

# The European Nucleotide Archive: database of record, platform for data sharing and data brokering hub

Guy Cochrane

<https://globalbiodata.org/>



<http://www.ebi.ac.uk>



Collaboration

Sustainability

# EMBL European Bioinformatics Institute: data resources



## Chemicals, molecules and drug discovery

ChEBI  
ChEMBL  
MetaboLights  
Open Targets  
SureChEMBL



## Genes, genomes and RNA

ArrayExpress  
Ensembl  
European Nucleotide  
Archive  
Expression Atlas  
HGNC  
MGnify  
Rfam  
RNACentral  
VectorBase  
WormBase



## Proteins

AlphaFold DB  
Enzyme Portal  
InterPro  
PDBe  
PDBe-KB  
Pfam  
PRIDE  
UniProt  
UniProtKB



## Imaging and cellular structure

BioImage Archive  
Electron Microscopy  
Data Bank  
Electron Microscopy  
Public Image Archive



## Genetic variation and disease data

COVID-19  
Data Platform  
European  
Genome-phenome  
Archive  
European Variation  
Archive  
Mouse informatics



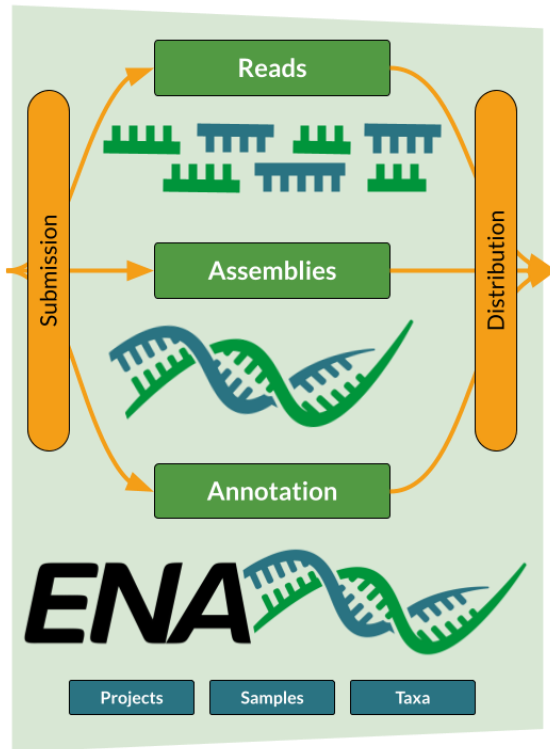
## Literature and knowledge management

BioModels  
BioSamples  
BioStudies  
Complex Portal  
Europe PMC  
GWAS Catalog  
IntAct  
OmicsDI  
Ontologies  
Reactome



# ***ENA and INSDC***

# The European Nucleotide Archive (ENA)

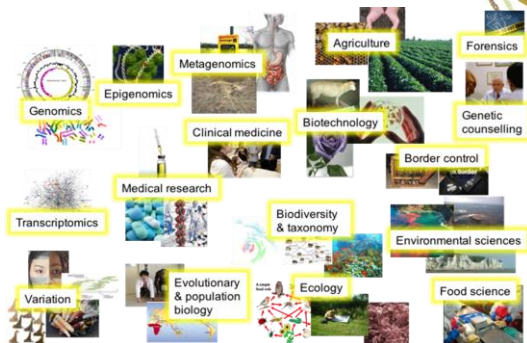


- The ENA provides the **scientific database of record** for sequencing data
  - Established in the early 1980s, extended for **new technologies and applications**
  - **Free and open access** for all users
  - **Broad scope** covering raw data, through layers of interpretation and context
  - Sequence data **foundation**
  - Rich submission, discovery and retrieval **software, tools and services**
  - **Support** through Helpdesk and training
- 
- **Management**, sharing, integration and dissemination of sequence data
  - **Data coordination** including project-specific data hubs and portals

Archive

Platform

<http://www.ebi.ac.uk/ena/>



The graph illustrates the rapid increase in genomic data over a 12-year period. Submissions and samples are the most prevalent data types, both exceeding 10 million by 2022. Other analyses and runs also show significant growth, reaching nearly 5 million. Genomes and studies, while growing, remain at lower volumes, around 3 million and 50,000 respectively by 2022.

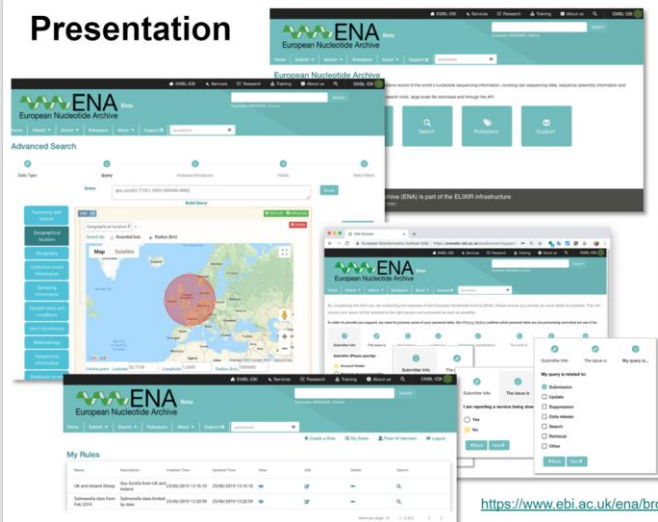
Year	submissions	samples	other analyses	runs	genomes	studies
2010	15,000	12,000	-	20,000	-	400
2012	80,000	100,000	-	120,000	-	1,500
2014	250,000	300,000	1,000	400,000	-	8,000
2016	600,000	1,000,000	80,000	1,200,000	20,000	15,000
2018	1,500,000	2,000,000	250,000	2,500,000	70,000	25,000
2020	3,000,000	4,000,000	1,000,000	4,000,000	250,000	40,000
2022	12,000,000	15,000,000	4,000,000	5,000,000	3,000,000	50,000

- **Rate:** 1 new dataset every 6 minutes
- **Data:**  $2 \times 10^9$  sequences and  $1 \times 10^{16}$  base pairs of read data across  $2 \times 10^6$  taxa
- **Usage:** 2,000 submitters; 10x thousands monthly consumers; 10x millions of monthly hits, many times this globally
- **Support:** 46 tickets per day and in-person training delivered to more than 350 users per annum
- **Adoption:** *the sequence database of record for the broadest scientific community*

Usage: map showing access to [www.ebi.ac.uk/ena](http://www.ebi.ac.uk/ena), year to October 2022, based on unique hosts

# ENA services

## Presentation



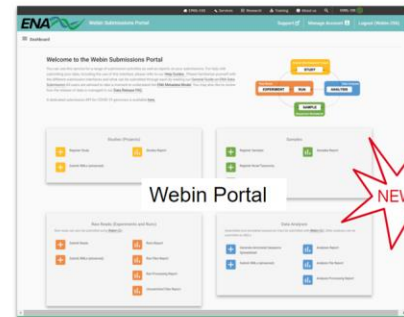
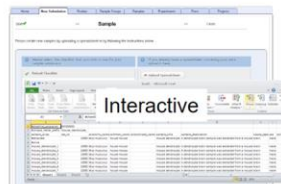
- Search
  - Sample characteristics
  - Function
  - Sequence similarity
- Browse
  - Study
  - Taxonomy
  - Sequencing method
  - Historic versions
- Filter
  - Reusable/shareable rules
  - Synchronisation with latest data
- Retrieve
  - Download tools

<https://www.ebi.ac.uk/ena/browser/home>

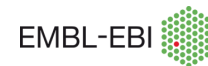


## Data submissions services

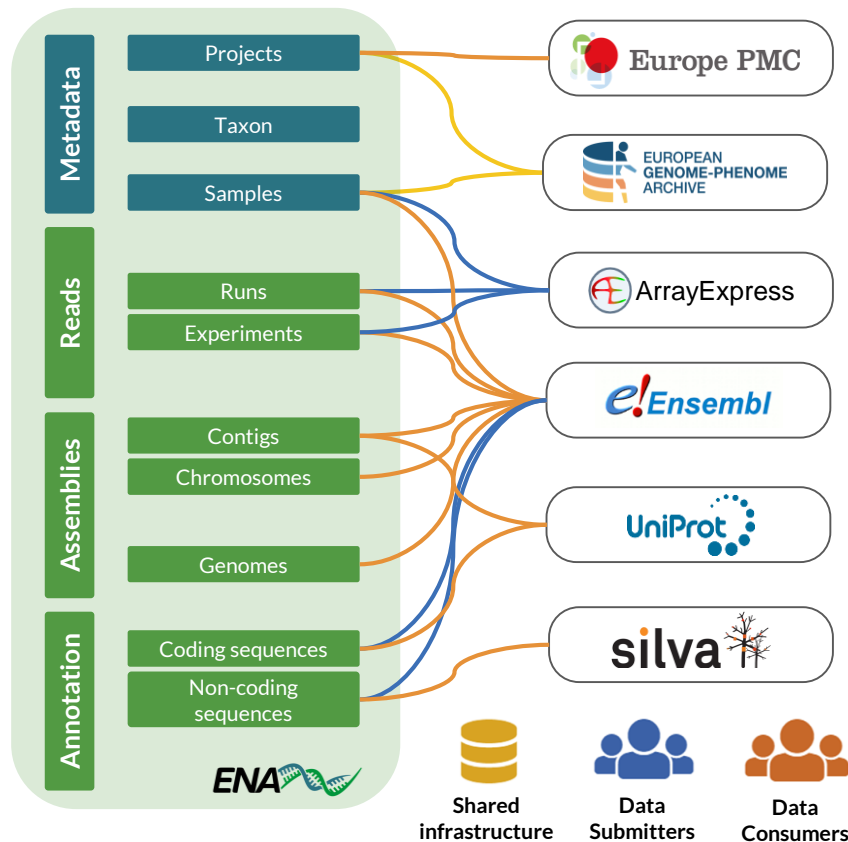
- Support across all ENA data types
- Submission, update, linking and tracking tools
- Programmatic
  - Webin-CLI, Webin-CLI REST and Metadata REST API
- Interactive
  - Family of interactive web tools
- Entry point Webin Portal
  - <https://www.ebi.ac.uk/ena/submit/webin/>
- Brokering toolkit



**ENA**   
European Nucleotide Archive



# Connectivity with other data resources



- The ENA is part of the wider ecosystem of biodata resources
- Recognised as an ELIXIR CDR and Global Core Biodata Resource by the Global Biodata Coalition
- Integration into a system of biodata resources increases value in genomics data



# International Nucleotide Sequence Database Collaboration



## Values

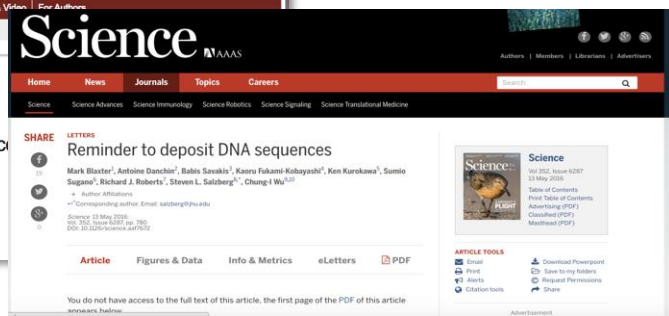
- open access for all
- globally comprehensive
- spanning life science domains
- permanent database of record
- public forum for the scientific process



<http://www.insdc.org/>

## Organisation

- established early 1980s
- major ongoing investment
- structure and governance
- model for scientific collaboration



## Instruments

- regular data exchange
- accession scheme
- data standards
- mandatory submission agreement
- services and software (node-level)

# ***Data coordination***

# European COVID-19 Data Platform

## Access



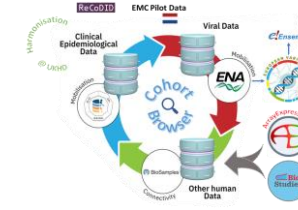
- 25M records across molecular, literature, imaging and social science, backed up by a network of 11 national Data Portals

## Data Hubs

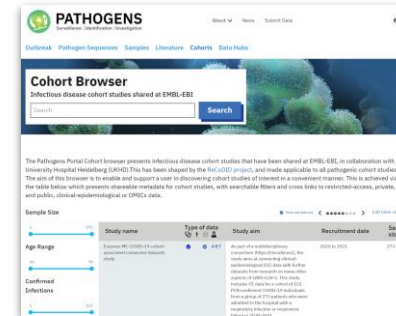
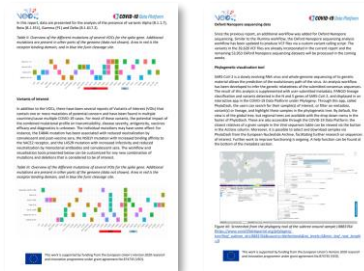
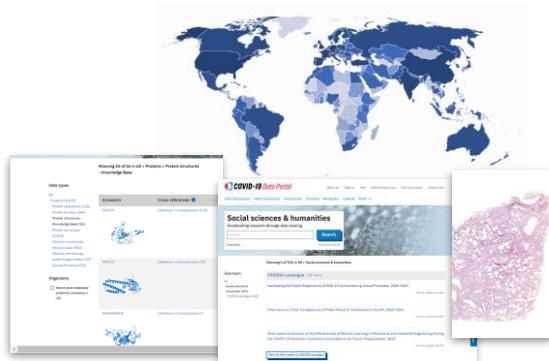


- Mobilisation (>6M) and systematic analysis (>4M) data sets from SARS-CoV-2 isolates from 121 countries
- 15 VEO public health reports

## Human data



- Demonstrated protocols to link between sensitive research subject and pathogen data, leveraging (federated) European infrastructure



# EarlyCause - early life stress

- Data portal, search functionalities
- Biodata to support investigations into lifelong effects of early-life stress
- <https://portal.earlycause.eu/>
- Data on organism level or data type (Mouse, Rat, Human, Cell Lines, Literature, Cohorts)
- Reusable infrastructure and framework to bring forward biodata (e.g. soil biodata)

The screenshot displays the EarlyCause portal interface. The top navigation bar includes links for Home, About, News, Partners, FAQ, Useful information, and Submit data. Below this, a secondary navigation bar lists categories: Cell Lines, Mouse, Rat, Human, Literature, Cohorts, and Tools.

The main content area is divided into two sections. The top section, titled "Early Cause", describes the portal's mission: "Investigating the lifelong effects of early life stress on health". It states that the portal aims to bring together various datasets to promote research on early life stress and its short- and long-term effects on psychology, cardiology, and metabolism. The portal enables the upload, searching, sharing, and analysis of relevant mouse, rat, human, and cell-line datasets. A "Read more" link with a right-pointing arrow is provided.

The bottom section, titled "Cohorts", provides a search interface. It includes a search bar with the placeholder text "Enter your search terms" and a "Search" button. Below the search bar, it lists related examples: "Birth cohort", "Generation", and "Pregnancy".

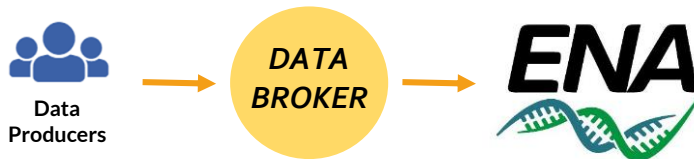
The "Cohorts" section also displays a table of results. The table has columns for "Data types", "Cohorts", "Study title", "Cross-references", and "Description". The table shows 7 results, with the first three being "Prenatal (3)", "Newborn (3)", and "Infant (3)". The table also includes a "Download" button and a "More..." link.

Data types	Cohorts	Study title	Cross-references	Description
Prenatal (3)	PID	Parents of NFBCL1966		
Newborn (3)	NFBCL1966GO	Avon Longitudinal Study of Parents and Children	Literature > Publications (2)	Based at Longitud ALSPAC a world-
Infant (3)	ALSPAC			

# ***Data submissions brokering***

# Data brokering

- A *data submissions broker* is a group that provides tools, services and/or infrastructure that end-users access and through which their data flows into ENA



- Long history of brokering with demonstrated mutual benefits
- Can be **generalist** or have **data-type** or **domain specific** expertise

- **32 active brokers**, with brokered data accounting for **16% of total data** submitted to ENA (*excluding institutional brokers registered as regular submitters*)
- **9% growth of active brokers** and **43% growth of brokered data** during the last year compared to the previous year's activity

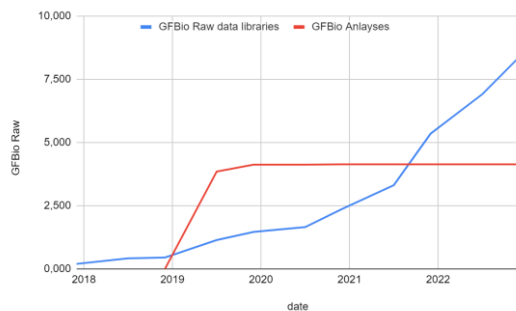
# Example: GFBIO

- FAIR data in biodiversity, ecology and environmental sciences

**ENA**  
European Nucleotide Archive

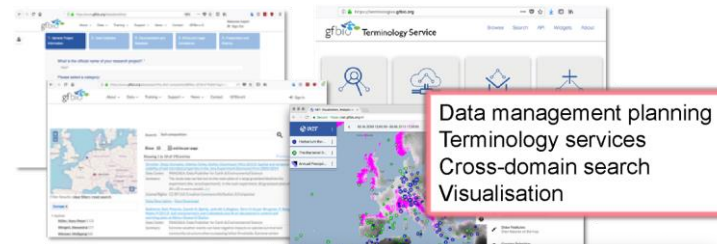
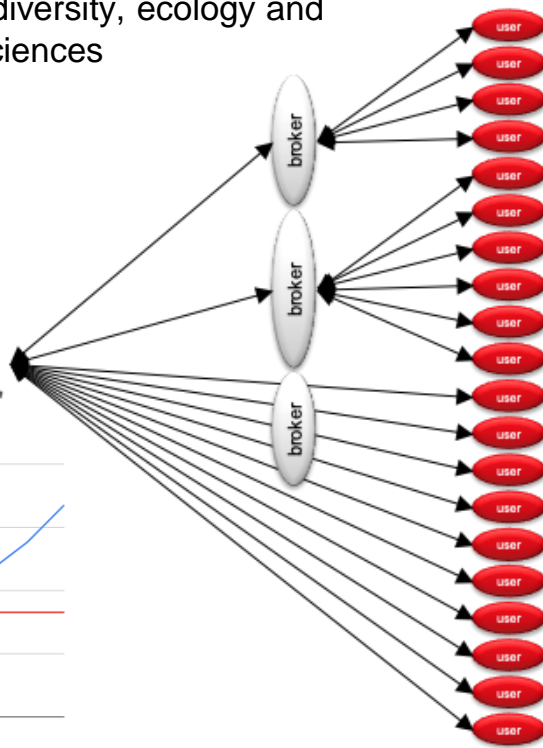
**EMBL-EBI**

**elixir**  
Core Data Resource



← Submissions

Delivery →



Data management planning  
Terminology services  
Cross-domain search  
Visualisation

Data upload  
& publish

**gfbio**  
GERMAN FEDERATION  
FOR BIOLOGICAL DATA

**NFDI4**  
BIODIVERSITY

Frontend  
Submission form  
Users

Helpdesk

Submission System

<https://>  
REST API  
Programmatic Access

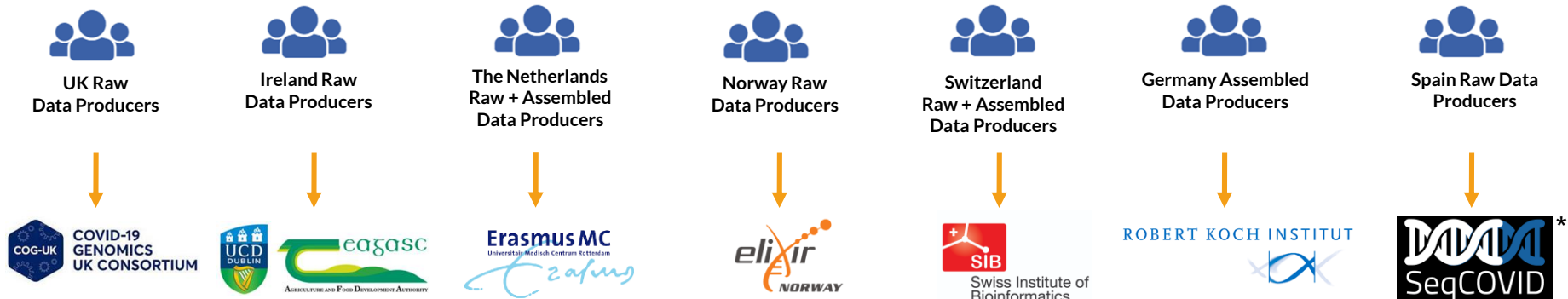
**ENA**  
European Nucleotide Archive

**PANGAEA**  
Data Publisher for Earth &  
Environmental Science

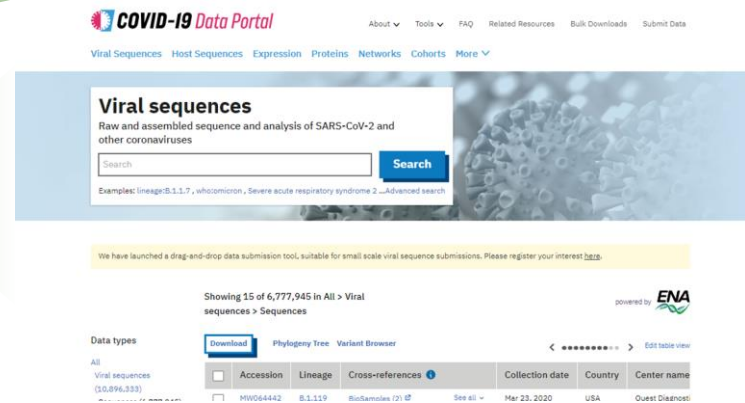
Collection Data Centers

**EMBL-EBI**

# Example: Global health and SARS-CoV-2



- Active network growth, especially national public health data brokers providing SARS-CoV-2 data
- Since COVID-19, 16 new brokers were established of which 6 remained active (\* +1 to be resumed)





***Future***

# Expansion of brokering network

## Brokering data to the ENA

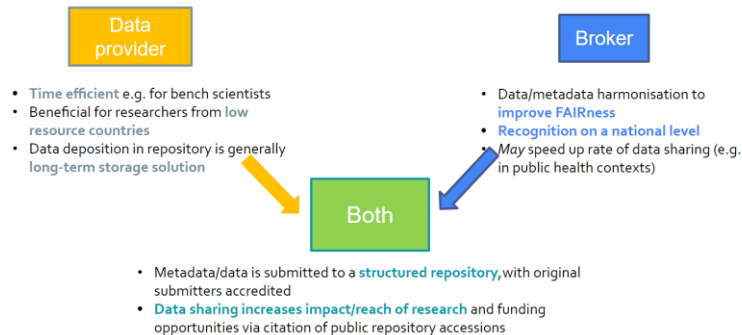


- An open-access, global nucleotide sequence repository, covering **raw sequence data**, **sequence assembly** information and **functional annotation** for all **non-human organisms**



- We are the European arm of the International Nucleotide Sequence Database Collaboration (INSDC)

## Benefits of data brokering



## Some of the possible considerations of a data broker

	Aspects to consider
<b>Data/metadata collection</b>	<ul style="list-style-type: none"> <li>File formats</li> <li>All must comply with national regulations</li> </ul>
<b>Data/metadata validation</b>	<ul style="list-style-type: none"> <li>Broker-side validation</li> <li>Will the broker <b>verify whether information provided is correct?</b></li> </ul>
<b>Processing data and metadata</b>	<ul style="list-style-type: none"> <li>Curation/masking sensitive metadata</li> <li>QC of data files</li> <li>Analysis of data?</li> <li>Beware legal implications -&gt; <b>broker role changes to data processor</b></li> <li>Or, is preference to receive data/metadata in mostly repository compliant format</li> </ul>
<b>Data storage</b>	<ul style="list-style-type: none"> <li>What happens to data/metadata after submission – is there a timeframe for storage?</li> </ul>
<b>Helpdesk support</b>	<ul style="list-style-type: none"> <li>In case of queries, <b>where can users seek help</b> – both broker + repository?</li> </ul>
<b>Data accreditation</b>	<ul style="list-style-type: none"> <li><b>How will the data provider be credited?</b></li> <li>What are deposition databases' policies on this?</li> </ul>

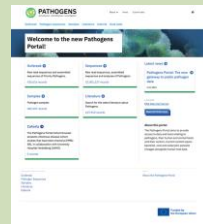


# All pathogens...

- Following success of the EU COVID-19 Data Platform, focus switching more to:
  - General pathogens
  - Pandemic preparedness
- **Pathogens Platform**
  - Two main components
  - Pathogens Portal includes an ‘Outbreaks’ page on selected ‘focus pathogens’
  - Pathogen Data Hubs - extend from COMPARE and SARS-CoV-2 counterparts → themed activities
- **Pathogens Portal:**  
<https://www.pathogensportal.org/>

## Pathogens Portal

Interface for  
pathogen life  
sciences data



## Pathogen Data Hubs

Tools to support submission,  
analysis, visualisation and  
presentation of pathogen  
sequence data in the  
Pathogens Portal



# Participating in INSDC



NCBI



Not just ENA...

- INSDC exploring how to engage global players
- Enabling broader participation

# ***Global Biodata Coalition***

# Global Biodata Coalition

- A coalition of funders of life science research
  - Recognising the essential nature of biodata resources in enabling and supporting life science research
  - Understanding that threats that exist to the sustainability of biodata resources
  - Embracing the need to cooperate globally to address sustainability
- Membership
  - 11 member and 5 observer\* life science research funding organisations around the world ***\*including National Research Foundation of Korea***
- Landscape analysis and classification
  - Inventory of 3112 biodata resources
  - 37 selected Global Core Biodata Resources
- Board and other Working Groups
  - Open data strategies
  - Biodata resource sustainability
  - Sustainability indicators



<https://globalbiodata.org/>

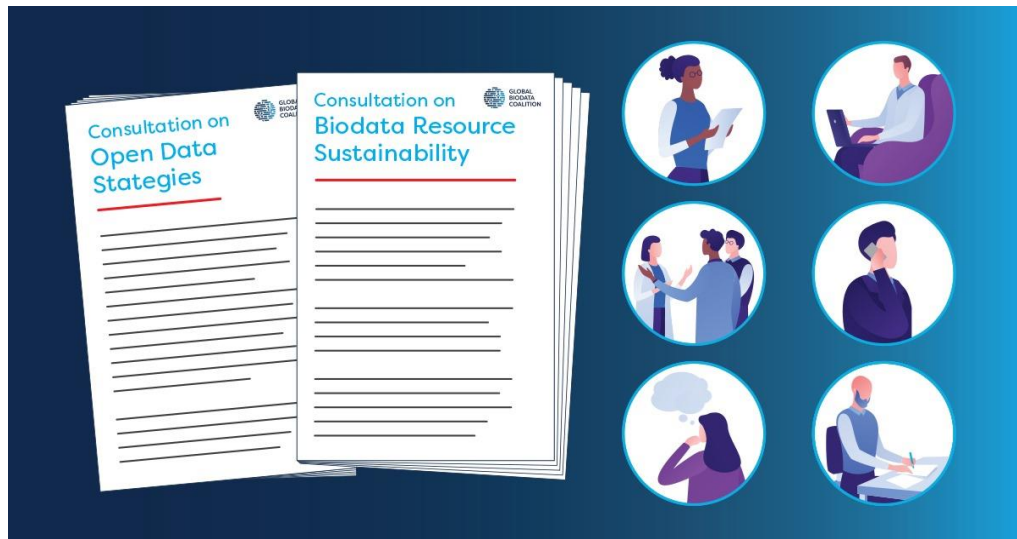
# GBC Board Working Groups

## Sustainability

- Premises and principles for sustainability
- What does sustainability mean to a funder?
- First explorations of mechanisms to support biodata resources

## Open Data Strategies

- Openness is key - an enabler and amplifier
- Exploration of policy
- Towards an environment supportive of international cooperation



<https://globalbiodata.org/what-we-do/sustainability/>

<https://globalbiodata.org/what-we-do/open-data-strategies/>

# Acknowledgements

## ***Data Coordination and Archiving (Cochrane)***

Alisha Ahamed, Carla Cummins, Khadim Gueye, Vikas Gupta, Maira Ihsan, Suran Jayathilaka, Vishnukumar Balavenkataraman Kadhivelu, Manish Kumar, Ankur Lathi, Jasmine McKinnon, Lili Meszaros, Colman O'Cathail, Joana Paupério, Stéphane Pesant, Nadim Rahman, Gabriele Rinck, Swati Suman, Zahra Waheed, Peter Woollard, Ahmad Zyoud, Guy Cochrane

## ***Archiving and Infrastructure Technology (Burdett)***

Dipayan Gupta, Eugene Ivanov, Muhammad Haseeb, Rajkumar Devraj, Rasko Leinonen, Milena Mansurova, Sandeep Selvakumar, Yanisa Sunthornyotin, Senthilnathan Vijayaraja, David Yuan, Tony Burdett

## ***Funding organisations***



## ***Member organisations***

## ***Observer organisations***

## ***Human Frontiers Science Programme***

## ***GBC Board Working Groups***

## ***GCBR Forum***

## ***Secretariat***

Chuck Cook, Rachel Drysdale, David Carr, Lindsey Crosswell, Dana Černošková, Heidi Imker, Ken Schackart, Guy Cochrane