EPTRI

EUROPEAN PAEDIATRIC TRANSLATIONAL RESEARCH INFRASTRUCTURE

EPTRI-ELIXIR common service "Paediatric Data Interoperability"

EPTRI Open Meeting – April 2, 2020

This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No 777554

BIG DATA IN BIOLOGY: DATA COLLECTION, ANALYSIS AND INTEGRATION

Genome

EpiGenome

Transcriptome

Proteome

Metabolome

Microbiome

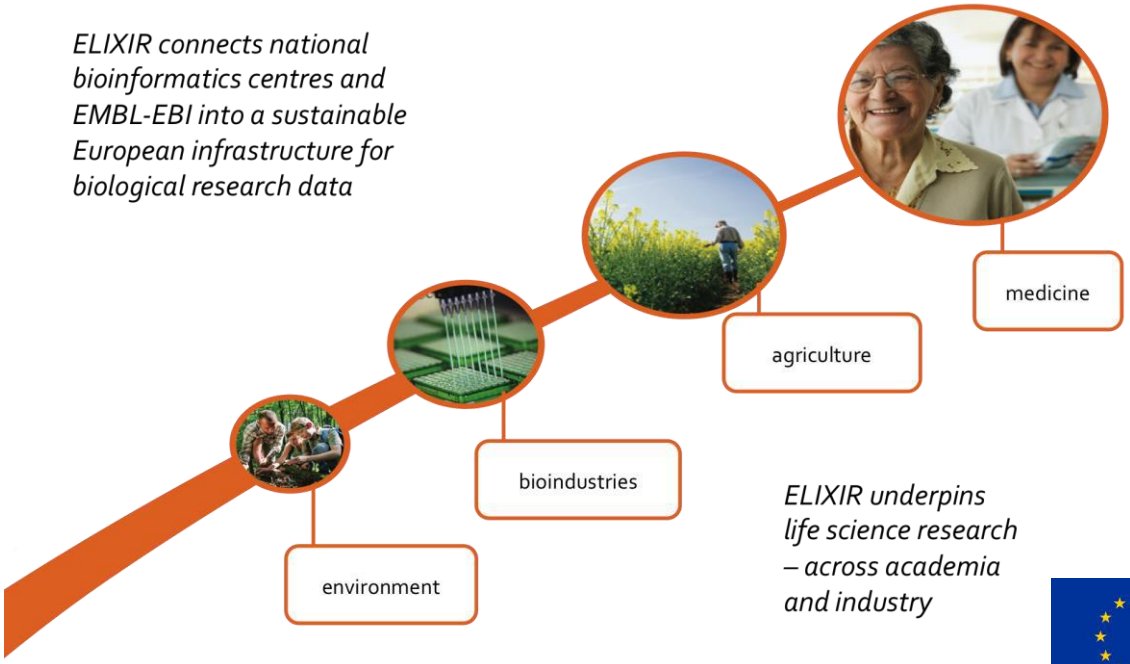
A pan-European sustainable European infrastructure for biological information (e.g. Omics data) is thus critically needed for supporting life science research and its translation to medicine, agriculture, bioindustries and society.

ELIXIR: a Research Infrastructure to face the Big Data challenge in Biology in Biology

ELIXIR is an **intergovernmental organisation,** formally established in 2016 as a Landmark European Research Infrastructure, that brings together "**bioinformatic resources" for life sciences** from across Europe. These resources include **databases, software tools, training materials, best practices, cloud storage** and **supercomputers**.

The goal of ELIXIR is to coordinate these resources so that they form **a single infrastructure**. This infrastructure makes it easier for scientists **to find and share data, exchange expertise, and agree on best practices.** Ultimately, it will help them gain **new insights** into how living organisms work.

ELIXIR connects national bioinformatics centres and EMBL-EBI into a sustainable European infrastructure for biological research data

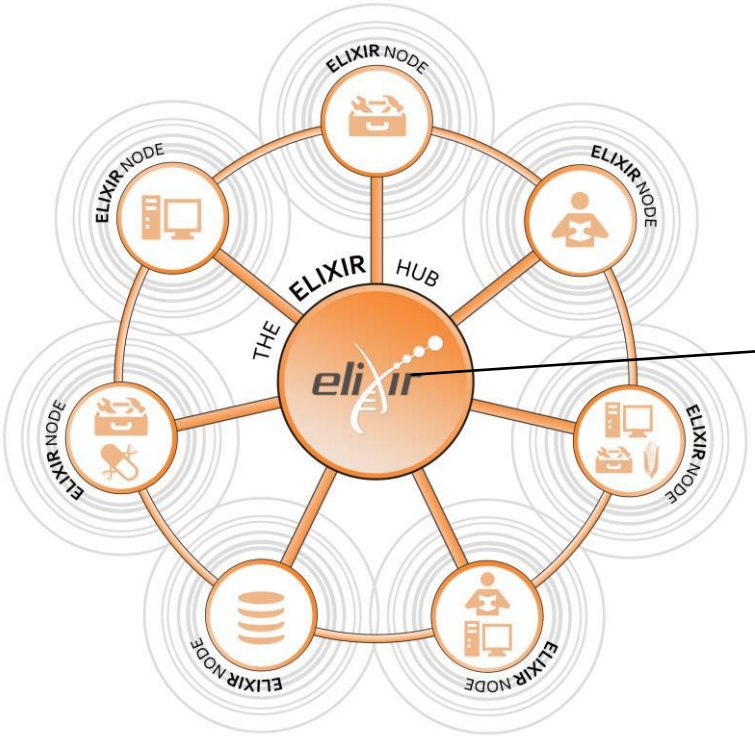


ELIXIR underpins life science research – across academia and industry

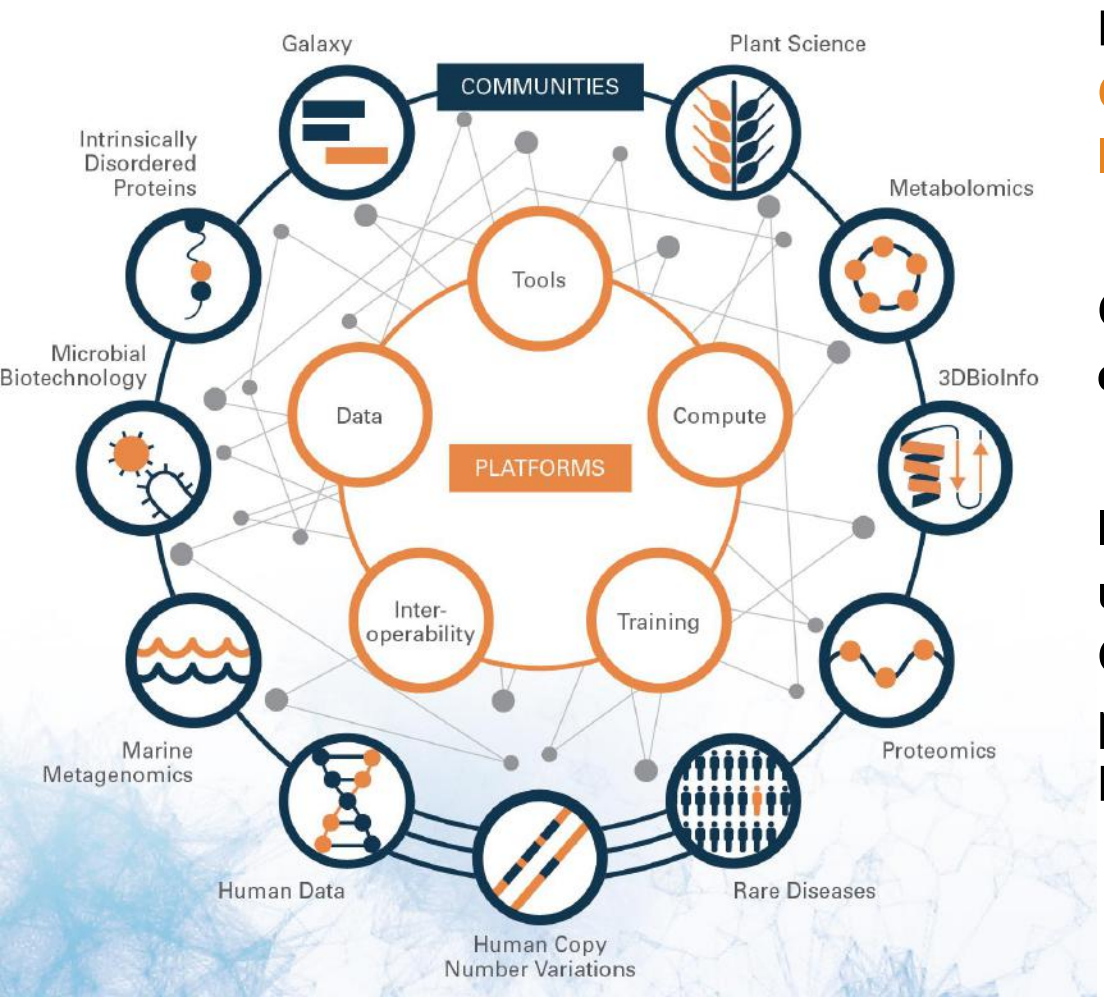
EPTRI
European Paediatric Translational Research Infrastructure

ELIXIR: A pan-european distributed Infrastructure for Bioinformatics

ELIXIR is structured as a central hub, located in the Wellcome Genome Campus (Hinxton, UK) **and 23 national nodes including over 160 Research Organizations.**



ELIXIR Organization



Five technical **platforms** for Compute, Data, Tools , Interoperability and Training

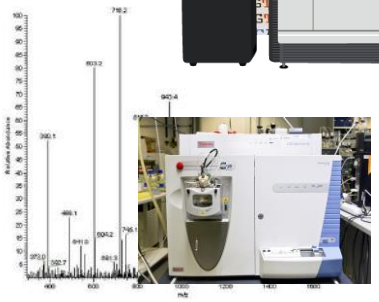
Complemented by several **user communities**

In the 2019-23 Scientific Programme use cases evolved in "User Communities" enlarging the ELIXIR portfolio such as Proteomics, Metabolomics, Galaxy, ..

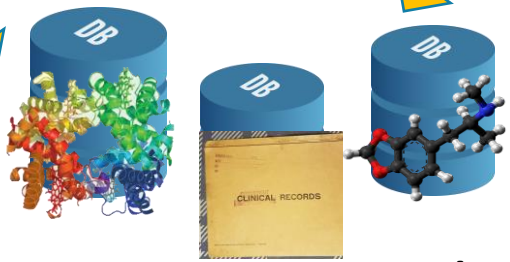
Big Data-driven innovation requires complex eco-systems

DATA ***COMPUTE*** ***TOOLS***

CLINICAL CASES

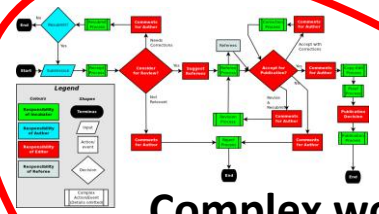


Heterogeneous and fragmented data



Heterogeneous and fragmented databases

Heterogeneous computational systems



Complex workflows

Different tools

ePTRI
EUROPEAN PAEDIATRIC TRANSLATIONAL RESEARCH INFRASTRUCTURE

What is needed to efficiently connect the ecosystem?

INTEROPERABILITY

- Standard formats
- Standard description of concepts (Ontologies)
- Standard and stable identifiers
- Rich, standard and machine-readable description of resources(data and tools) with metadata
- Clear access/privacy policies
- Technologies to deploy tools on different computational architectures
- Languages to easily connect different tools/data in workflows

FAIR principles

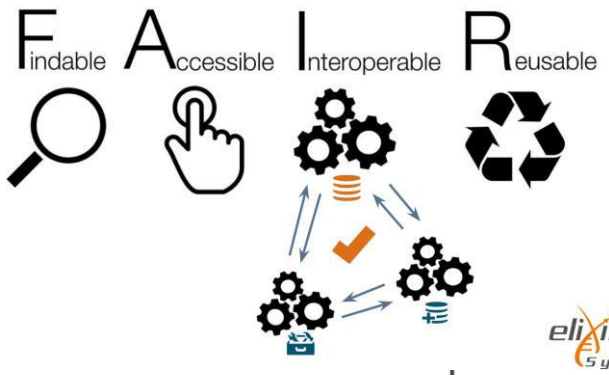
Box 2 | The FAIR Guiding Principles

To be Findable:
F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource

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 Reusable:
R1. meta(data) are richly described with 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 detailed provenance
R1.3. (meta)data meet domain-relevant community standards



https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4792175/pdf/sdata201618.pdf

Benefits of FAIR data principle adoption

Process automation through machine readability (of data and metadata)

Effective **streamlined acquisition, integration and analysis of data**

Minimization/elimir of data wrangling

Scientific queries are answered more rapidly in a flexible way.

time-to-value are significantly reduced

R&D can be accelerated.

developing more-segmented or -personalized medicines

e-resources for EPTRI data

Electronic documents and data e-library	Resources to ***store, cure and preserve all the digital documents and data*** produced as a result of research activities during EPTRI, offering a central location where authorised users can upload and download files, in several different forms and formats.
"Data interoperability" Common Service	Tools for ***discovering, accessing, integrating and analysing biological data to facilitate sharing and re-use*** of data according to the FAIR principles.
Text mining and Natural Language Processing tools	Tools ***for semantic search and classification to index all documents and tag them with the appropriate metadata***. The final goal is to extract quality and coherent knowledge from digital documents and data. Elixir can offer some solutions

Elixir-IT, Elixir-LU

Common service for data interoperability

Standards: formats, reporting guidelines, ontologies

Metadata services: ontology, annotation, validation, harvesting, Indexing

Register services and datasets

Search engine for datasets.

Identifier resolution & management

Identifier mapping services

Describing and sharing **workflows** between different systems

Harmonisation of **tools and pipelines**

Common **Programmable Interfaces**

Best practice.

https://elixir-europe.org/platforms/interoperability

FAIRsharing.org
standards, databases, policies

Bioschemas

Identifiers.org
Resolution service

BioContainers

elixir bio.tools

COMMON WORKFLOW LANGUAGE

Beacon

research object.org

EPTRI
European Paediatric Translational Research Infrastructure

FAIRsharing.org
standards, databases, policies

FAIRsharing

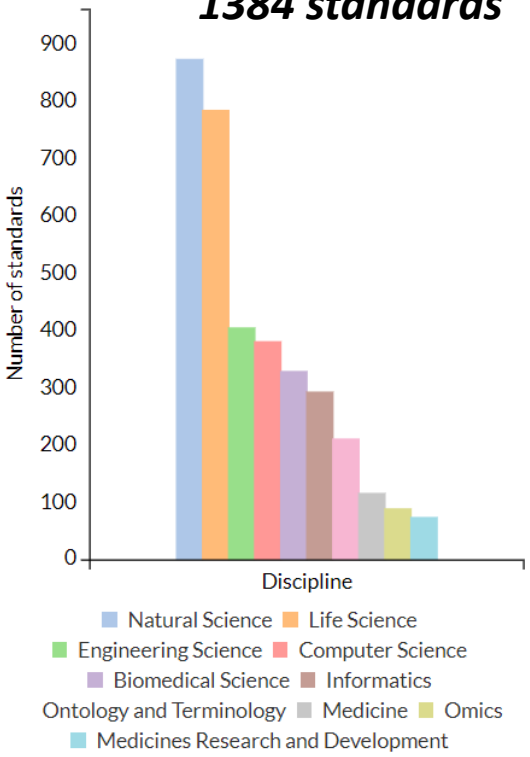
A curated, informative and educational resource on data and metadata *standards*, inter-related to *databases* and data *policies*.

https://fairsharing.org

Identify and cite the standards, databases or repositories that exist for your discipline

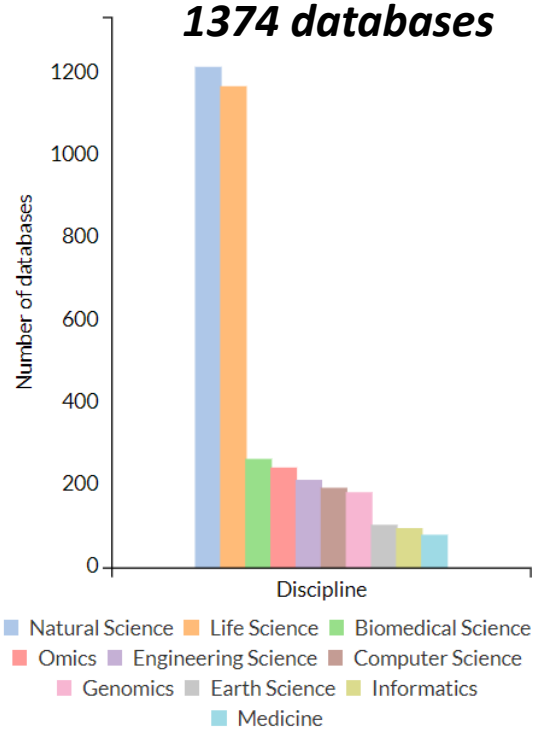
Top 10 disciplines covered by standards

1384 standards



Top 10 disciplines covered by databases

1374 databases

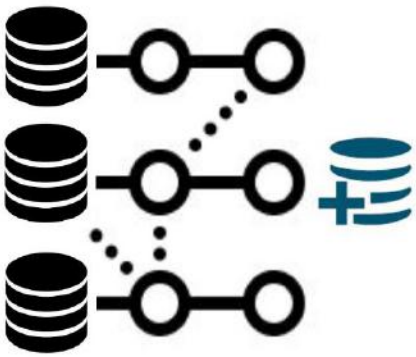


EPTRI
European Paediatric Translational Research Infrastructure

FAIRsharing.org
standards, databases, policies

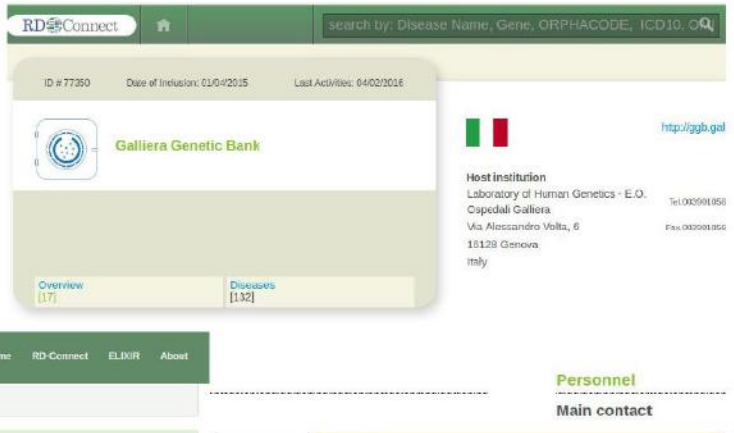
An example

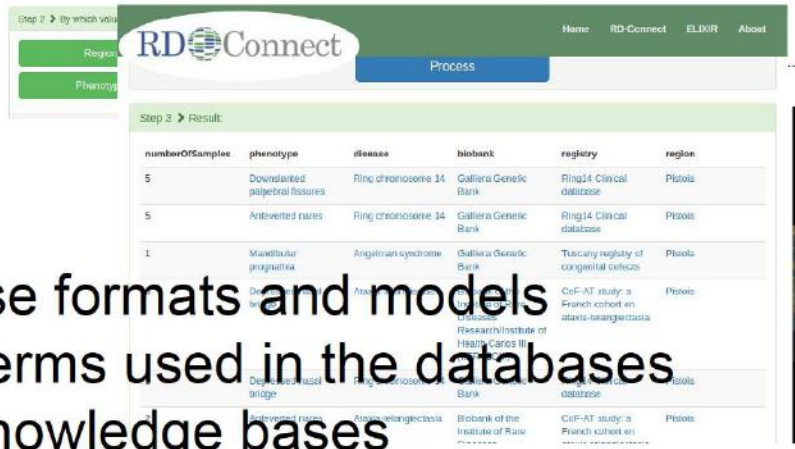
Rare Disease research

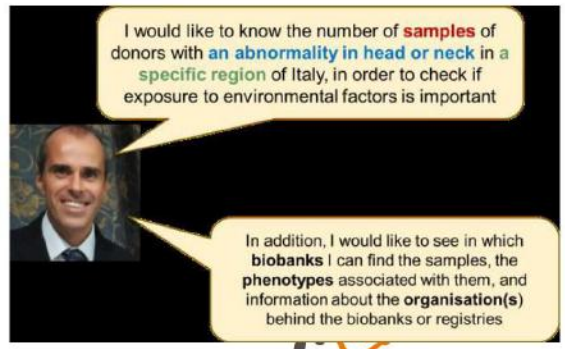


Harmonise database formats and models
Map between the terms used in the databases
Link to reference knowledge bases

Retrieval and analysis across resources







elixir

EPTRI
European Paediatric Translational Research Infrastructure

https://fairsharing.org

The service will collaborate in defining NEW STANDARDS for:

Formats.

Metadata. In short, ***it's data about data***. Many distinct types of metadata exist, including descriptive metadata, structural metadata, administrative metadata, reference metadata and statistical metadata.

Ontologies. An ontology encompasses a ***representation***, ***formal naming and definition of the categories, properties and relations between the concepts, data and entities*** of a particular domain of knowledge: e.g. Gene Ontology offers a controlled vocabulary and relationship among terms for describing gene/protein functions

Strong interaction with all the community is required

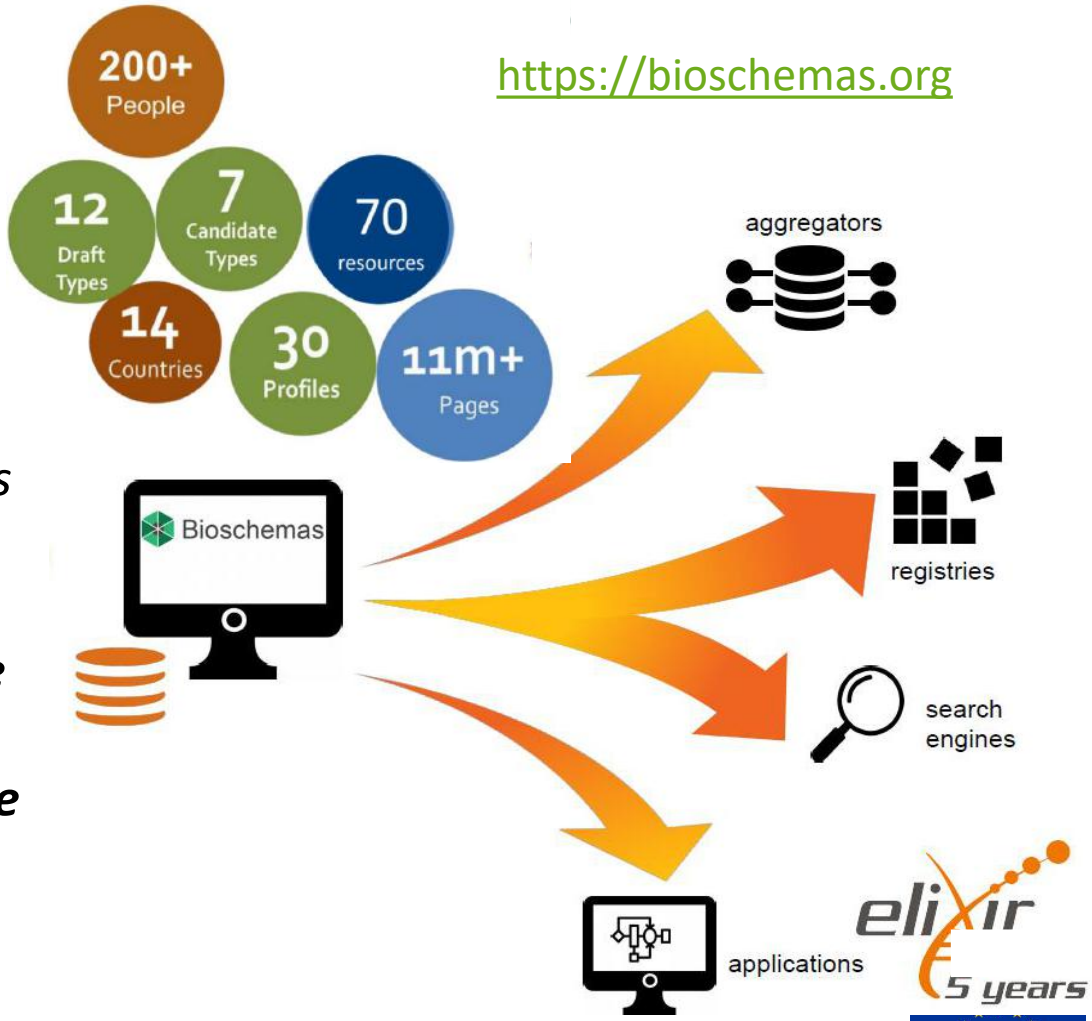
Bioschemas Data/resource findability

An Interoperability Resource for findability and metadata exchange for all of ELIXIR's Resources

https://bioschemas.org

Metadata *for web based resources using a widely adopted web standard in a community agreed way.*

This ***structured information*** *enables machines to understand what your metadata is in advance, making it easier to* ***find, integrate, and re-use*** *your data in their websites so that they are* ***discoverable*** *and* ***indexable by search engines*** *and other services.*

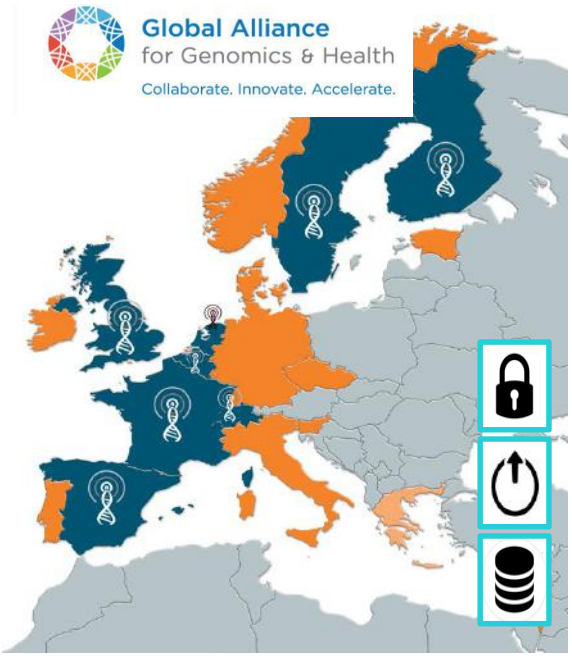


EPTRI
European Paediatric Translational Research Infrastructure

Data availability

The Beacon Network is a search engine across the world's public beacons. It enables global discovery of genetic mutations, federated across a large and growing network of

www.elixir-europe.org/beacons



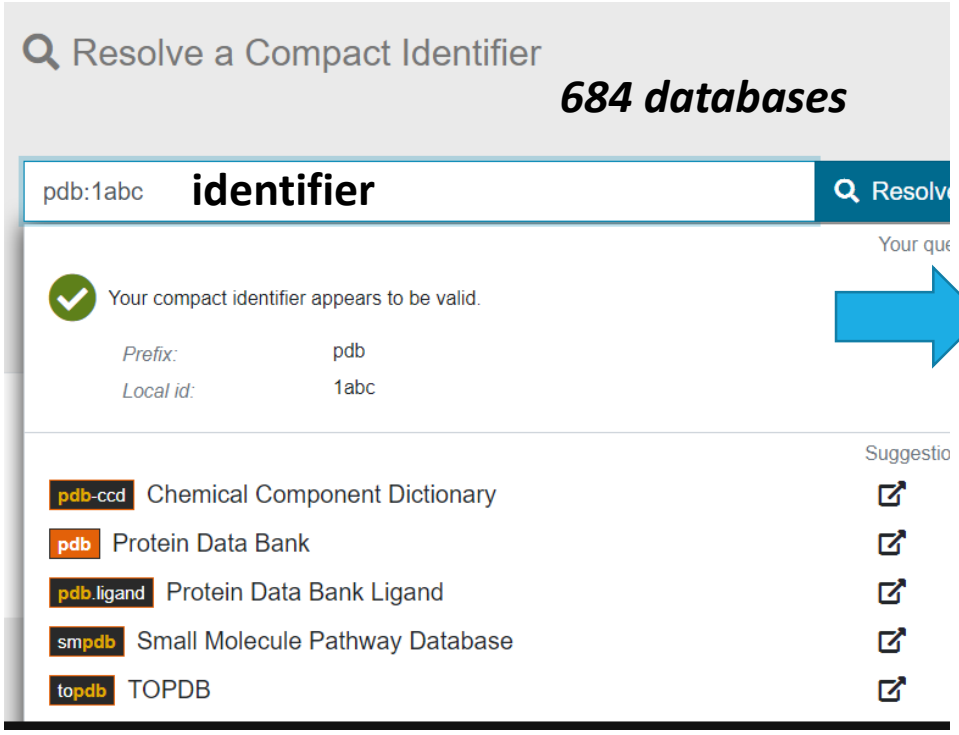
The system allows a user to ask whether a specific genomic variation has been documented in a given beacon (hospital, research center,..), while keeping all other sequence data concealed. This would allow a clinician to check whether a patient's mutation had been discovered in other patients without needing access to those other patients' genomes.

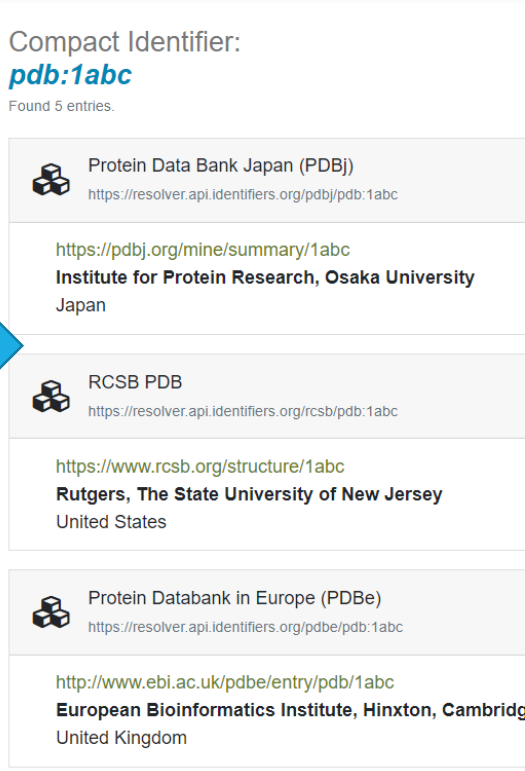
Consistent references

The Identifiers.org **Resolution Service** provides **consistent access** to life science data using **Compact Identifiers**.

http://identifiers.org/

*It handles **persistent identifiers** in the form of URIs and CURIEs. This allows the referencing of data in both a **location-independent** and **resource-dependent** manner.*





Reproducibility of analyses

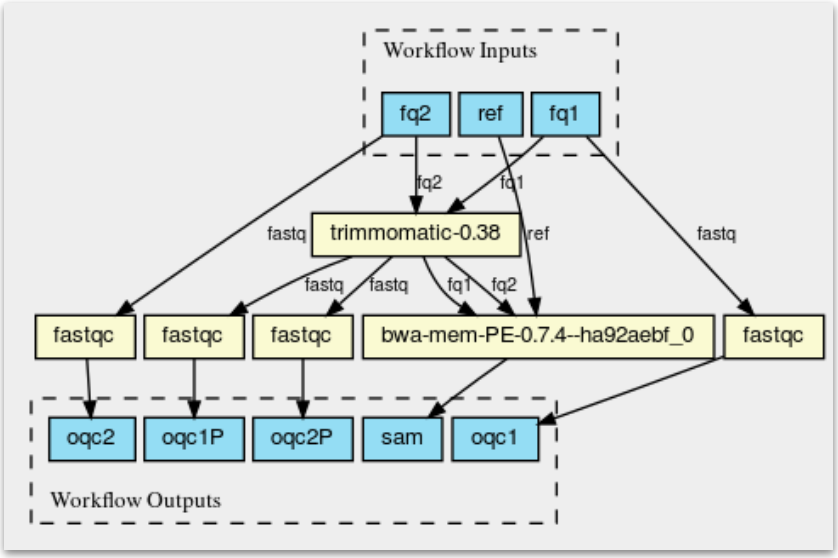
Open standards for distributing software and describing analysis workflows to make them portable and scalable across a variety of software and hardware environments

Biocontainers **provide ready-to-use packages and tools that can be easily deployed** and used on local machines, HPC and cloud architectures.

https://biocontainers.pro

CWL offers workflow descriptions that can be exported for re-executing analyses, ensuring consistency and reproducibilty on different environments, from workstations to cluster, cloud, and high performance computing (HPC) environments

https://www.commonwl.org



Common service for data interoperability

- Training on interoperability best practices to people involved in data management
- Definition of standard formats, metadata, ontologies (FAIRsharing)
- Implementation of new formats and ontologies dedicated to specific domains
- Best practices in database building and interoperable cross-reference (Identifiers)
- Annotation of existing and new resources to make them findable (Bioschemas)
- Inclusion in federated resources for sharing data on human variations (Beacon)
- Training on best practices to release new software and analysis workflows ensuring reusability (Biocontainers, CWL)

EPTRI

European Paediatric Translational Research Infrastructure