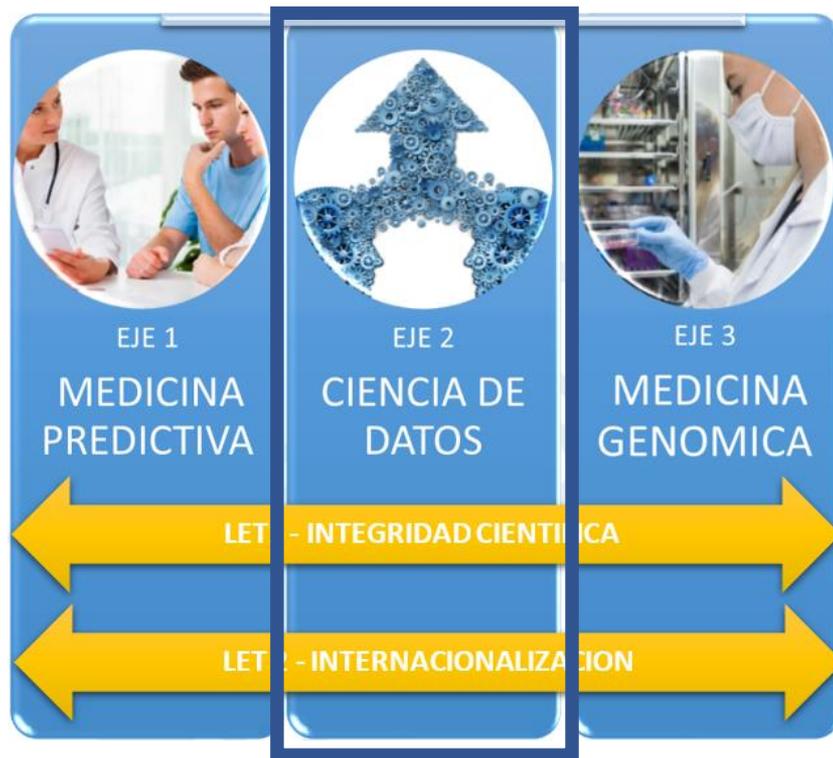


IMPACT-Data

28/06/2021 (online)



Ejes Estratégicos IMPaCT

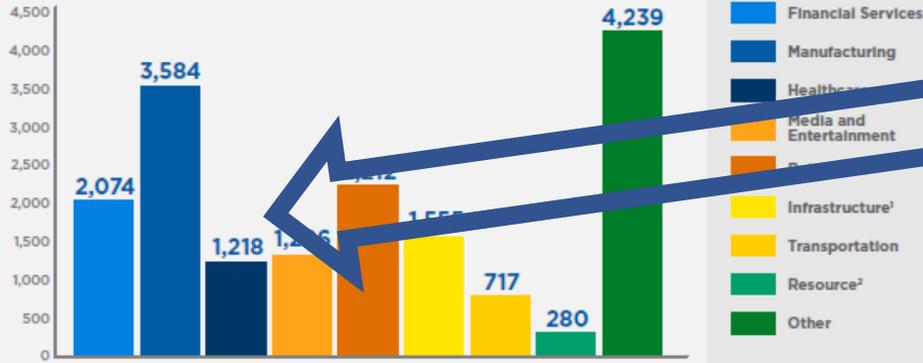


IMPACT-Data - Imagen Clínica & Informática Médica (21/06/2021)

Healthcare data is the fastest growing

- Fast growing
- Heterogenous
- Small providers
- Noise
- Complex
- Interlinked

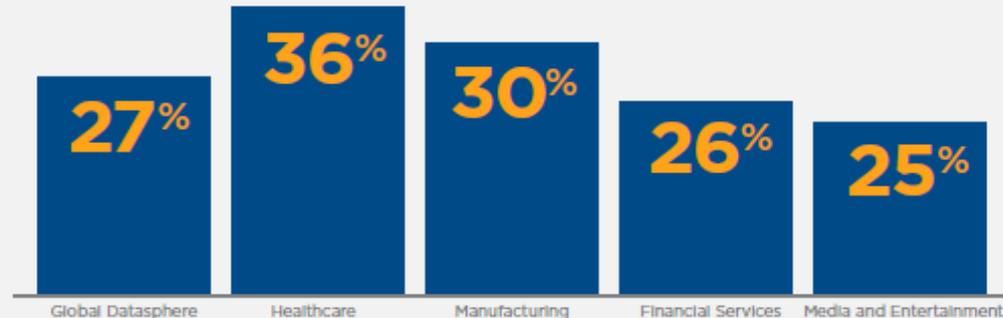
2018 Enterprise Datasphere by Industry (EB)



¹ Infrastructure includes Utilities, Telecommunications

² Resource includes Oil and Gas (Mining), Transportation of oil & gas through pipelines or shipping, Resource industries, Petroleum and coal manufacturing/refining

2025 CAGR

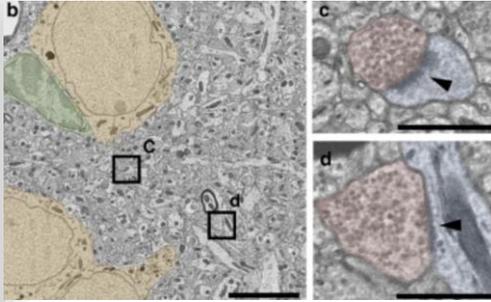


Bio-data



Magnetic Resonance Imaging

7 Tesla MRI of the ex vivo human brain at 100 micron resolution.
The scans, performed over ~100 hours
8 TB dataset



Electron microscopy

Imaging of a 1 mm³ of mouse neocortex by piTEAM. Over 26,500 ultrathin tissue sections from the same block were imaged, yielding a dataset of more than 2 petabytes
2 PB experiment

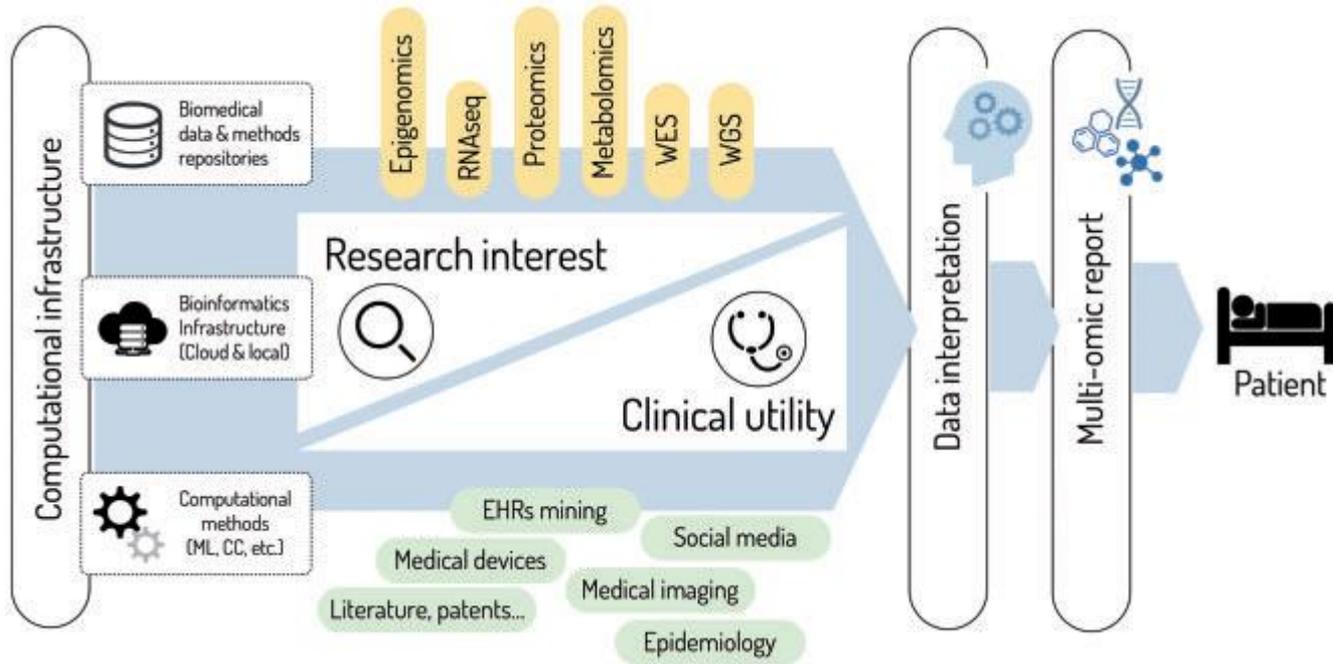
Declaration for delivering cross-border access to **genomic database**

- 1 million **genomes accessible** in the EU by 2022
- Linking access** to existing and future genomic database across the EU
- Providing a sufficient scale for **new clinically impactful** associations in research

European '1+ Million Genomes' Initiative
Cooperation involving all 24 countries
120 PB data set
EGA 13.9 PB

1 ExaByte = 10³ Petabytes = 10³ Terabytes = 10¹⁵ bytes ó 10 Millones de iPhones

Personalized Medicine schema



Precision medicine needs pioneering clinical bioinformaticians

Gonzalo Gómez-López, Joaquín Dopazo, Juan C Cigudosa, Alfonso Valencia, Fátima Al-Shahrour ✉

Review | Published: 30 July 2012

Getting personalized cancer genome analysis into the clinic: the challenges in bioinformatics

Alfonso Valencia ✉ & Manuel Hidalgo

Genome Medicine 4, Article number: 61 (2012) | [Cite this article](#)

2019

Objetivos

Programa de Ciencia de Datos: IMPaCT-Data

El proyecto se centra en el desarrollo de un [entorno de integración y análisis de datos](#) que incluya la capacidad tanto para resolver preguntas provenientes de grupos clínicos, incluyendo las relativas a un caso de uso sobre COVID-19 y formuladas por los Programas de Medicina Predictiva y Medicina Genómica. Para ello se propone un *ciclo de prueba, evaluación e implementación de sistemas computacionales en los entornos sanitarios*.

IMPaCT-Data tiene la vocación de crear **la infraestructura para el uso secundario** (investigación y formulación de políticas) de los datos de los sistemas sanitarios españoles: *historias clínicas electrónicas, imagen médica y repositorios genómicos* y contribuir al sistema sanitario con el conocimiento y metodología generado.

Objetivos (I)

- 1** Desarrollo de la **primera iteración de una plataforma de datos federada** que comprenda datos médicos, genómicos y de fenotipo, normalizada y multipropósito.
- 2** Desarrollo e implementación de una **primera versión de una infraestructura de computación en la nube** para dar soporte inicialmente a los desarrollos de IMPaCT.
- 3** Desarrollo e implementación de **protocolos, métodos y sistemas integrados de análisis de datos** basados en la infraestructura de bases de datos, métodos, sistemas de evaluación y mecanismos de FAIRificación.
- 4** Desarrollo inicial para **monitorizar los procesos de tratamiento y evaluación de la calidad de los datos** durante los procesos de transformación y manejo de los mismos.

Objetivos (II)

5 Desarrollo inicial de sistemas para la **extracción (semi-)automática y segura - con garantía de confidencialidad** - de información mínima normalizada y pseudonimizada a partir de los sistemas de información sanitarios, incluyendo la **Historia Clínica Electrónica (HCE)**, con el desarrollo de **protocolos de interoperabilidad y normalización basados en estándares internacionales**.

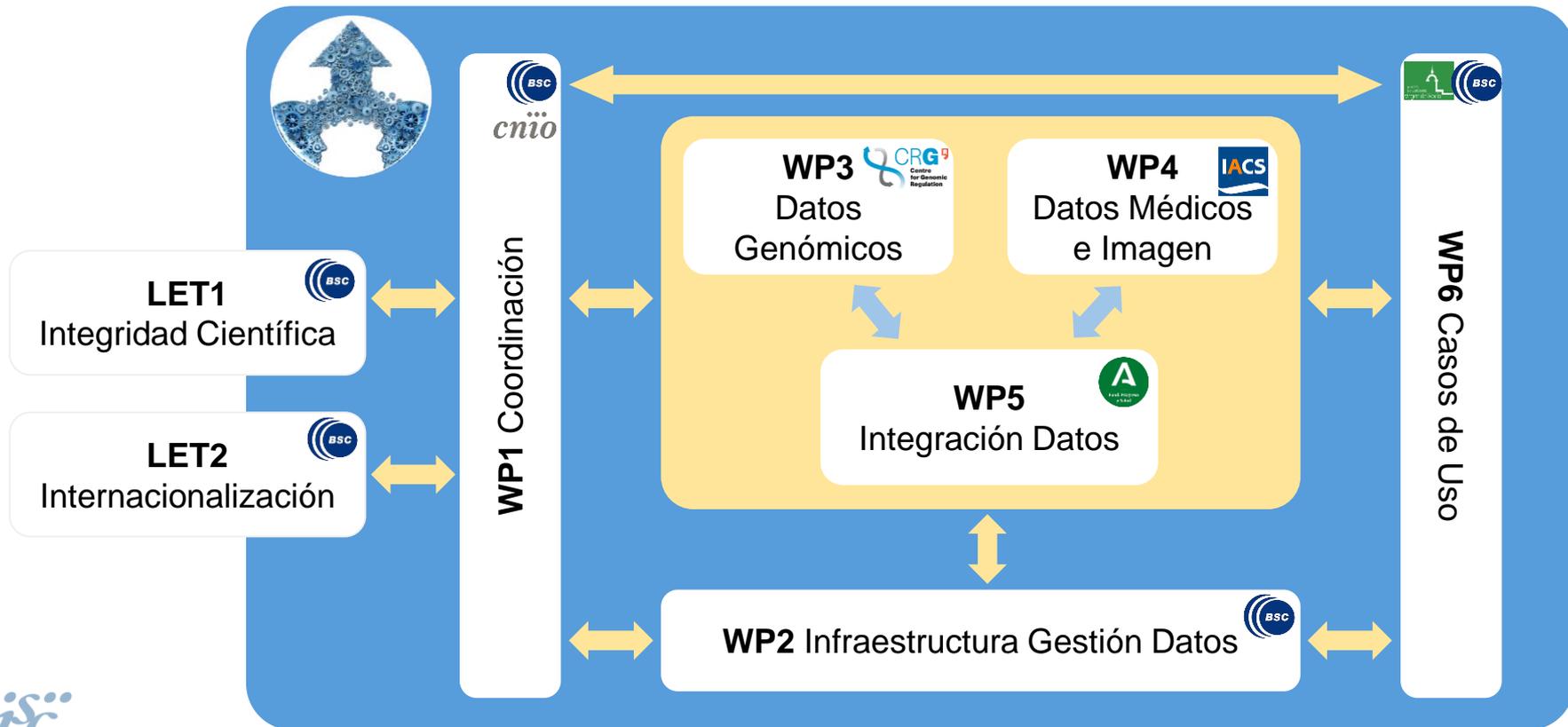
6 **Incorporación de información genética y genómica**, tratamiento de la información para su **consulta rápida, encriptación** para mantener la confidencialidad e **indexación** y agregación de datos.

7 **Liderar el portafolio de recursos bioinformáticos ofrecidos por España a ELIXIR**, incluyendo software, bases de datos y plataformas enfocados al ámbito de salud humana, y organizados en consonancia con el objetivo general de IMPaCT-Data.

Objetivos (III)

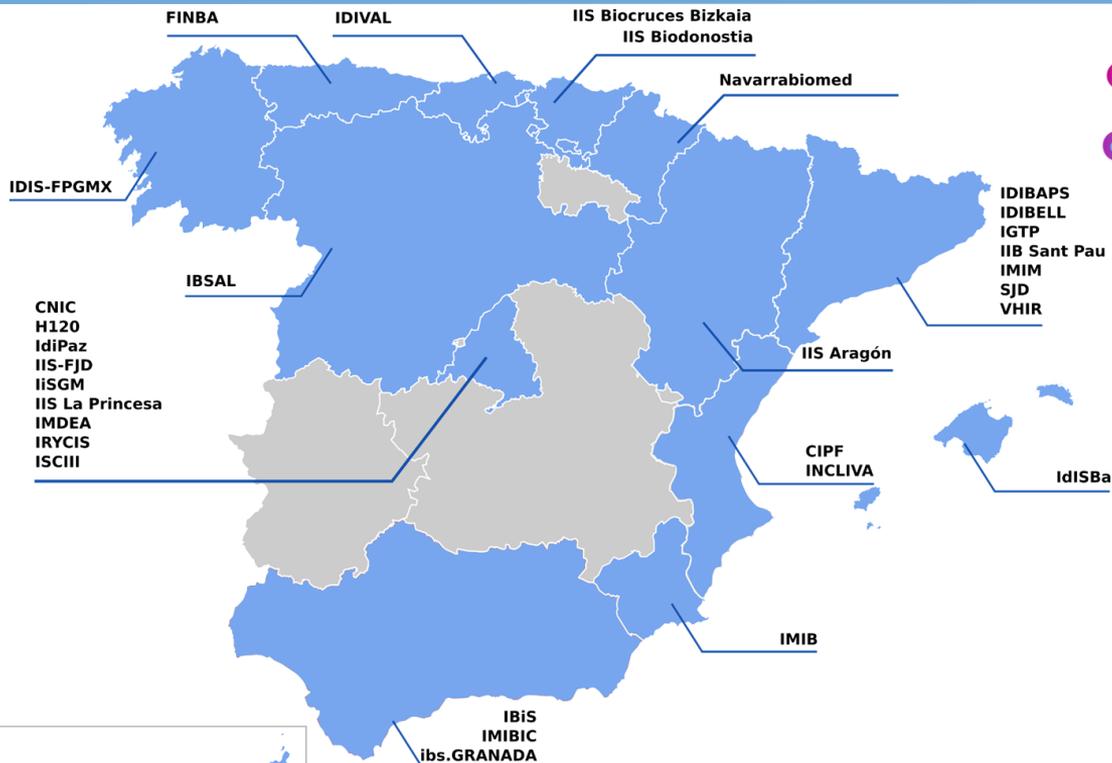
- 8 Extracción de información cuantitativa de imágenes médicas**, tanto basadas en características de la imagen (radiómica) como las que están relacionadas con fenómenos patofisiológicos, para su uso e integración con otras fuentes de datos.
- 9 Desarrollo de prototipos de integración de los resultados del análisis genómico y de imagen con la información normalizada extraída de las HCE**, en los formatos adecuados para el análisis humano y computacional.
- 10 Implementación de ecosistemas de “demostradores” sobre funciones avanzadas de interoperabilidad de información traslacional**, permitiendo la inclusión de aportaciones de empresas, entidades y grupos académicos que, comprometidos con el proyecto, complementen la infraestructura de interoperabilidad.
- 11 Evaluación y ejecución concertada** en distintas localizaciones de los demostradores de gestión, en colaboración directa con la red **TransBioNet** y con otros actores del **sistema de salud**.

Plan de trabajo



Entidades y grupos

Nodos TransBioNet @ IMPaCT-Data



Grupos TransBioNet

- *CIPF*
- *CNIC*
- *FIISC*
- *FINBA*
- *FPS-GENyO*
- *H12O*
- *IBiS*
- *ibs.GRANADA*
- *IBSAL*
- *IDIBAPS*
- *IDIBELL*
- *IdiPaz*
- *IDIS-FPGMX*
- *IdISBa*
- *IDIVAL*
- *IGTP*
- *IIB Sant Pau*
- *IIS Aragón*
- *IIS Biocruces*
- *IIS Biodonostia*
- *IIS La Princesa*
- *IIS-FJD*
- *IiSGM*
- *IMDEA Alimentación*
- *IMIB*
- *IMIBIC*
- *IMIM*
- *INCLIVA*
- *IRYCIS*
- *Navarrabiomed*
- *SJD*
- *VHIR*
- *ISCIII - Unidad de Bioinformática*¹

Grupos INB/ELIXIR-ES

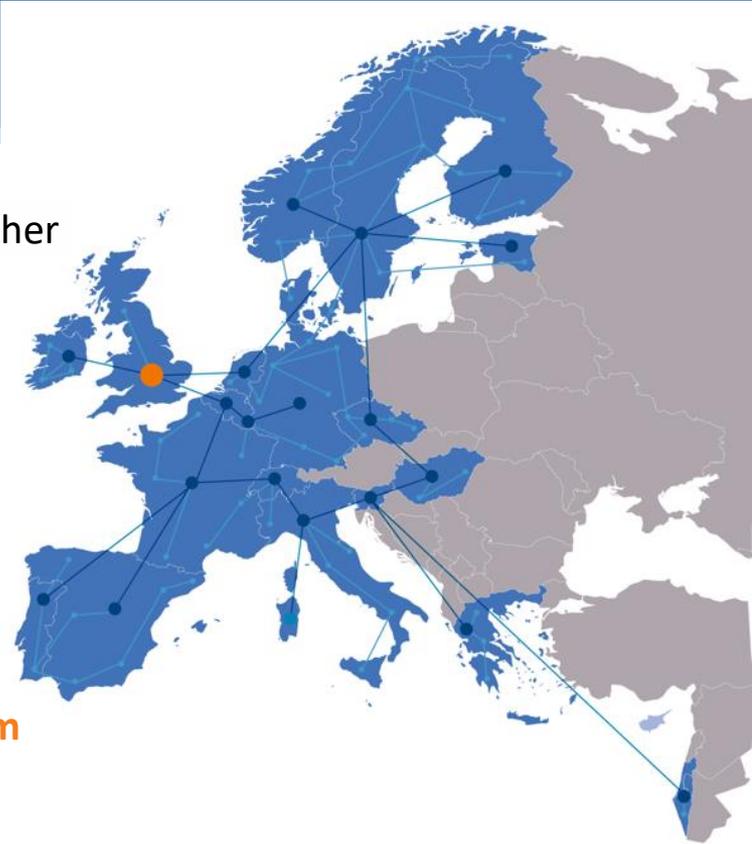
- *BSC*
- *CRG*
- *FPS*
- *CNIO*
- *CSIC*
- *IRB Barcelona*
- *UMA*
- *UPF*

ELIXIR - data for life

ELIXIR is an intergovernmental organisation that brings together life science resources such as

- databases
- software tools
- training materials
- data standards
- compute resources

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

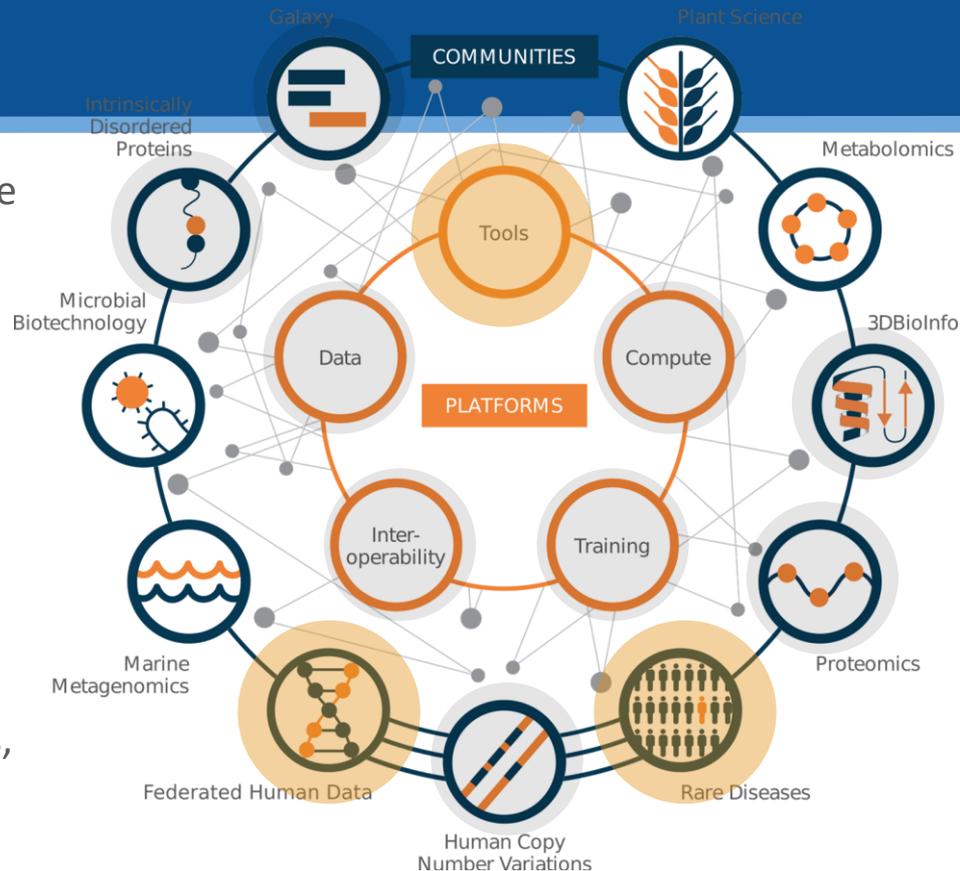


ELIXIR Structure

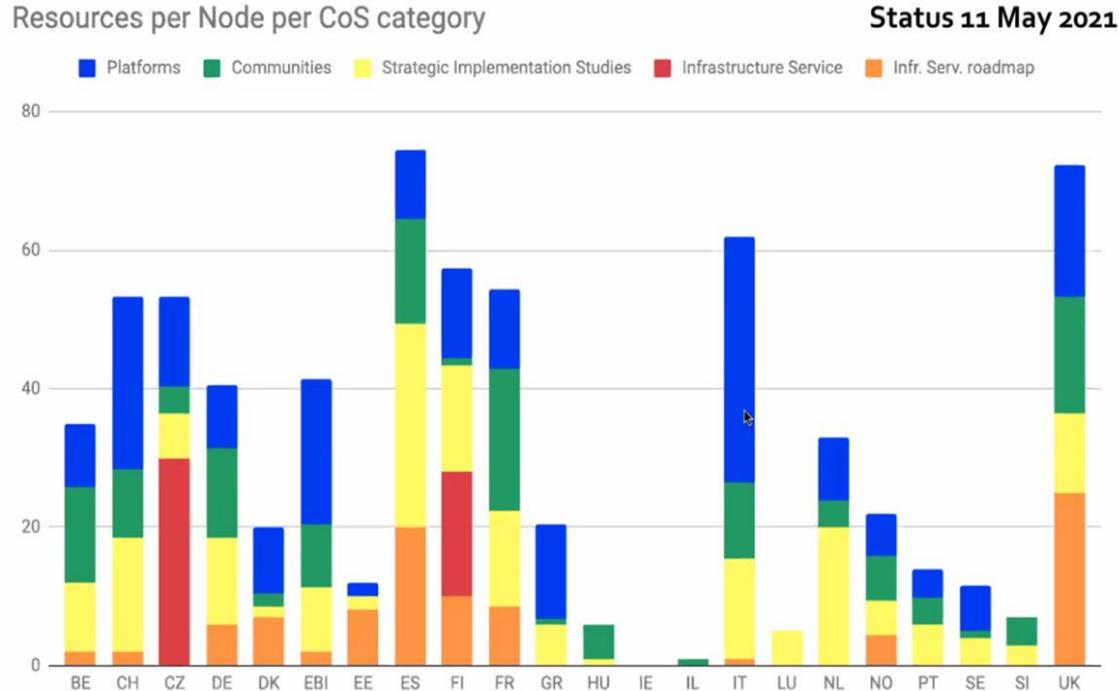
ELIXIR coordinates activities through at least one of the five 'areas of activity' called **Platforms**:

- Compute
- Data
- Interoperability
- Tools
- Training

These Platforms are driven by eleven ELIXIR **Communities** which develop standards, services, and training within their life science domains.



ELIXIR implementation studies



INB/ELIXIR-ES has a strong participation in **Human Genomics and Translational Data** projects & **Tools platform** projects.

- ELIXIR Beacon & Federated Human Data.
- Deploying Reproducible Containers and Workflows across Cloud Environments.
- Expanding the Galaxy: meeting (the needs of) ELIXIR Communities.
- ...

INB/ELIXIR-ES. Services portfolio to the broad community

ELIXIR Core Data Resource
and Deposition database



ELIXIR Recommended
Interoperability Resources



FAIRtracks

INB/ELIXIR-ES. Services portfolio to the broad community

**Strategic Resources
for ELIXIR**

**ELIXIR products supported
by ELIXIR-ES**

RD  Connect



Beacon



INB/ELIXIR-ES contribution towards enabling research using biomedical data



Imagen Clínica & Informática Médica @ IMPaCT-Data

IMAGEN CLÍNICA

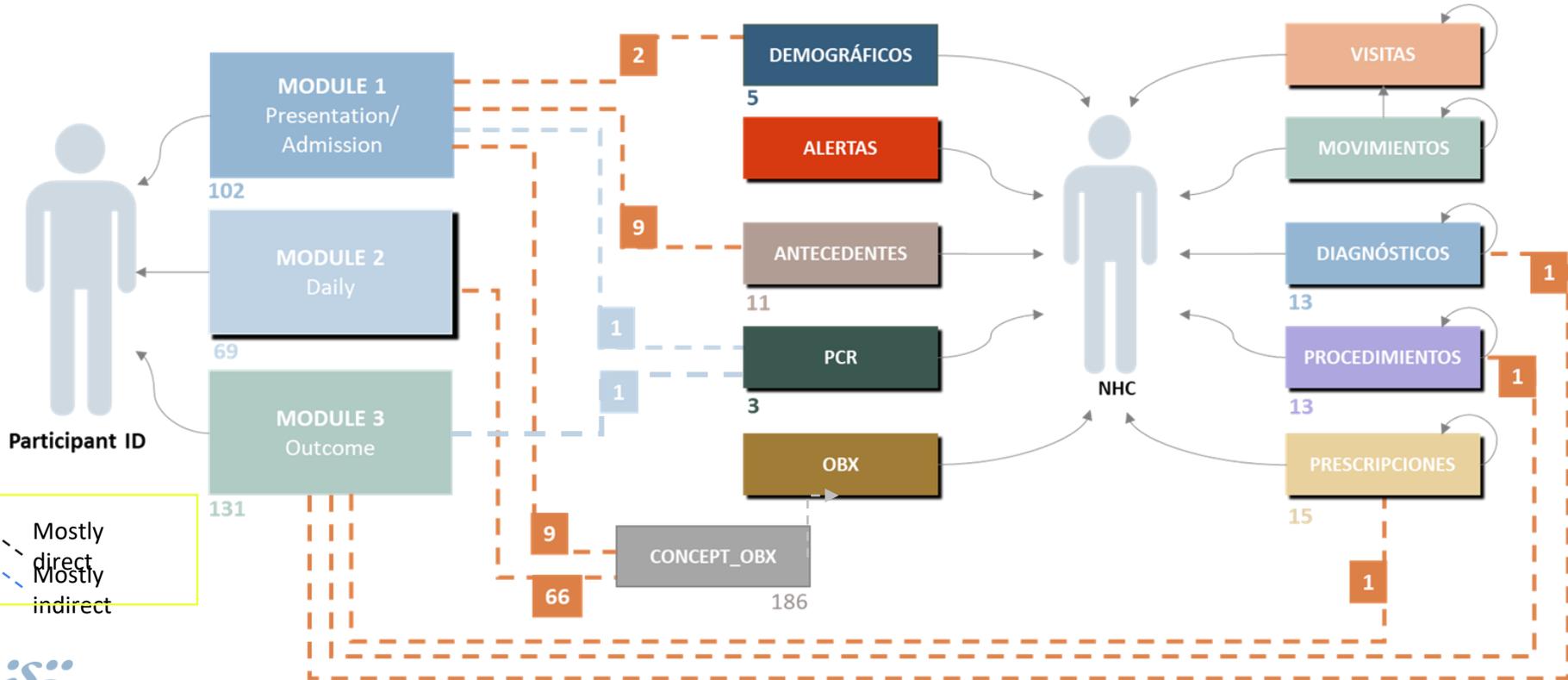
- FISABIO
- IIS La Fe - Imagen Clínica
- INIBICA

INFORMÁTICA MÉDICA

- H12O - Informática Médica
- HCB
- IACS
- IDIAPJGol
- IIS La Fe - Informática Médica
- Navarrabiomed - Informática Médica
- SAS-HUVR
- ISCIII - Informática Médica ¹

¹ Financiación intramural prevista del ISCIII

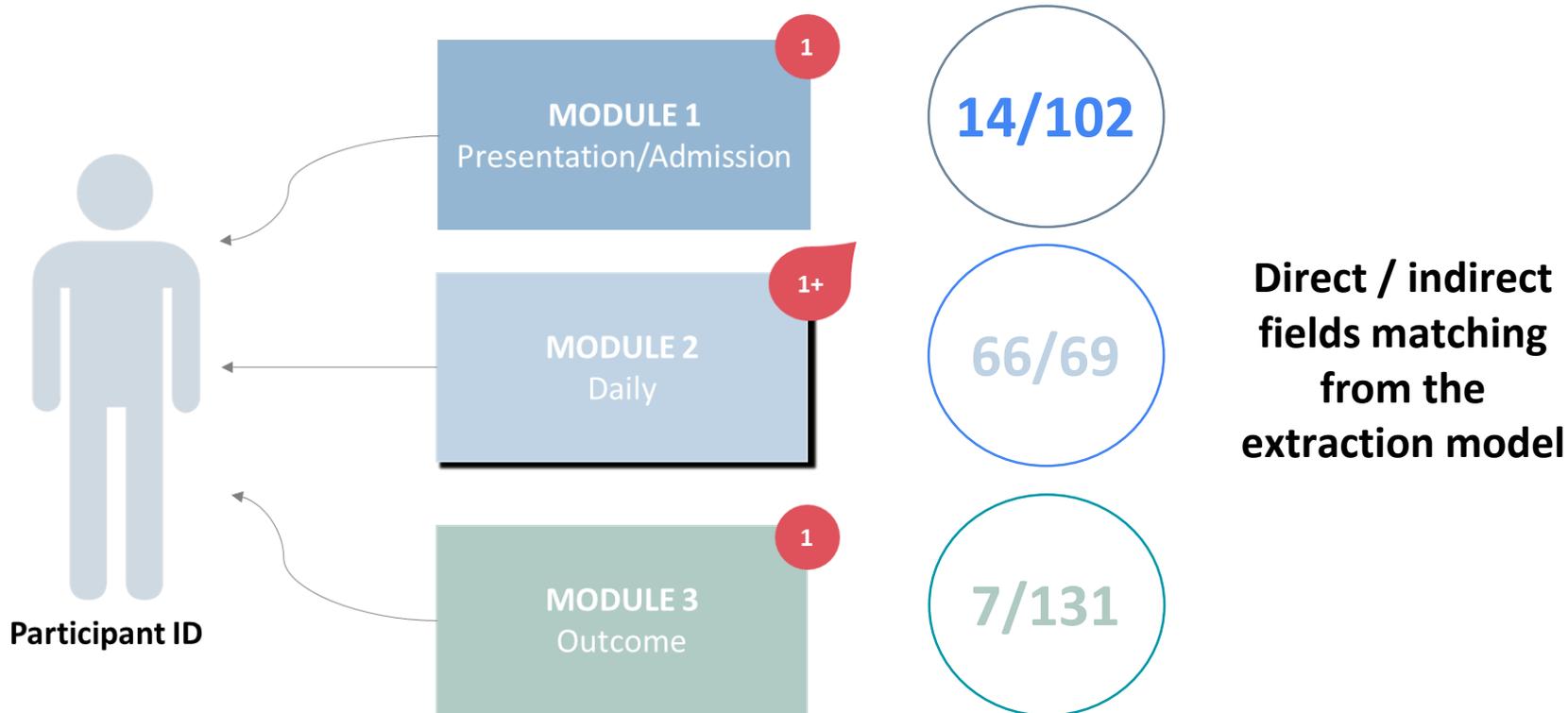
Correspondences



Mapping

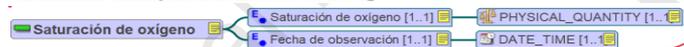


Hospital Universitario
12 de Octubre



A methodology EHR-derived datasets for COVID-19 research based on a Detailed Clinical Model

Functional data-specification and modeling



Standardized information model



Standardized EHR extract

```

<DATA_RECORD>
<CODE>103228002</CODE>
<DESCRIPTION>Saturación de oxígeno</DESCRIPTION>
<DATE>2020/05/16</DATE>
<TIME>07:31:29</TIME>
<VALUE>100</VALUE>
<UIDS>%</UIDS>
</DATA_RECORD>
<DATA_RECORD>
<CODE>103228002</CODE>
<DESCRIPTION>Saturación de oxígeno</DESCRIPTION>
<DATE>2020/05/15</DATE>
<TIME>07:28:18</TIME>
<VALUE>97</VALUE>
<UIDS>%</UIDS>
</DATA_RECORD>
<DATA_RECORD>
<CODE>103228002</CODE>
<DESCRIPTION>Saturación de oxígeno</DESCRIPTION>
<DATE>2020/05/14</DATE>
<TIME>07:37:34</TIME>
<VALUE>99</VALUE>
<UIDS>%</UIDS>
</DATA_RECORD>
    
```

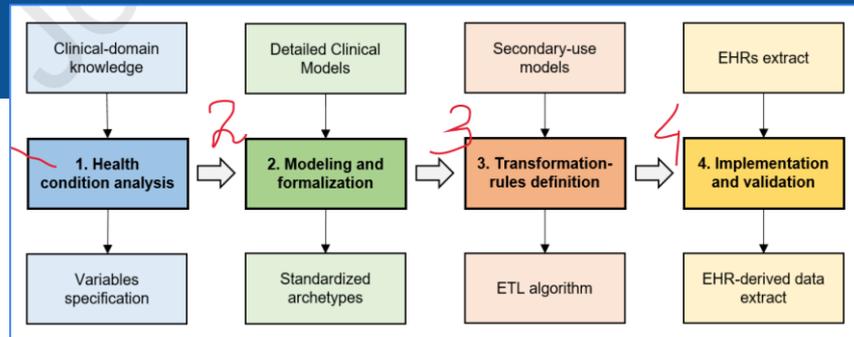
EHR-derived information model



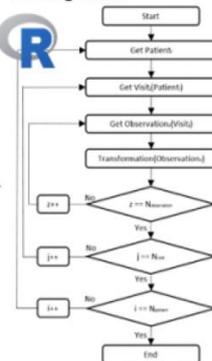
Regular data-transformation rules

```

start_date<-VISITS%>%filter(TYPE='AM')>%select(STARTDATE)
end_date<-VISITS%>%filter(TYPE='AM')>%select(ENDDATE)
sao2_min<-OBSERVATIONS%>%filter(CODE=='103228002')>%
filter(MRN==mrn, (DATE>start_date & DATE<end_date))>%
filter(VALUE==min(VALUE))
sao2_max<-OBSERVATIONS%>%filter(CODE=='103228002')>%
filter(MRN==mrn, (DATE>start_date & DATE<end_date))>%
filter(VALUE==max(VALUE))
    
```



ETL algorithm



EHR-derived data extract

day_max	day_min	day_max	day_min	spci_max	spci_min	glucose	hemoglobin	hematocrit	white_platelets	platelet_gi	pt	alt	albumin	ast
140	140	95	95	100	100		13.94	3.7	141	29.124	1.08	24.0	7	23
140	132	85	75	99	98	12.9	37.5	9.7	179	19.122	1.06	23.0	6	24
124	114	82	86	96	97	13.4	33.9	5.7	184	30.113	1.02	40.0	6	24
132	119	76	86	96	97	13.2	32.4	6	237	29.122	1.06	57.0	5	41
117	109	87	89	96	96	18.8	31.1	3.7	338	31.124	1.08	90.0	9	39
106	104	75	75	97	97	13.1	32.8	4	229	30.113	1.02	93.0	2	4
110	110	89	89	100	99	13.3	32.9	4.6	225	30.124	1.08	39.0	2	17
127	123	76	75	97	96	13.4	32.8	2.9	176	28.0	1.02	29	2	29
130	131	84	83	96	97	13.1	32.2	2.7	194	30.117	1.02	28.0	2	29
107	89	86	89	99	97	13.4	34.4	2.9	204	27.0	1.02	29	2	29
114	114	80	80	97	97	13.6	33.9	3.5	228	30.0	1.02	29	2	29
112	112	89	89	99	96	12.9	33.9	4.7	276	27.0	1.02	32	2	29
111	88	76	85	100	96	13.1	35.6	6	285	30.0	1.02	29	2	29
119	113	75	76	100	99	13.8	34.3	3.6	297	32.0	1.02	29	2	29
130	94	82	88	100	96	13.9	34.4	4.3	187	29.124	1.08	27.0	2	24
127	123	76	75	97	96	13.4	32.8	2.9	176	28.0	1.02	29	2	29

Obtaining EHR-derived datasets for COVID-19 research within a short time: a flexible methodology based on Detailed Clinical Models

Miguel Pedrera Jiménez, Noelia García Barrio, Jaime Cruz Rojo, Ana Isabel Terriza Torres, Elena Ana López Jiménez, Fernando Calvo Boyero, María Jesús Jiménez Cerezo, Alvar Javier Blanco Martínez, Gustavo Roig Domínguez, Juan Luis Cruz Bermúdez, José Luis Bernal Sobrino, Pablo Serrano Balazote, Adolfo Muñoz Carrero

PII: S1532-0464(21)00026-5
 DOI: <https://doi.org/10.1016/j.jbi.2021.103697>
 Reference: YJBIN 103697

To appear in: *Journal of Biomedical Informatics*

Extracción de información no estructurada PLN



A

Medical document selection and automatic annotation

Paciente con fiebre ingresa en neumología
Patient with fever is admitted to the pneumology

Unannotated Medical Documents



Artificial Intelligence baseline systems

Paciente con fiebre ingresa en neumología
Patient with fever is admitted to the pneumology ward

Automatically Annotated Documents



B

Medical Experts correct the automatic annotations

Automatically Annotated Documents



Multilingual Terminologies

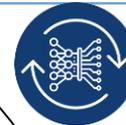


Medical Experts

SCTID: 386661006
Paciente con fiebre ingresa en neumología
Patient with fever is admitted to the pneumology

"tag a little, train a little"

Iterative correction of automatic annotations until system is perfected powered by HPC resources



use human knowledge



to learn automatically

C

Automatic Annotation of new medical documents

El paciente con disnea fue colocado en decúbito prono.
The patient with dyspnea was placed in prone position.

Unannotated Medical Document



Artificial Intelligence Systems powered by HPC resources



SCTID: 2670360073
El paciente con disnea fue colocado en decúbito prono.
The patient with dyspnea was placed in prone position

Automatically Annotated Text [symptom / disease / medical procedure / medication + code]



Record #01		
Affirmed	Negated	Uncertain
fever	dyspnea	COVID-19

Structured medical records

Use Cases



Detecting similar patients



Identifying successful treatments



Doc. indexing - Power complex searches

Grupo de anotadores	Fármacos	Procedimientos	Patógenos	Síntomas	Enfermedades
Anotadores	2	2	2	2	1
Especialidades	Neurología, farmacia	Anestesia, documentación clínica	Microbiología, epidemiología	Pediatría, neuroradiología	Documentalista
Documentos anotados	85	40	37	69	42
Anotaciones manuales	2127	1340	789	2010	728

Modelo	Fármacos	Procedimientos	Patógenos*	Síntomas	Enfermedades
Iteración V0 (Precision, Exhaustividad y Valor F)	92,4 73,9 82,1	84,3 58,7 69,2	93,9 92,7 93,3	36,2 45,5 40,3	66,5 47,9 55,7
Iteración V1 (Precision, Exhaustividad y Valor F)	-	75,6 66,7 70,9	-		-

Investigación

Centros e Inst. Investigación

- GENyO
- IFISC
- INIBIC
- IRB Lleida
- ITER

Universidades

- U.Lleida
- U.Politécnica Catalunya
- U.Politécnica Madrid
- U.Salamanca
- U.Valladolid
- U.Zaragoza

Demostradores Tecnológicos (Empresas)

- BITAC
- Fujitsu
- GMV
- IBM
- INTEL
- IDIS
- *llumina*
- *Procter&Gamble*
- *Roche Pharma*

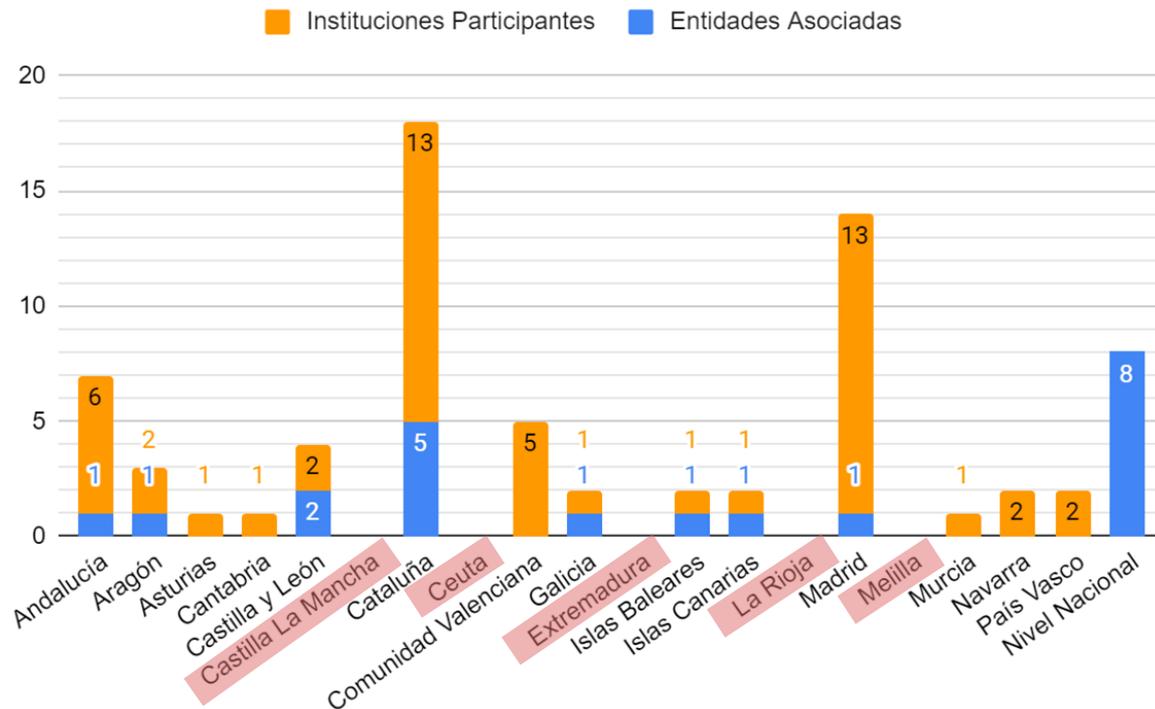
Otras Entidades Relevantes

Agencias Evaluadoras

- **AEMPS** - Agencia Española del Medicamento y Productos Sanitarios
- **RedETS** - Red Española de Agencias de Evaluación de Tecnologías
- **CTI-SEIS** - Comité Técnico de Interoperabilidad de la Sociedad Española de Informática de la Salud
- **AQuAS** - Agència de Qualitat i Avaluació Sanitàries de Catalunya

Distribución en el Territorio

Representación Comunidades Autónomas

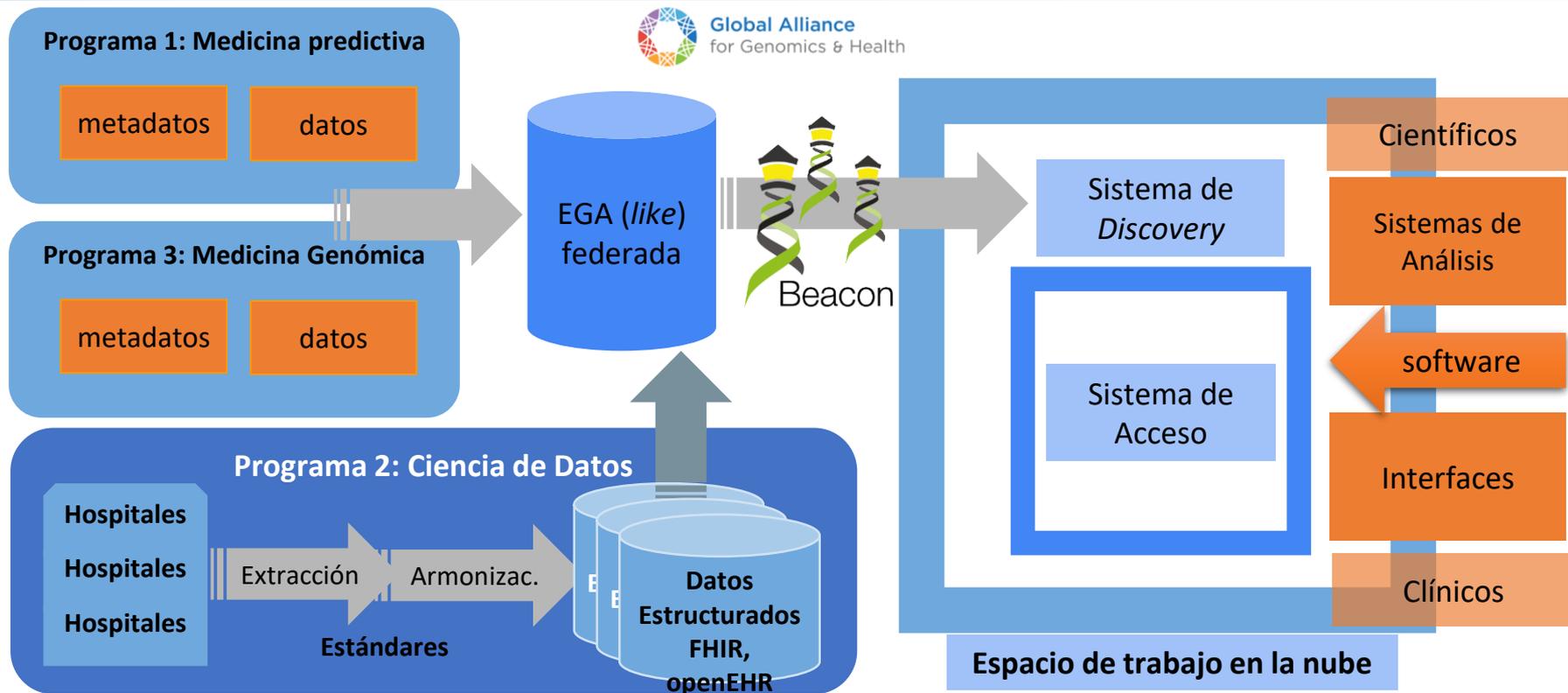


Presentaciones

Entidad	Reps
BSC	Alfonso Valencia, Salvador Capella-Gutierrez, Eva Alloza, Lidia López
CNIO	Fátima Al-Shahrour - WP1
CRG-EGA	Arcadi Navarro - WP2
FPS	Joaquín Dopazo - WP3
IACS	Sandra García-Armesto - WP4
SAS-HUVR	Carlos Parra - WP5

Plan de Trabajo

IMPACT-Data: Esquema



WP1 - Coordinación

BSC-CNS (Alfonso Valencia & Salvador Capella-Gutiérrez)

CNIO (Fatima Al-Shahrour)

WP1 - Coordinación

Visión: Facilitar el trabajo interno de IMPaCT-Data y contribuir a asegurar el alineamiento e impacto de sus actividades más allá del propio programa.

- **Objetivo(s) vinculados a IMPaCT-Data**

- Coordinación para facilitar se consigan con éxito todos los objetivos del proyecto (**O1 ... O9**)
- Coordinación tanto con los actores internos como externos del proyecto
- Coordinación de la difusión y comunicación de los resultados del proyecto
- **Proveer de recursos formativos a las diferentes instituciones**

- **Resultados**

- Garantizar el buen funcionamiento interno
- Establecer y mantener relaciones con los otros dos programas de IMPaCT así como otros agentes relevantes, p. ej. industria, universidades, representantes de sistemas sanitarios, etc.
- Promover la difusión y comunicación de los resultados de IMPaCT-Data
- Gestión de las Líneas Estratégicas Transversales en el contexto del Programa de Ciencia de Datos

WP1 - Visión inicial Tareas

- **T1.1.** Coordinación con la Comisión de Dirección y Seguimiento de IMPaCT y los otros dos programas de IMPaCT
- **T1.2.** Coordinación con los servicios de salud de las CCAA
- **T1.3.** Coordinación y gestión interna del programa
- **T1.4.** Comunicación y difusión de las actividades del programa
- **T1.5.** Coordinar la participación de las entidades asociadas en las actividades del programa
 - T1.1 ... T1.5: Coordinación tanto del correcto funcionamiento de programa de Ciencia de Datos de IMPaCT como de todas las colaboraciones necesarias dentro y fuera de IMPaCT
- **T1.6. Formación Especializada (CNIO)**
 - Diseño y desarrollo de programas formativos para investigadores y profesionales del SNS con el objetivo de capacitarlos para acceder y usar las herramientas e infraestructura IMPaCT-Data con fin asistencial e investigación.

WP1 - Primeros Entregables (M12, M24 y M36)

- **E1.0.** Indicadores de IMPaCT-Data (T1.3, M12)
 - Informe de indicadores de seguimiento del proyecto (E1.0).
- **E1.1.** Coordinación Seguimiento Proyecto (T1.1, M12)
- **E1.2.** Coordinación con Servicios de Salud (T1.2, M12)
- **E1.3.** Coordinación Interna (T1.3, M12)
- **E1.5.** Coordinación con Entidades Asociadas (T1.5, M12)
 - Resumen anual de todas las acciones de coordinación
- **E1.4.** Informe de Comunicación y Difusión (T1.4, M12)
 - Resumen anual de las actividades de comunicación y difusión de las diferentes áreas del proyecto
- **E1.6.** Informe de Actividades Formativas (T1.6, M12)
 - Diseño de un catálogo de cursos y programas docentes que conecten los 3 ejes estratégicos IMPACT (medicina predictiva, ciencia de datos, medicina genómica) y que cubran las necesidades docentes del SNS (ej: cursos acreditados para residentes).
 - Promover la creación de unidades docentes acreditadas en las instituciones colaboradoras. Fomentar la participación de los grupos TransBioNet como nodos de enlace intrahospitalarios.

WP2 - Infraestructura Gestión Datos

BSC-CNS (Alfonso Valencia & Salvador Capella-Gutiérrez)

WP2 - Infraestructura Gestión Datos

Visión: Diseñar la arquitectura de un **sistema computacional federado** que combine computación en la nube con computación de altas prestaciones. Realizar una primera implementación de este sistema durante este ciclo de IMPaCT-Data y servir de referencia para otras implementaciones.

- **Objetivo(s) vinculados a IMPaCT-Data**

- **O1.** Desarrollo de la primera iteración de una **plataforma de datos federada** que comprenda datos médicos, genómicos y de fenotipo, normalizada y multipropósito.
- **O2.** Desarrollo e implementación de una primera versión de una **infraestructura de computación en la nube** para dar soporte inicialmente a los desarrollos de IMPaCT.
- **O7.** Liderar el portafolio de **recursos bioinformáticos ofrecidos por España a ELIXIR**, incluyendo software, bases de datos y plataformas enfocados al ámbito de salud humana, y organizados en consonancia con el objetivo general de IMPaCT-Data.

- **Resultados**

- **R1.** Arquitectura de referencia para facilitar la interoperabilidad a nivel computacional.
- **R2.** Establecer las guías y buenas prácticas para la generación de contenedores software y su uso como parte de flujos de trabajo para contribuir a la diseminación de software y la reproducibilidad de resultados.

EOSC-Life: an open collaborative space for digital biology in Europe



- Publish FAIR life science data resources in EOSC
- Populate an ecosystem of innovative life science tools in EOSC
- Enable data-driven research in Europe by connecting life scientists to EOSC via open calls for participation

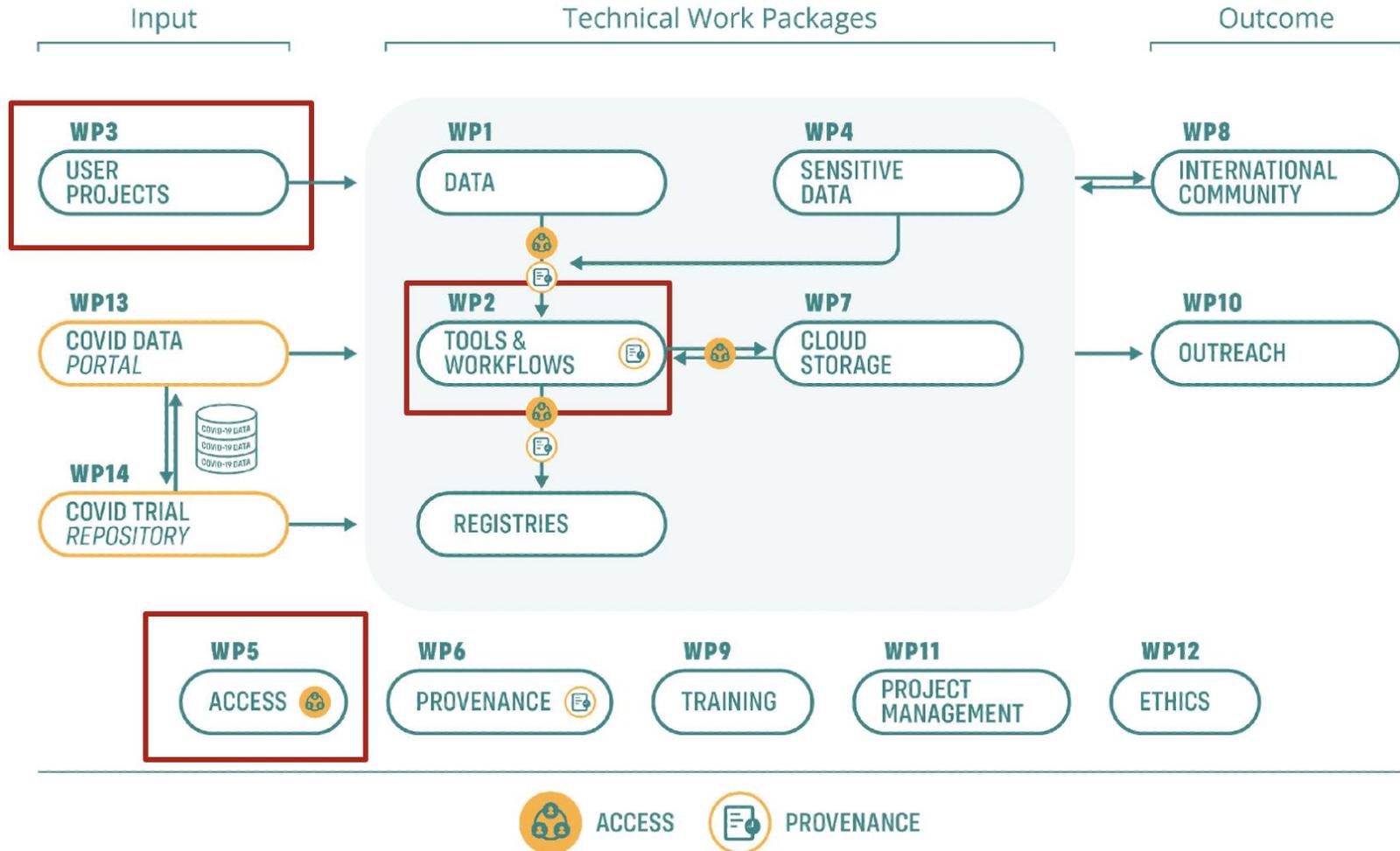
13 ESFRI Health and
Food Research Infrastructures
47 Partners and 17 linked 3rd parties

EOSC Association
Advancing Open Science in Europe

nática Médica (21/06/2021)



EOSC-Life: Instruct is co-leading 3 of the 14 work packages



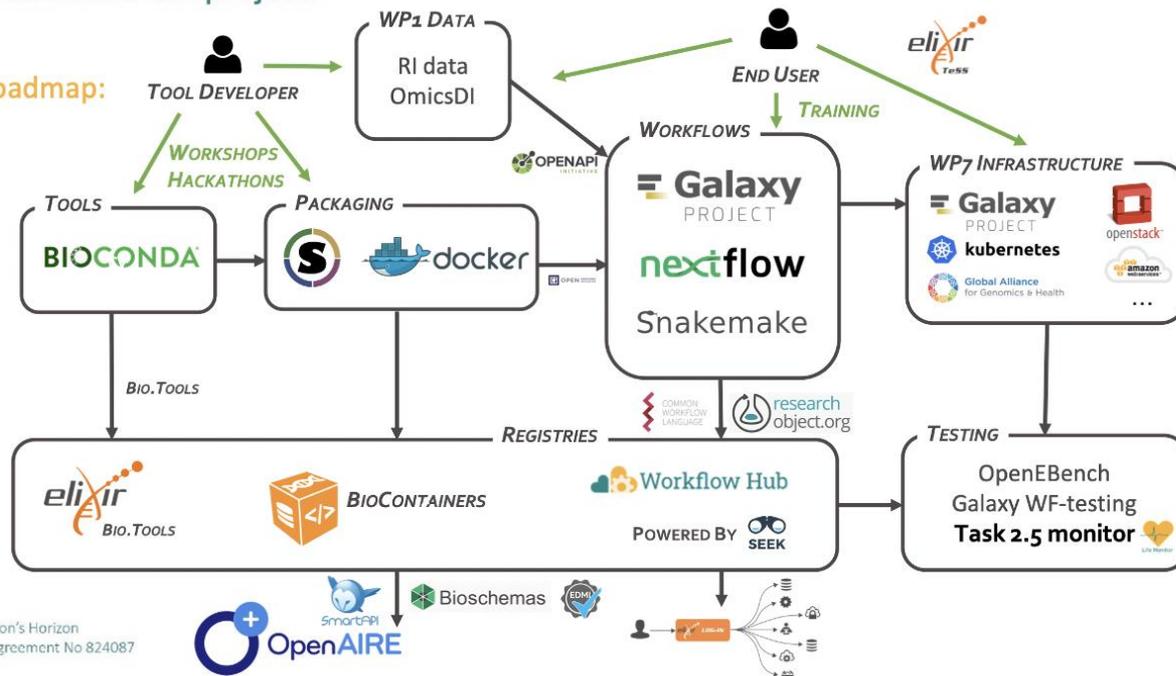
EOSC-life WP2: tools and workflows



WP2 Co-lead: Antonio Rosato (Instruct-ERIC – CIRMMMP)

- Development of an environment enabling the cloud deployment of workflows for the analysis and integration of life science data.
- Support EOSC-Life pilot projects on the cloud deployment of workflows using tools, platforms and standards supported in the EOSC-life project.

WP2 Tools Collaboratory roadmap:



WP2 - Visión inicial Tareas

- **T2.1:** Implementación inicial y evaluación de un **sistema de computación para datos biomédicos** asociados a IMPaCT.
 - Evaluar los distintos **entornos computacionales existentes** en Europa, e.g. de.NBI cloud - GAIA-X.
 - Proponer la arquitectura inicial del **sistema computacional en la Nube** en biomedicina incluidos las necesidades de hardware y de recursos humanos.
 - Proponer una arquitectura computacional que combine el sistema en la nube con computación de altas prestaciones.
- **T2.2:** Adopción e implantación de protocolos para el uso de **software containers y flujo de trabajos** siguiendo las recomendaciones de ELIXIR, EOSC-life y GA4GH
 - Identificar las guías de buenas prácticas para la generación de contenedores software y flujo de trabajo.
 - Evaluar la implantación de distintas **alternativas de contenedores software** y gestores de flujo de trabajo.
 - Identificar las necesidades de **formación específica para facilitar la adopción** de estas guías de buenas prácticas.

WP2 - Primeros Entregables

- **E2.1. Evaluación de Interfaces Disponibles (T2.1, M12)**
 - Establecer qué interfaces OA son necesarias para IMPaCT-Data basado en otros sistemas existentes.
 - Establecer las necesidades iniciales de los casos de uso (WP6) junto con los otros programas de IMPaCT.
 - Estudiar las posibles alternativas para proporcionar acceso a la estructura computacional de IMPaCT-Data.
- E2.2. Diseño Inicial de la Arquitectura de la Infraestructura (T2.1, M24)
- E2.3. Implementación de la Infraestructura (T2.1, M36)
- E2.4. Guía de Buenas Prácticas para el Desarrollo y Mantenimiento de Software (T2.2, M18)
- E2.5. Informe Indicadores de Software (T2.2, M36)

WP3 - Datos Genómicos

CRG-EGA (Arcadi Navarro & Jordi Rambla)

WP3 - Datos Genómicos

Visión: Facilitar la gestión y posterior reutilización (segura, ética, fácil y efectiva) de los datos genómicos (en su contexto).

- **Objetivos directos**

- **O1.** Desarrollo de la primera iteración de una **plataforma de datos federada** que comprenda datos médicos, **genómicos y de fenotipo**, normalizada y multipropósito.
- **O6.** Incorporación de información genética y genómica, **tratamiento** de la información para su consulta rápida, **encriptación** para mantener la confidencialidad, e **indexación** y **agregación** de datos.
- **O7.** Liderar el **portafolio de recursos** bioinformáticos ofrecidos por España a ELIXIR, incluyendo software, bases de datos y plataformas enfocados al ámbito de salud humana, y organizados en consonancia con el objetivo general de IMPaCT-Data.>> *(Local EGA + Analysis System + GA4GH Beacons + IMPaCT Beacon Network)*

- **Vínculos con otros objetivos**

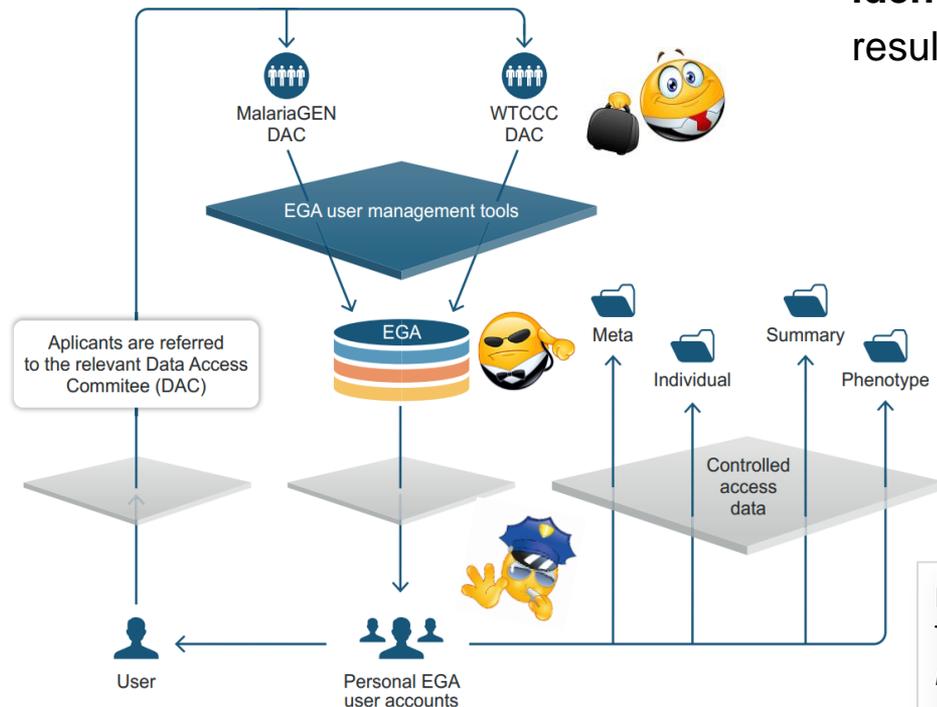
Conexión con **O2** (Infraestructura de computación en la nube), Conexión con **O3** (Análisis de datos y FAIRificación), Conexión con **O4** (Calidad de los datos), Conexión con **O9** (Prototipos de integración de resultados del análisis genómico, de imagen y HCE), Conexión con **LET02** (Internacionalización)

- **Resultados**

- R1 - Contribución a la gestión de los datos genómicos >> *(Local EGA + Analysis System)*
- R2 - Entorno de consulta de datos disponibles >> *GA4GH Beacons + IMPaCT Beacon Network*
- R3 - Integración internacional >> *Federated EGA + GA4GH Beacons + ELIXIR Beacon Network*

The European Genome-phenome Archive

The EGA is a resource for permanent **secure archiving** and sharing of all types of **potentially identifiable** biomolecular and phenotypic data resulting from biomedical research projects.



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 healthcare institutions.

Project goal:

To transform the EGA to a joint project (*in the context of ELIXIR Europe*) to have a real impact in the development of **personalized medicine**

The EGA contains a

- Studies > 2,700
- Datasets > 6,250
- DACs > 1,200
- Submitters > 1,500
- Requesters ~19,000
- Files > 3,300,000
- Archive size > 13.5 Petabytes

Updated February 26th



EGA SUCCESS:

Federation: *Discover globally, manage locally*

One Central EGA, many Federated Nodes.

To go into production in 2021 with:

Finland,
Germany,
Norway,
Spain,
Sweden

... and then the 2nd, 3rd wave

Federation of human data

- Many national datasets from human research participants needs to be stored locally
- ELIXIR developing a federation with shared metadata (FAIR) and local data store (secure)
- Linking local EGA to national clouds – and international access (ELIXIR-AAI)



1+ Million Genomes



DIGITAL DAY
APRIL 30, 2021

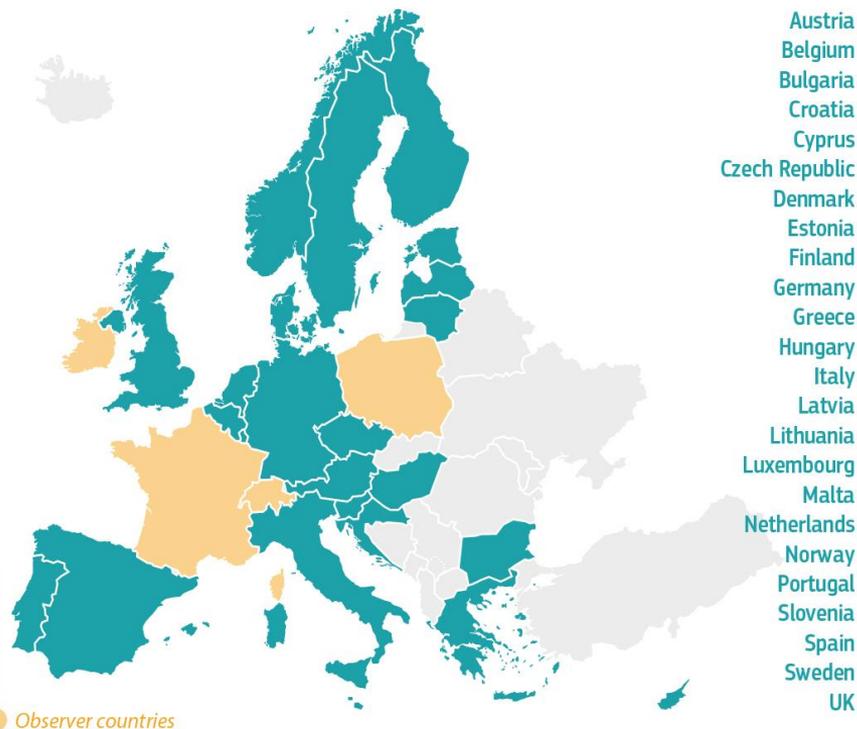
DECLARATION OF COOPERATION

Towards access to at least 1 million sequenced genomes in the European Union by 2022

Mariya Gabriel, Commissioner for Digital Economy and Society welcomed this new step forward:

"EU cooperation on linking genomic health data becomes ever more important to advance personalised health care, better prevention and excellence in research. We are happy to see Latvia joining the other 18 signatory Member States and look forward to enhance this European collaboration. "

Countries that have signed the 1+MG Declaration since 2018



updated on 2 February 2021



nática Médica (21/06/2021)



ELIXIR and 1+MG > Beyond 1MG

Discover globally,
manage locally

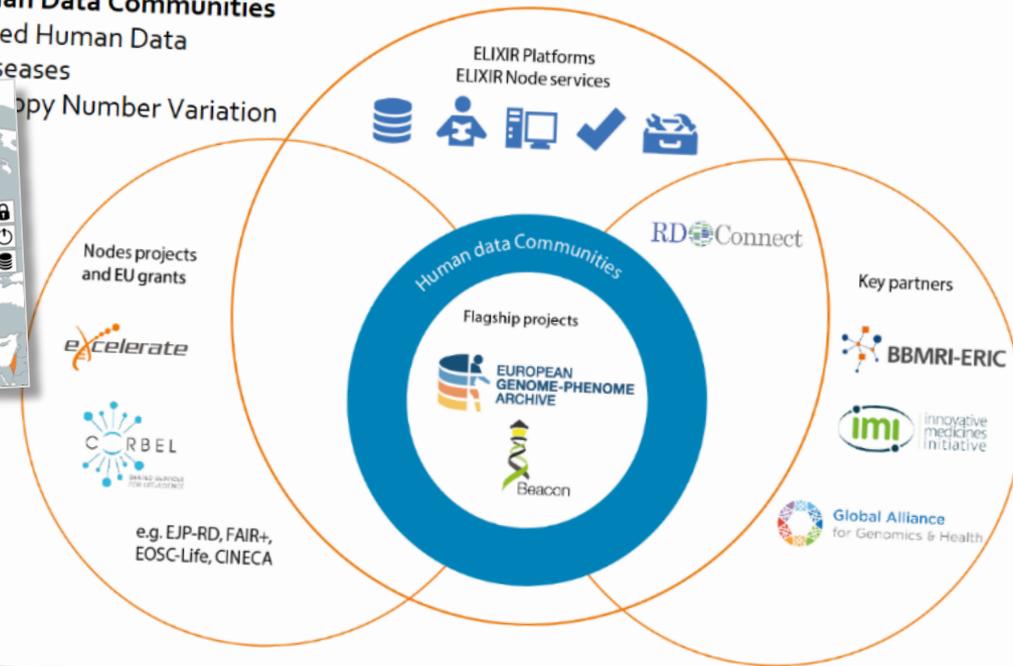
Partnerships and community formation

ELIXIR Human Data Communities

- Federated Human Data
- Rare Diseases

Federation of human data

- Many national datasets from human research participants needs to be stored locally
- ELIXIR developing a federation with shared metadata (FAIR) and local data store (secure)
- Linking local EGA to national clouds – and international access (ELIXIR-AAI)



What is a Beacon?

Beacon Test Instance Datasets Filtering Terms Login

Subject ID	Datasets IDs	Taxon	Sex	Ethnicity	Geographic Origin	SRA Family ID	Race	Date Of Birth	Weight (kg)	Height (cm)	E T
NA24631	GiaB	Homo sapiens	male	Han Chinese	United States of America	3150	asian	1979-08-07	57	173	A
NA24694	GiaB	Homo sapiens	male	Han Chinese	United States of America	3150	asian	1949-03-19	64	167	A
NA24695	GiaB	Homo sapiens	female	Han Chinese	United States of America	3150	asian	1949-07-08	48	160	C

Search all

Search by ID

○ Show individual
○ Show samples from individual
○ Show variants from individual

Filters

Variant **basic** advanced

Assembly **+**

chr : pos ref > alt

Datasets

None selected -

Ontologies

Search

Individual NA24631

Phenotypic Features

phenotypicid	severity	onsetType	ageOfOnset	dateOfOnset
NCIT:C3001	-	-	-	-

Diseases

diseaseid	stage	severity	onsetType	ageOfOnset	dateOfOnset	familyHistory
HP:0000545	-	-	-	-	-	-
HP:0004789	-	-	-	-	-	-

Medications

medid	start	end
-------	-------	-----

Measures

measureid	modifier	date	ageAtMeasure
-----------	----------	------	--------------

Exposures

exposureid	modifier	date	ageAtExposure
------------	----------	------	---------------

Degrees

degreeid	numSubjects	pedigreeRole	affectedStatus	pedigreeDisease
id1	3	NCIT:C64435	-	-

Handovers

note	handoverType	url
Personal Genome Project Public Profile hu91BD69	CUSTOM.pgp_public_profile	https://my.pgp-hms.org/profile/hu91BD69
GET-Evidence variant report for subject hu91BD69	CUSTOM.pgp_variant_report	https://evidence.pgp-hms.org/genomes?display_genome_id=4832c699ec9b961e5cf20cddd73aa66fd19a

Cohorts

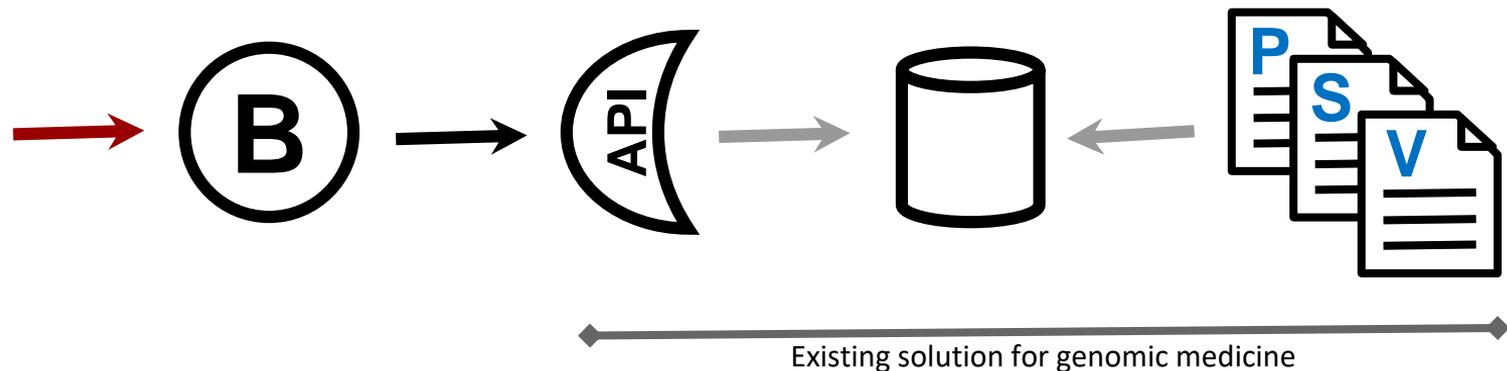
Families

Individuals

Biological samples

Mutations

What is a Beacon?



Fundación Progreso y Salud
CONSEJERÍA DE SALUD



Beacon Network: Aggregator of Beacons



Discover globally, manage locally

Catalan hospitals organizing a Beacon Network

*Discover globally,
manage locally*

CLÍNIC
BARCELONA
Hospital Universitari


Vall d'Hebron
Hospital


ICO
Institut Català d'Oncologia

Hospital
del Mar

Parc
de Salut
MAR
Barcelona


HOSPITAL DE LA
SANTA CREU I
SANT PAU
UNIVERSITAT AUTÒNOMA DE BARCELONA


SJD **Sant Joan de Déu**
Barcelona · Hospital


Parc Taulí
Hospital Universitari


Fundació Puigvert



WP3 - Visión inicial Tareas

- **T3.1:** Implementación de sistemas de almacenamiento de información genómica primaria mediante instancias de local EGAs
 - Facilitar la gestión de datos genómicos (internamente)
 - ...integrados en IMPaCT-Data (con imágenes, HCE, infraestructura...)
 - ...e integrados en otras redes (EGA, ELIXIR,...)
- **T3.2:** Adaptación, instalación y uso de software de código abierto para el análisis e integración de distintas fuentes de datos
 - ¿Qué necesita el entorno sanitario/hospitalario en el proceso de diagnóstico genómico?
 - ¿Cómo varía en función de los recursos humanos y técnicos disponibles en cada centro?

WP3 - Primeros Entregables

- **E3.1.** Requisitos de un nodo integrado en EGA** (T3.1, M12)
 - Requisitos técnicos, humanos y acuerdos legales para la puesta en marcha de un nodo EGA-like
- **E3.2.** Descripción de los interfaces de las instancias EGA Platform** (T3.1, M18)
- **E3.4.** Análisis genómico en entornos sanitarios (T3.2, M18)
 - Informe con las adaptaciones necesarias para el potencial uso de los sistemas de análisis por parte del personal sanitario

- E3.3. Informe final de los requisitos y contexto de un nodo EGA** (T3.1, M36)
- E3.5. Comparativa de instalación en entornos hospitalarios** (T3.2, M36)

* (vinculado a EGA-Community)

** (editados respecto al Plan Estratégico): Requisitos técnicos, humanos y acuerdos legales para la puesta en marcha de un nodo de local EGA

WP4 - Datos Médicos e Imagen

IACS (Sandra García-Armesto)

WP4 - Datos Médicos e Imagen

Objetivo(s) vinculados a IMPaCT-Data

- **01.** Desarrollo de la primera iteración de una **plataforma de datos federada que comprenda datos médicos, genómicos y de fenotipo, normalizada y multipropósito.**
- **03.** Desarrollo e implementación de protocolos, métodos y sistemas integrados de análisis de datos basados en la infraestructura de bases de datos, métodos, sistemas de evaluación y mecanismos de **FAIRificación.**
- **04.** Desarrollo inicial para monitorizar los procesos de **tratamiento y evaluación de la calidad de los datos durante los procesos de transformación y manejo de los mismos.**
- **05.** Desarrollo inicial de sistemas para la **extracción (semi-)automática y segura - con garantía de confidencialidad - de información mínima normalizada y seudonimizada** a partir de los sistemas de información sanitarios, incluyendo la Historia Clínica Electrónica (HCE), con el desarrollo de protocolos de interoperabilidad y normalización basados en estándares internacionales.
- **08. Extracción de información cuantitativa de imágenes médicas,** tanto basadas en características de la imagen (radiómica) como las que están relacionadas con fenómenos patofisiológicos, para su uso e integración con otras fuentes de datos.

● Resultados

- **Guía para la extracción de dato clínico e imagen médica** desde lo sistemas de información clínica de los sistemas de salud
- **ETLs de extracción** de dato clínico e imagen médica en socios de referencia

WP4 - Visión inicial Tareas

- **T4.1:** Adaptación, instalación y uso de **software de código abierto para la extracción de variables a partir del HCE**

Despliegue de soluciones basadas en código abierto para la extracción y tratamiento de dato clínico estructurado a partir de sistemas de información clínicos

- Mapeo de los sistemas de información médica existentes en el SNS y selección de los estándares para la interoperabilidad recomendados
- Propuestas de **diseño de los procesos de extracción, transformación y carga (ETL)**

- **T4.2:** Adaptación, instalación y uso de herramientas de software de código abierto para el seguimiento y validación de los procesos de extracción y tratamiento de los datos de imagen médica

Despliegue de soluciones basadas en código abierto para **la extracción y tratamiento de imagen médica desde los sistemas de información clínicos**

- Mapeo de los sistemas de información médica existentes en el SNS y selección de los estándares para la interoperabilidad recomendados
- Propuestas de diseño de de los **procesos de extracción, transformación y carga (ETL)**

WP4 - Primeros Entregables

- **E4.1. Normas Internacionales de Anotación de Información de HCE (T4.1, M15)**
Estándares internacionales de codificación y almacenamiento de dato clínico para su explotación secundaria
 - Evaluación de los estándares, terminologías y ontologías de dato sanitario con la perspectiva de un uso secundario del mismo
 - Mapeo de los **formatos, estándares y esquemas de codificación en uso en el SNS**
- **E4.4. Normas Internacionales de Anotación de Información de Imagen Médica (T2.2, M15)**
Estándares internacionales de codificación y almacenamiento de imagen médica para su explotación secundaria
 - Evaluación de los formatos de almacenamiento y anotación de imagen médica normalizados internacionalmente
 - Mapeo de los formatos almacenamiento y anotación de imagen médica en uso en el SNS
- **E4.2. Comparación de Técnicas de Gestión de Información de HCE (T2.1, M21)**
- **E4.3. Pruebas de Concepto de Extracción de Información de HCE (T2.1, M36)**
- **E4.5. Comparación de Técnicas de Gestión de Información de Imagen Médica (T2.2, M21)**
- **E4.6. Pruebas de Concepto de Extracción de Información de Imagen Médica (T2.2, M36)**

WP5 - Integración Datos Biomédicos

FPS (Joaquín Dopazo)

WP5 - Integración Datos

Visión: Exploración de la forma óptima en la que se **pueden integrar datos genómicos, de imagen médica y clínicos para su uso secundario**. Federación versus centralización. **Evaluación de problemas técnicos y barreras legales**. Propuesta de soluciones.

- **Objetivo(s) vinculados a IMPaCT-Data**

- **O1.** Desarrollo de la primera iteración de una **plataforma de datos federada que comprenda datos médicos, genómicos y de fenotipo, normalizada y multipropósito**.
- **O6.** Incorporación de información genética y genómica, tratamiento de la **información para su consulta rápida, encriptación para mantener la confidencialidad e indexación y agregación de datos**.
- **O8.** Extracción de información cuantitativa de imágenes médicas, tanto basadas en características de la imagen (radiómica) como las que están relacionadas con fenómenos patofisiológicos, para su uso e integración con otras fuentes de datos.
- **O9.** **Desarrollo de prototipos de integración de los resultados del análisis genómico y de imagen con la información normalizada extraída de las HCE**, en los formatos adecuados para el análisis humano y computacional.

- **Resultados**

- R1. Informe sobre la integración de datos y pros y contras de los distintos modelos
- R2. Demostrador de Beacons en centros escogidos

WP5 - Visión inicial Tareas

- **T5.1: Integración de datos genómicos y radiómicos con datos médicos estructurados para su uso secundario**
 - Comparación de modelos de **integración centralizados o federados** en relación al uso secundario de los datos
 - Posibilidad de aprovechar estructuras de redes existentes que comparten datos (ej. CIBERER)
- **T5.2: Implantación de mecanismos para la identificación automática de información genómica relevante, o contenido de las imágenes médicas que den respuestas anonimizadas al estilo Beacon**
 - **Demostrador de catálogo** de información genómica relevante para i) diagnóstico en X enfermedades raras, ii) recomendación de tratamiento en X cánceres, iii) patógenos (COVID) y árboles de decisión incluyendo el manejo de VUSes
 - Evaluación de los **mecanismos de priorización de anotaciones** de imagen médica y propuesta de automatización

WP5 - Primeros Entregables

- **E5.4.** Requisitos Técnicos para Puesta en **Marcha de Sistemas Beacon** (T5.2, M12)
 - Informe sobre requisitos técnicos y equipamiento mínimo necesario para la puesta en marcha de un sistema Beacon
 - Informe sobre soluciones en redes cooperativas existentes (ej. CIBERER)
- **E5.1. Técnicas de Integración de Datos Biomédicos¹** (T5.1, M18)
 - Propuestas de estructuración de datos genómicos
 - Propuestas de estructuración de datos de imagen médica
 - Informe sobre su combinación con datos clínicos estructurados para hacerlos consultables
- **E5.2.** Demostrador de Integración Datos Biomédicos (T5.1, M30)
- **E5.3.** Procesos de Instalación de Sistemas de Análisis de Datos Biomédicos (T5.1, M36)
- **E5.5.** Especificaciones Revisadas para la Inclusión de Marcadores Biomédicos de Imagen Médica (T5.2, M18)
- **E5.6.** Informe Final sobre el Funcionamiento de la Red IMPaCT de Beacons (T5.2, M36)

WP6 - Casos de Uso

BSC (Alfonso Valencia & Salvador Capella-Gutierrez)

SAS-HUVR (Carlos Parra)

WP6 - Casos de Uso

Visión: Contribuir a la **definición de los casos de uso propios de IMPaCT-Data** así como los casos de uso asociados a los otros programas de IMPaCT. Poner en valor y relevancia la calidad del dato como elemento vertebrador de las actividades de IMPaCT-Data.

- **Objetivo(s) vinculados a IMPaCT-Data**

- **O09. Desarrollo de prototipos de integración** de los resultados del análisis genómico y de imagen con la información normalizada extraída de las HCE, en los formatos adecuados para el análisis humano y computacional.
- **O10 Implementación de ecosistemas de “demostradores”** sobre funciones avanzadas de interoperabilidad de información traslacional, permitiendo la inclusión de aportaciones de empresas, entidades y grupos académicos que, comprometidos con el proyecto, complementen la infraestructura de interoperabilidad.
- **O11 Evaluación y ejecución concertada en distintas localizaciones de los demostradores de gestión**, en colaboración directa con la red **TransBioNet** y con otros actores del sistema de salud.

- **Resultados**

- **R1** Demostración práctica de los casos de uso planteados.
- **R2** Plan de Gestión de Datos de la estrategia IMPaCT implantado de acuerdo al Modelo de Madurez de datos FAIR de la RDA (Research Data Alliance).

WP6 - Visión inicial Tareas

- **T6.1:** Desarrollo de los **casos de uso alineados con IMPaCT (BSC)**
 - **COVID-19 como demostrador** de los sistemas para analizar conjuntamente datos de genómica (EGA y ENA), información médica (variables ISARIC o complementarias) e imágenes en respuesta a cuestiones relacionadas con el curso clínico de la enfermedad, relaciones con otras enfermedades o efecto de medicamentos.
 - **Cáncer y Enfermedades raras, farmacogenómica.** Siguiendo los desarrollos y datos generados por el programa de Medicina Genómica para convertirlos en protocolos ejecutables de modo sostenido en la infraestructura computacional del proyecto.
 - **Interacción con el Programa de Medicina Preventiva,** para conectar los desarrollos y contribuir al mantenimiento de sistemas comunes.
- **T6.2:** Evaluación de aspectos de **seguridad en el manejo de datos sensibles**, como procesos de pseudonimización, accesibilidad, trazabilidad y privacidad de los datos. (**SAS-HUVR**)
 - Esta tarea se encargará del diseño e implementación del modelo de evaluación de la gestión de datos de IMPaCT de acuerdo al nivel de cumplimiento de los **principios FAIR** y de acuerdo a los objetivos de la **Línea Estratégica Transversal 1 (LET1) de Integridad Científica**, integrando en la evaluación dimensiones relacionados con la protección de datos personales, aspectos éticos, regulatorios, organizativos y sociales de los datos de diferente naturaleza y origen manejados en los casos de uso, así como las reglas de acceso y uso para los datos de la cohorte del programa de medicina predictiva.

WP6 - Primeros Entregables

- **E6.1.** Descripción de los Casos de Uso (T6.1, M12)
 - Descripción inicial de los casos de uso a desarrollar durante el ciclo de vida de IMPaCT-Data.
- **E6.4.** Aspectos de Seguridad en el Manejo de Datos Sensibles (T6.2, M18)
 - Descripción inicial de los aspectos de seguridad para la FAIRificación de datos sensibles, como procesos de pseudonimización, custodia, accesibilidad, trazabilidad y privacidad de los datos.
- **E6.2.** Diseño de Procesos de Evaluación de los Casos de Uso (T5.1, M24)
- **E6.3.** Resultados de los Casos de Uso (T5.1, M36)
- **E6.5.** Guías de Buenas Prácticas en el Manejo de Datos Reales (T5.2, M36)

LET1 - Integridad Científica (IMPACT)

- **LET1:** Identificar las cuestiones ético-legales
- **LET1.2: Establecer procedimientos para garantizar los principios FAIR**
- **LET1.3:** Establecer las provisiones para la adecuada protección de los datos personales de los participantes de la cohorte CIBER-SNS
- **LET1.4:** Identificar las normas legales y recomendaciones éticas relacionadas con datos ómicos
- **LET1.5:** Establecer recomendaciones y herramientas a aplicar en IMPACT para la obtención de datos genómicos
- **LET1.6:** Identificación y revisión ético-legal de las estructuras existentes para el manejo de datos de carácter sensible
- **LET1.7:** Análisis del grado de implementación de los procedimientos ético-legales
- **LET1.8:** Análisis del grado de implementación de los procedimientos ético-legales, sus dificultades, y propuestas de resolución, entre los clínicos, investigadores y sistemas de salud
- **LET1.9: Catálogo del software y sistemas de investigación desarrollado y/o utilizado en el marco de IMPACT**

LET1 - Integridad Científica (BSC)

Tareas

- **LET1.2.** Establecer **procedimientos para garantizar los principios FAIR**
- **LET1.9.** **Catálogo del software** y sistemas de investigación desarrollado y/o utilizado en el marco de IMPaCT

Entregables

- **E.1.** Sistema de Evaluación de principios FAIR (M12)
- **E.8.** Informe de Buenas Prácticas para Desarrollo de Software y Principios de Ciencia Abierta (M36)

LET2 - Internacionalización (BSC)

Tareas

- **LET2.1.** Participación activa y posiciones de liderazgo en Infraestructuras de Investigación enfocadas al manejo y explotación de datos biomédicos
- **LET2.2.** Participación activa en la **Alianza Global para la Genómica y la Salud**
- **LET2.3.** Participación activa y posiciones de liderazgo en la Iniciativa de los Estados Miembros del **1+ Million Genomes bajo el liderazgo del ISCIII**
- **LET2.4.** Participación activa en la construcción del Espacio Europeo de Datos en Salud
- **LET2.5.** Situar a España como centro de coordinación y análisis de datos de **ICGC-ARGO** en Europa
- **LET2.6.** Participación en proyectos Europeos relevantes para IMPaCT

Entregables

- **E.9.** Informe sobre **Internacionalización de IMPaCT-Data** (M36)

Siguientes Pasos

Planificación

WP	T	Coordina	2021												2022												2023											
			1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36
WP1	T1.1	BSC-CNS											E1.0												E1.0												E1.0	
												E1.1												E1.1												E1.1		
	T1.2	BSC-CNS										E1.2												E1.2												E1.2		
	T1.3	BSC-CNS										E1.3												E1.3												E1.3		
	T1.4	BSC-CNS										E1.4												E1.4												E1.4		
	T1.5	BSC-CNS										E1.5												E1.5												E1.5		
	T1.6	CNIO									E1.6												E1.6												E1.6			
WP2	T2.1	BSC-CNS										E2.1						E2.2																E2.3				
	T2.2	BSC-CNS																E2.4																			E2.5	
WP3	T3.1	CRG										E3.1						E3.2																			E3.3	
	T3.2	CRG										E3.4																									E3.5	
WP4	T4.1	IACS																E4.1						E4.2												E4.3		
	T4.2	IACS																E4.4						E4.5												E4.6		
WP5	T5.1	FPS																E5.1													E5.2						E5.3	
	T5.2	FPS										E5.4						E5.5																		E5.6		
WP6	T6.1	BSC-CNS										E6.1												E6.2												E6.3		
	T6.2	SAS-HUVR																E6.4																		E6.5		
LET1	L1.2	BSC-CNS										E.1																										
	L1.9	BSC-CNS																																		E.8		
LET2	all	BSC-CNS																																		E.9		

Digital Twin for Future Medicine

