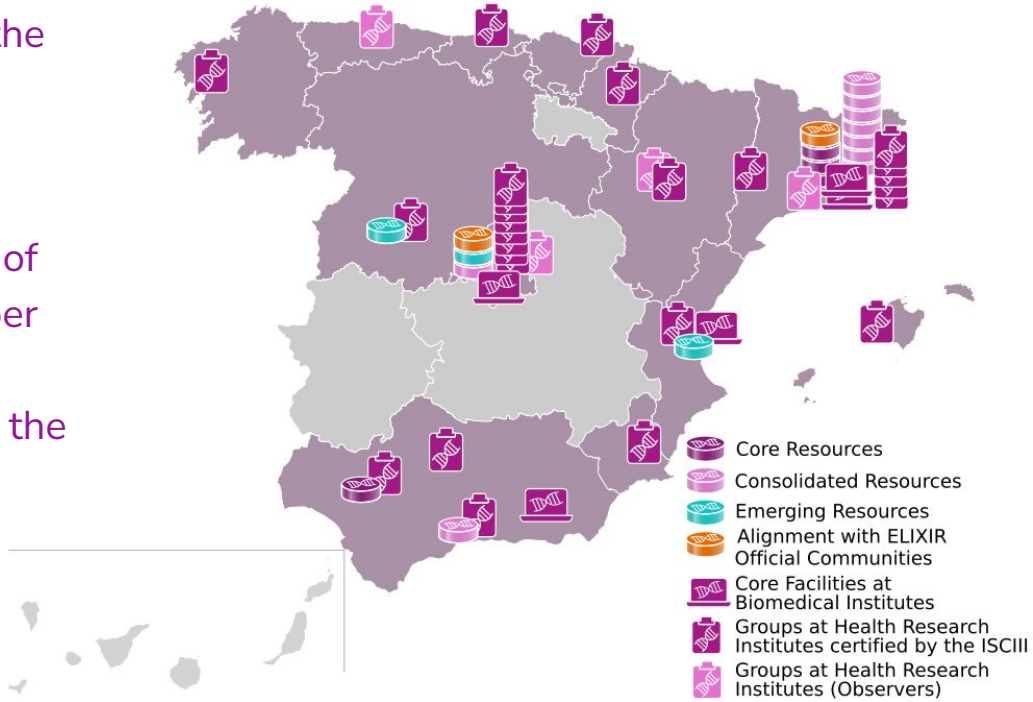
elixir SPAIN

INB - Spanish National Bioinformatics Institute

The **INB**, created in 2003, became one of the technological platforms of ISCIII (2018 - 2021) with two overarching objectives:

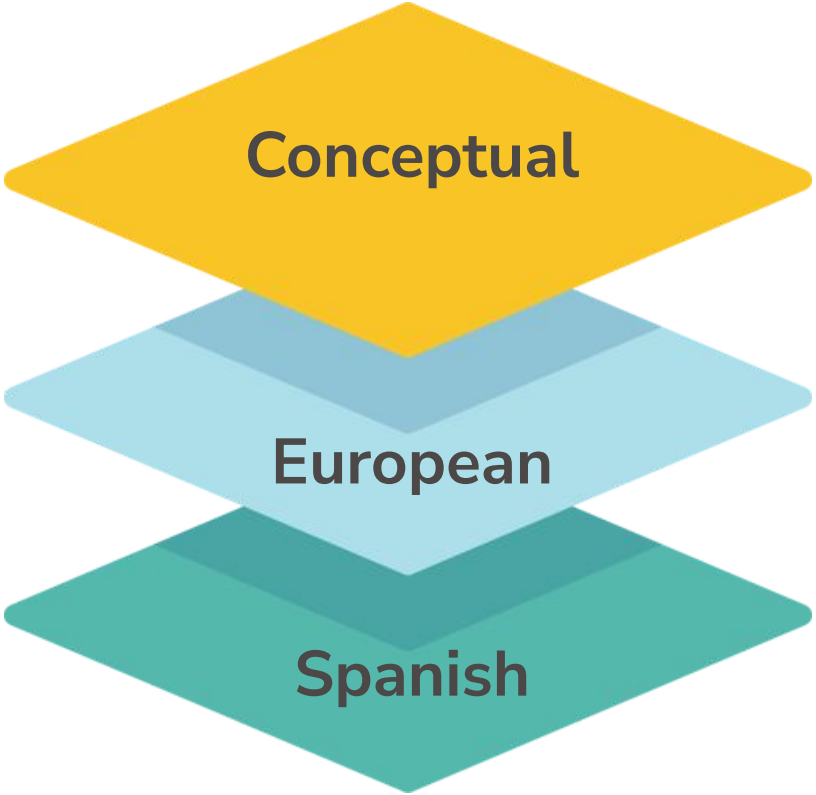
1. Maintain and increase the alignment of the INB with **ELIXIR** looking for deeper synergies.
2. Increase the translational capacity of the INB towards the **Spanish National Health System** (SNS).

INB is nowadays part of **IMPaCT-Data**



Instituto de Salud Carlos III
INB Spanish National Bioinformatics Institute
BSC Barcelona Supercomputing Center
Centro Nacional de Supercomputación
elixir SPAIN

(Some relevant) Projects



HEALTHYCLOUD
Health Research & Innovation Cloud

FAIRplus

European Genomic Data Infrastructure

EHDS
HealthData@EU Pilot

elixir CONVERGE

EUCAIM

eosc | cancer

iMPaCT | Data

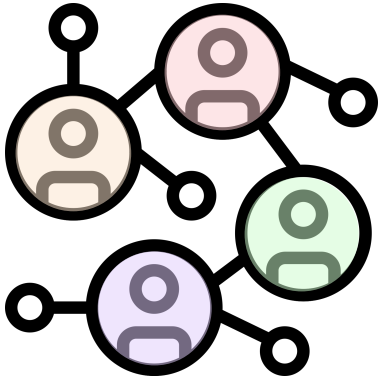
Instituto de Salud Carlos III

INB Spanish National Bioinformatics Institute

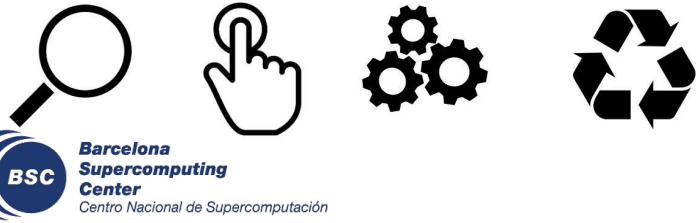
BSC Barcelona Supercomputing Center
Centro Nacional de Supercomputación

elixir SPAIN

Keep concepts on Research Health (Genomics) Data



Findable Accessible Interoperable Reusable



Instituto de Salud Carlos III

INB Spanish National Bioinformatics Institute

BSC Barcelona Supercomputing Center
Centro Nacional de Supercomputación

elixir SPAIN

Aspects to consider in the implementation of the FAIR Principles

- **To be Findable:**
  - F1. (meta)data are assigned a globally unique and eternally persistent identifier.
  - F2. data are described with rich metadata.
  - F3. (meta)data are registered or indexed in a searchable resource.
  - F4. metadata specify the data identifier.
- **To be Accessible:**
  - A1 (meta)data are retrievable by their identifier using a standardized communications protocol.
    - A1.1 the protocol is open, free, and universally implementable.
    - A1.2 the protocol allows for an authentication and authorization procedure, where necessary.
  - A2 metadata are accessible, even when the data are no longer available.
- **To be Interoperable:**
  - I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
  - I2. (meta)data use vocabularies that follow FAIR principles.
  - I3. (meta)data include qualified references to other (meta)data.
- **To be Re-usable:**
  - R1. meta(data) have a plurality of accurate and relevant attributes.
    - R1.1. (meta)data are released with a clear and accessible data usage license.
    - R1.2. (meta)data are associated with their provenance.
    - R1.3. (meta)data meet domain-relevant community standards.

Practical experience on the FAIRification of research data

Define FAIRification Goal

Determine goal for FAIRification, in terms of desired *usability of data* that isn't currently possible

Project Examination

Examine the current state of the project with respect to *FAIRification goal*

Identify Data Requirements

Determine the ***indicators*** that will specify how data needs to be presented/manipulated (wrt FAIR) to fulfil the expected usage

Identify FAIRification Capabilities & Resources

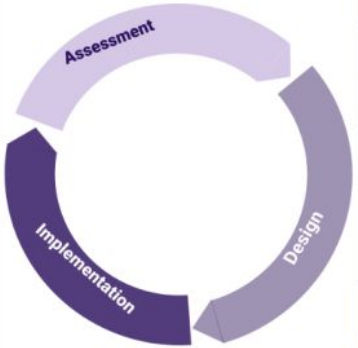
Determine data management ***capabilities*** *(e.g. data annotation, data search and indexing)* ***& resources*** (e.g. tools, databases, vocabulary services) needed to enable and support the FAIRification process

Produce the FAIRification Backlog

Define a prioritized list of target FAIRification outcomes, concentrating on the desired results for the project rather than the specifics of the solution.

Iterative FAIRification Cycles

Define and deliver *practical, achievable objectives* for a single 3 month release cycle that work towards the overall FAIRification goal



1- Assessment

Honest Evaluator to perform pre- and post-FAIRification assessments and to produce the summary report before and after every cycle

2- Design

Work with project representatives to develop a solution plan for the current cycle, which may involve developing new cookbook recipes or re-use existing ones.

3- Implementation

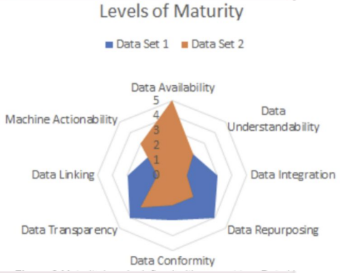
Execute the solution plan within the agreed time-frame. Prepare the resulting FAIRified data for the end-of cycle assessment and inform the Honest Evaluator when ready.

Post-FAIRification Review

Review *outcomes* and assess success against original goals

Disseminate learnings

Developed recipes are shared publicly in the cookbook. If applicable, update FAIRness levels in IMI catalogue. Integrate lessons learnt with other initiatives e.g. Pistoia FAIR toolkit, RDMKit



FAIRplus

imi innovative medicines initiative

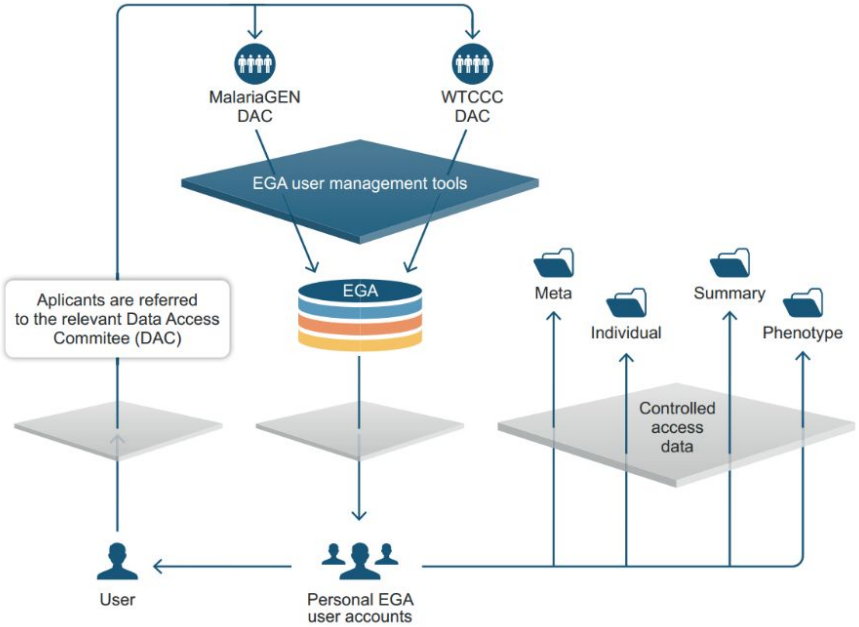
Instituto de Salud Carlos III

INB Spanish National Bioinformatics Institute

BSC Barcelona Supercomputing Center
Centro Nacional de Supercomputación

elixir SPAIN

European Genome-phenome Archive (EGA)



The EGA is a resource for permanent secure archiving and sharing of all types of potentially identifiable bio-molecular and phenotypic data resulting from **biomedical research** projects. EGA contributes and follows international standards.

- Data is provided by **research centers** and **health care institutions**.
- Access is controlled by Data Access Committees.
- Data requesters are researchers from other research or health care institutions.

EMBL-EBI

CRG Centre for Genomic Regulation

BSC Barcelona Supercomputing Center Centro Nacional de Supercomputación

https://ega-archive.org

Instituto de Salud Carlos III

INB Spanish National Bioinformatics Institute

elixir SPAIN

The EGA allows going from data to therapy all around the world

Published: 27 October 2010

The patterns and dynamics of genomic instability in metastatic pancreatic cancer

Peter J. Campbell, Shinichi Yachida, Laura J. Mudie, Philip J. Stephens, Erin D. Pleasance, Lucy A. Stebbings, Laura A. Morsberger, Calli Latimer, Stuart McLaren, Meng-Lay Lin, David J. McBride, Ignacio Varela, Serena A. Nik-Zainal, Catherine Leroy, Mingming Jia, Andrew Menzies, Adam P. Butler, Jon W. Teague, Constance A. Griffin, John Burton, Harold Swerdlow, Michael A. Quail, Michael P. Stratton, Christine Iacobuzio-Donahue & P. Andrew Futreal

Nature **467**, 1109–1113(2010) | Cite this article

3186 Accesses | **885** Citations | **29** Altmetric | Metrics

Cambridge, UK → Houston, T

Beijing, China

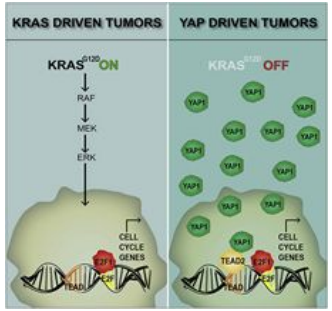
(1) Study deposited EGAS00000000064

★ The paper has over 700 citations

★ Datasets re-used many, many times

Cell

Yap1 Activation Enables Bypass of Oncogenic Kras Addiction in Pancreatic Cancer



(2) Molecular mechanism identified

(3) New therapeutic strategy shaped

Cancer Lett. 2017 Aug 28;402:61-70. doi: 10.1016/j.canlet.2017.05.015. Epub 2017 May 30.

A combinatorial strategy using YAP and pan-RAF inhibitors for treating KRAS-mutant pancreatic cancer

Xiao Zhao [1], Xiuchao Wang [2], Lijun Fang [3], Chungen Lan [2], Xiaowei Zheng [2], Yongwei Wang [1], Yinlong Zhang [4], Xuexiang Han [1], Shaoli Liu [4], Keman Cheng [1], Ying Zhao [1], Jian Shi [1], Jiayi Guo [1], Jihui Hao [2], He Ren [5], Guangjun Nie [6]

Affiliations + expand

PMID: 28576749 DOI: 10.1016/j.canlet.2017.05.015

Instituto de Salud Carlos III · INB Spanish National Bioinformatics Institute · BSC Barcelona Supercomputing Center, Centro Nacional de Supercomputación · elixir SPAIN

Federated EGA

Federated European Genome-phenome Archive

The Federated EGA is envisioned as a network fostering reuse of human biomedical data for research purposes in a federated context. Several ELIXIR nodes have engaged in establishing the federation together with the current EGA core institutions (CRG and EMBL-EBI)

Ciencia / Materia EL PAÍS

JESSICA MOUZO
Barcelona - 23 SEPT 2022 - 05:20 CEST

INVESTIGACIÓN CIENTÍFICA >

El almacén invisible que guarda datos genómicos de un millón de personas

El Archivo Europeo de Genomas y Fenomas, que dispone de 16 petabytes de datos de salud muy sensibles para investigación científica, está custodiado en el superordenador MareNostrum de Barcelona y en Cambridge

LA VANGUARDIA

CRISTINA SÁEZ
BARCELONA

HACIA LA MEDICINA PERSONALIZADA

Barcelona lidera un proyecto científico internacional para investigar enfermedades con datos de un millón de genomas

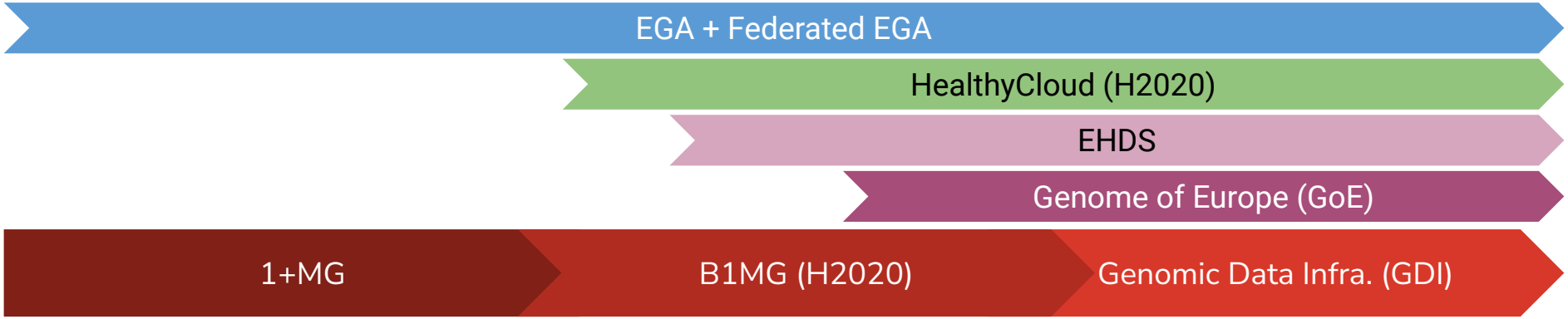
European institutes commit to data access across borders

Research institutes from five European countries have committed to improving the way researchers discover and access sensitive human data across national borders to enable more efficient health research.

Credit: Karen Arnott/EMBL

Instituto de Salud Carlos III · INB Spanish National Bioinformatics Institute · BSC Barcelona Supercomputing Center, Centro Nacional de Supercomputación · elixir SPAIN

Genomics Data at European scale



1+ Million Genomes

22 EU countries, the UK and Norway have signed the Declaration 'Towards access to at least 1 million sequenced genomes in the EU by 2022'.

Beyond 1 Million Genomes

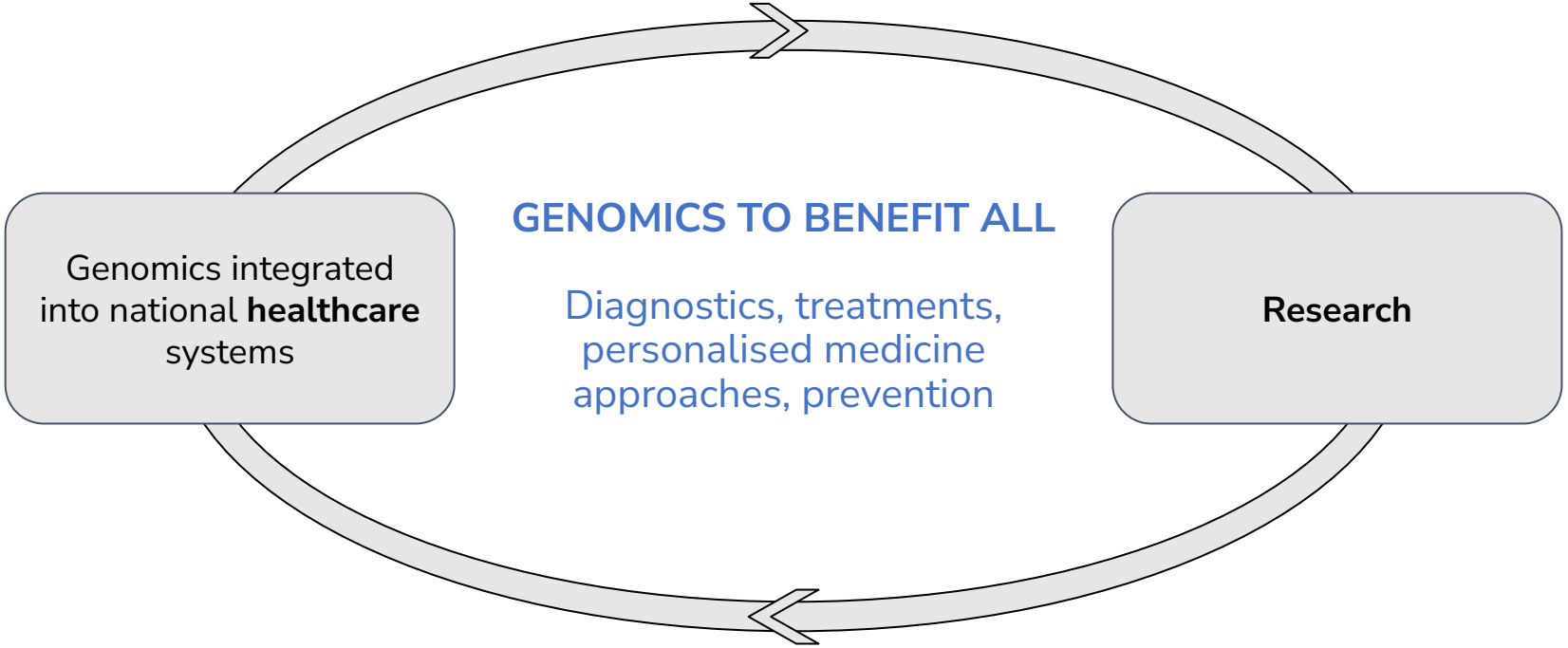
B1MG will go 'beyond' the 1+M genome target and 'beyond' the signatory countries, collaborating with an array of international initiatives and consult a range of stakeholders to support the creation of a pan-European genome-based health data infrastructure.

GDI

Will support the 1+MG Initiative to fulfil the goal to provide a cross-border federated network of national genome collections associated with other relevant data for advancing data-driven health and care solutions to benefit citizens of Europe.

Instituto de Salud Carlos III · INB Spanish National Bioinformatics Institute · BSC Barcelona Supercomputing Center, Centro Nacional de Supercomputación · elixir SPAIN

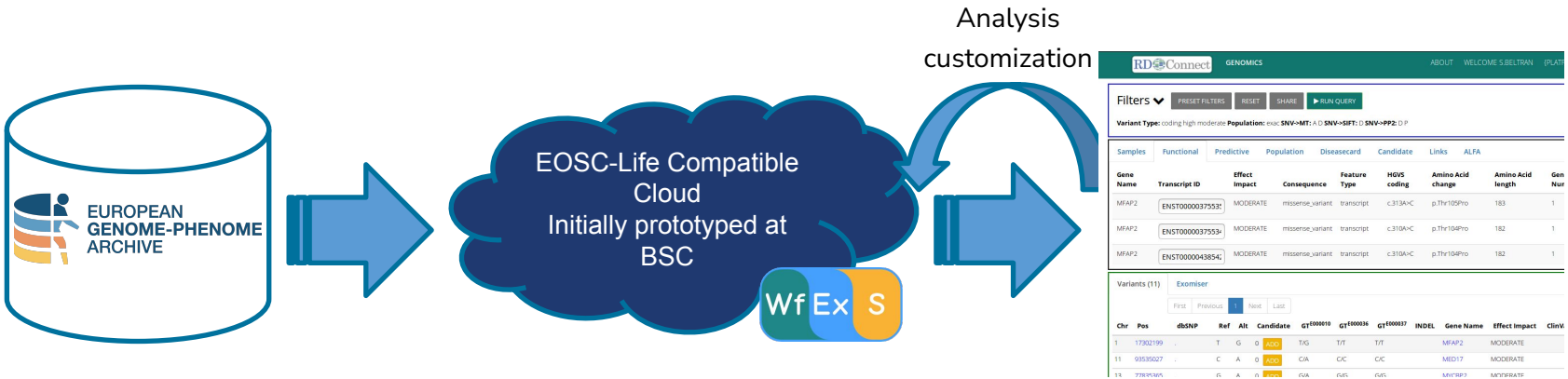
Overall ambition



Using access-controlled FAIR data for research

EOSC-Life

Demonstrator #7. An **integrative analysis pipeline** of genomic and transcriptomic human data for disentangling the genetic origin of a rare-disease in the context of the European Open Science Cloud.



- ➔ Controlled Access.
- ➔ Phenotypic data.
- ➔ SW Containers as part of metadata.

- ➔ Analysis provenance (WF)
- ➔ Workflows native execution.
- ➔ Use of SW containers.
  + selection of best WFs

- ➔ Controlled Access.
- ➔ Analysis interpretation.
- ➔ Analysis re-run.

Instituto de Salud Carlos III · INB Spanish National Bioinformatics Institute · BSC Barcelona Supercomputing Center, Centro Nacional de Supercomputación · elixir SPAIN

Spanish

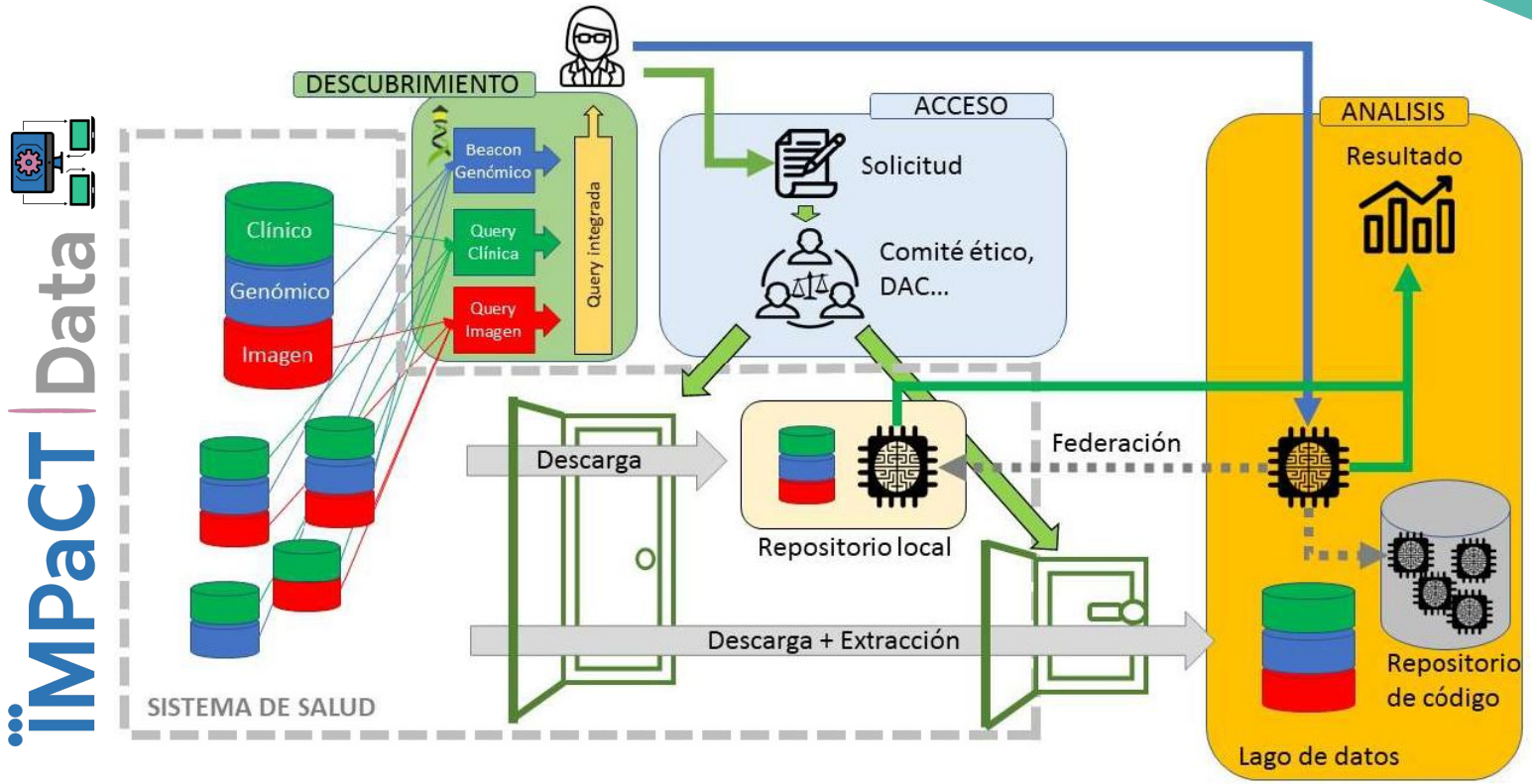
Health data at different scales

IMPaCT (Infrastructure for Personalized Medicine associated to Science and Technology) aims to set the foundations of the future national genome medicine in Spain.

- Divided into three complementary programs: Preventive Medicine, Genomics Medicine and **Data Science**.
- BSC coordinates the Data Science programme (46 partners + 20 other organizations), which aims to focus in the interception of clinical information, genomics data and biomedical imaging.
- The Data Science program will strongly rely in existing standards and mechanisms to favour **interoperability**, e.g. **OMOP**, **FHIR**, **openEHR**.
- Connects with other efforts like the **National Plan for Natural Processing Language** (among others domains from electronic health records).

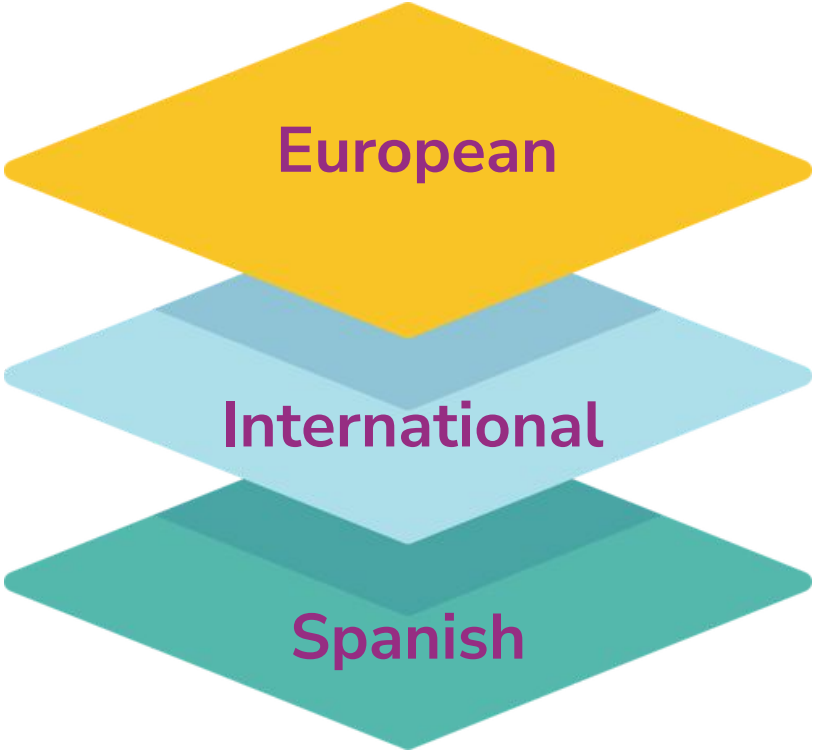
Spanish

Health data at different scales



Instituto de Salud Carlos III
INB Spanish National Bioinformatics Institute
BSC Barcelona Supercomputing Center
Centro Nacional de Supercomputación
elixir SPAIN

Perspective



- Genomics data would be generated by the healthcare systems.
- Strong focus on facilitating the pheno-clinical annotations to genomics data
- Existing solutions for federated access to genomics, e.g. EGA, beacon, can be potentially extended to other domains.
- Strong connection with 1+MG/B1MG & associated projects.

https://inb-elixir.es http://elixir-europe.org

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

FEDER
Fondo Europeo de Desarrollo Regional
UNIÓN EUROPEA
"Una manera de hacer Europa"

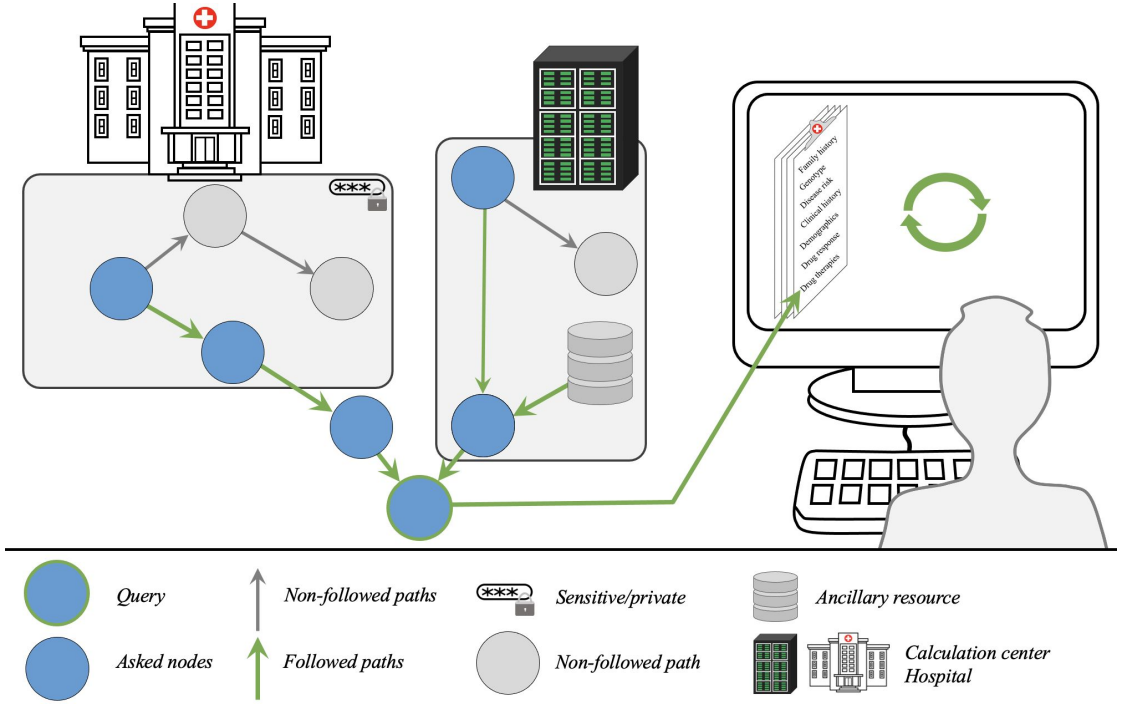
GOBIERNO DE ESPAÑA
MINISTERIO DE CIENCIA E INNOVACIÓN

Instituto de Salud Carlos III

inb Spanish National Bioinformatics Institute

elixir SPAIN

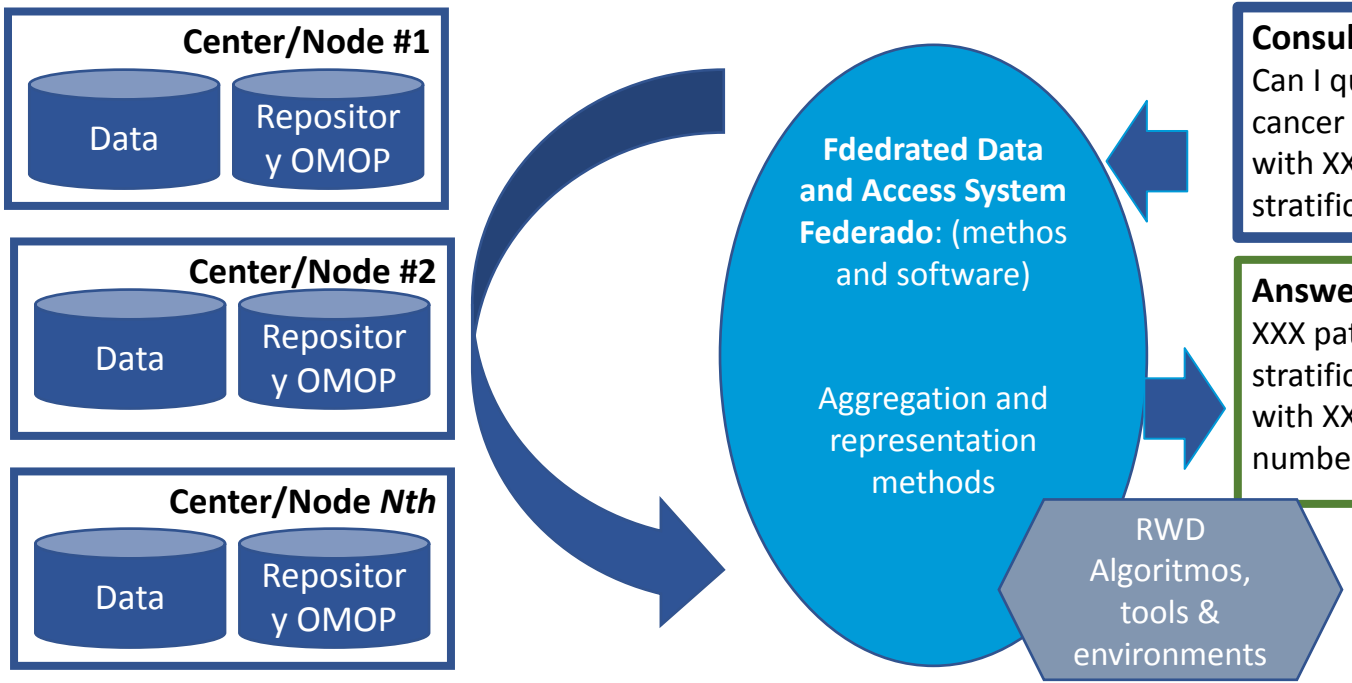
The overall goal :Transparent access to normalised and interoperable data and the appropriate compute



Patient dossier: healthcare queries over distributed resources

Miguel Vazquez[1,2] and Alfonso Valencia[1,3*]

Example of potential use: Building a Virtual Cohort in a Federated Environment



Consulting a real case:
Can I quantify the number of cancer free cases after treatment with XX and YY for a given stratification

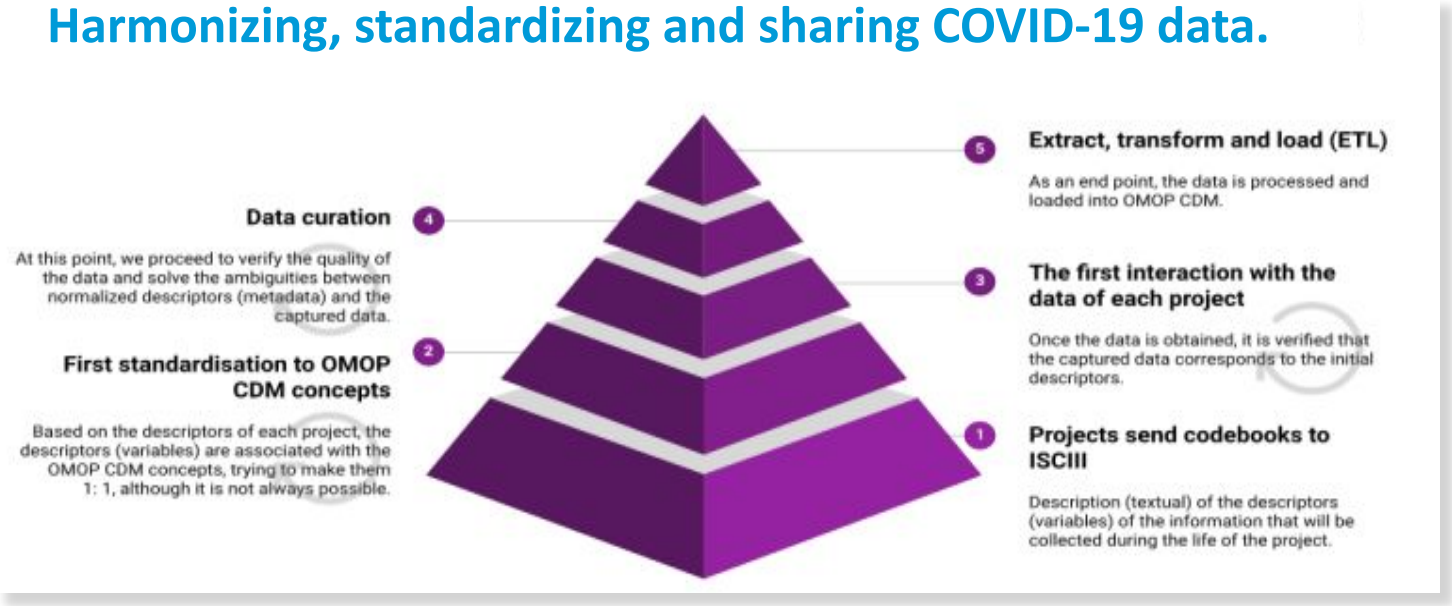
Answer:
XXX patients with the specific stratification have been treated with XX and YY resulting in YY number of recoveries

A federated operation is executed in each of the centers / nodes on their private OMOP repositories – after consultation with the ethic committee(s).

Requires interoperable data

Ejemplo de la dificultad de la interoperabilidad semántica.

Harmonizing, standardizing and sharing COVID-19 data.



60 funded projects by ISCIII: 51 clinic-epidemiological + 9 host+viral sequencing

A team of 6 experts curator for 2 years + informatics support to normalise 1/3 of them