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

Guy Cochrane

https://globalbiodata.org/

GLOBAL BIODATA COALITION

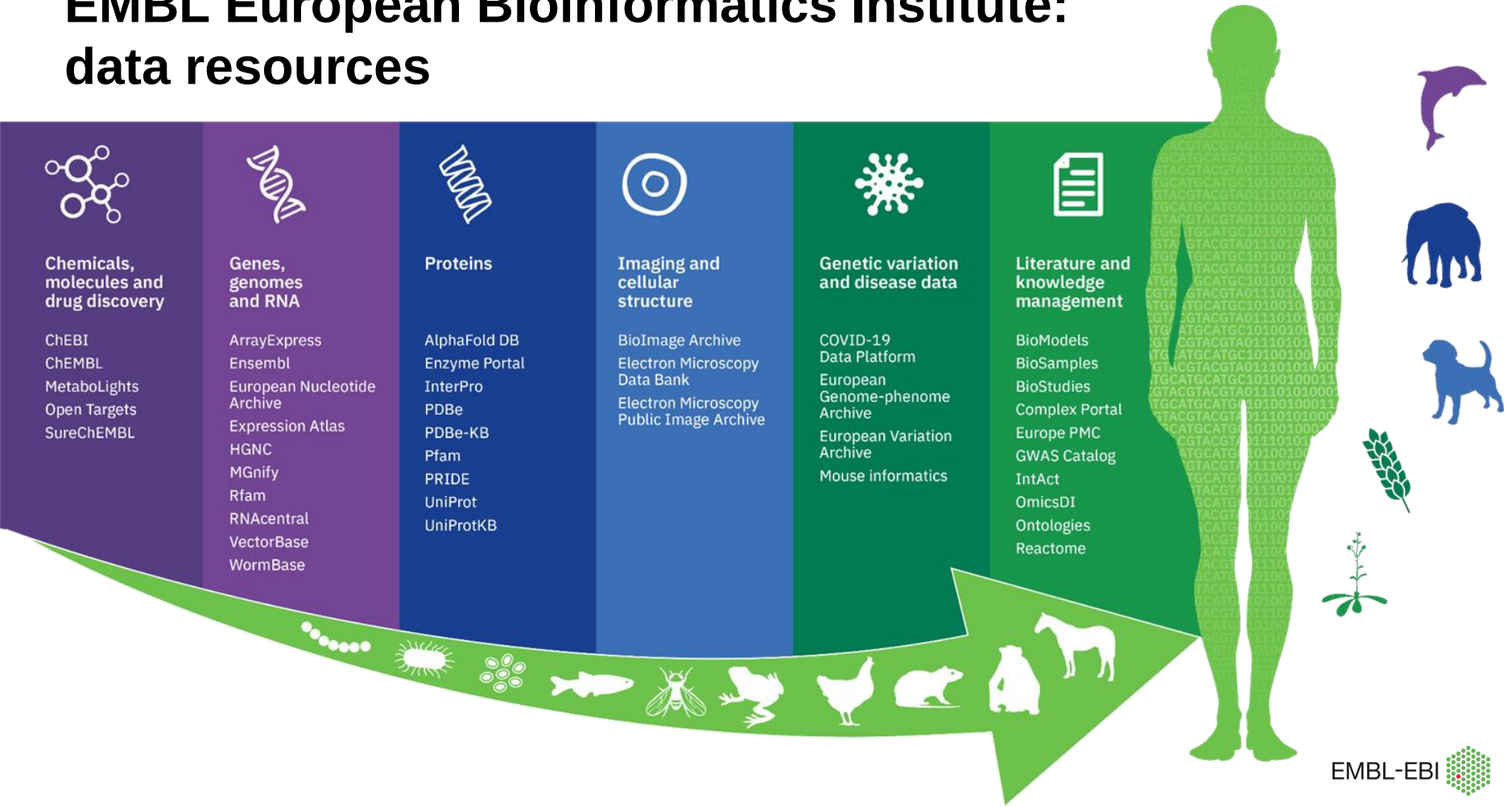
http://www.ebi.ac.uk

EMBL-EBI

Collaboration

Sustainability

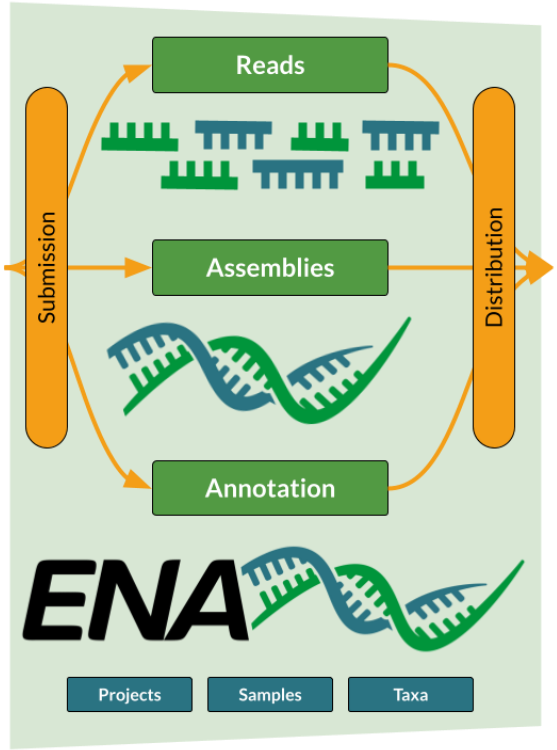
EMBL European Bioinformatics Institute: data resources



EMBL-EBI

ENA and INSDC

The European Nucleotide Archive (ENA)



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

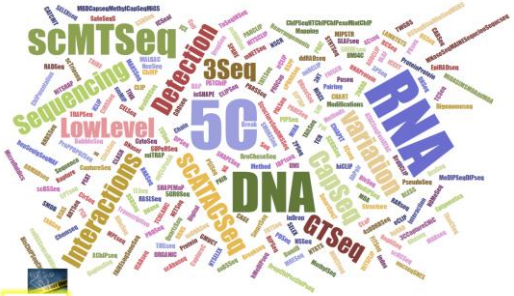
Archive

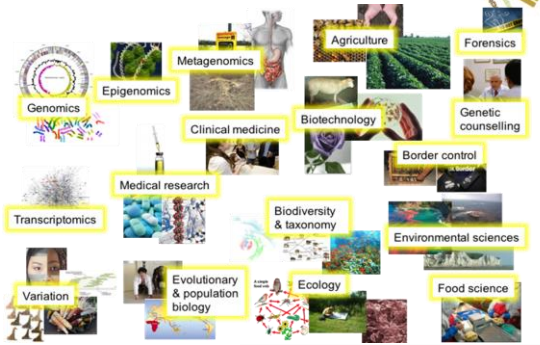
- **Management**, sharing, integration and dissemination of sequence data
- **Data coordination** including project-specific data hubs and portals

Platform

EMBL-EBI

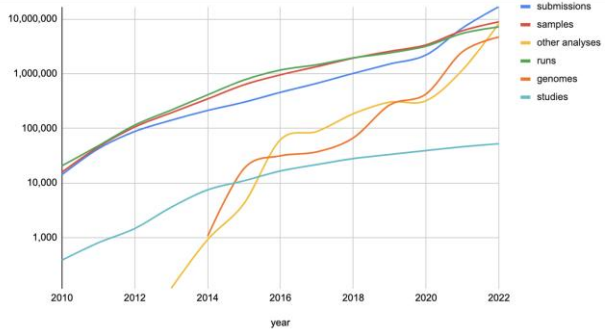
Scale





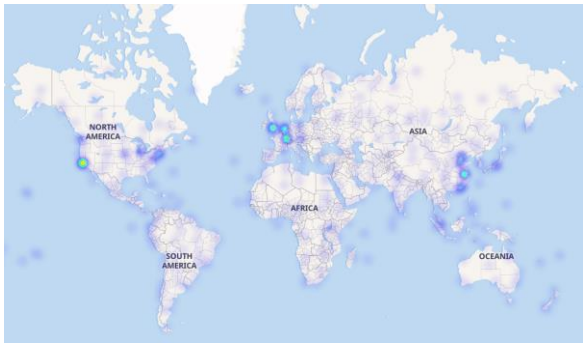
- **Rate**: 1 new dataset every 6 minutes
- **Data**: 2x10^{9} sequences and 1x10^{16} base pairs of read data across 2x10^{6} taxa
- **Usage**: 2,000 submitters; 10x thousands monthly consumers; 10x millions of monthly hits, many times this globally
- **Support**: 46 tickets per day and in-person training delivered to more than 350 users per annum
- **Adoption**: *the sequence database of record for the broadest scientific community*

Submission



Cumulative submission rates for EMBL-EBI routed samples, experiments, analyses and studies

Retrieval



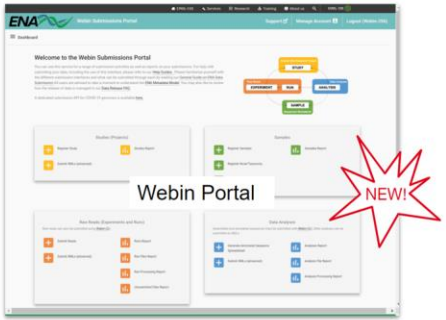
Usage: map showing access to www.ebi.ac.uk/ena, year to October 2022, based on unique hosts

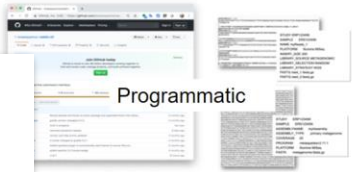
EMBL-EBI

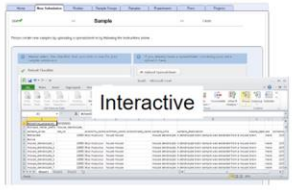
ENA services

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

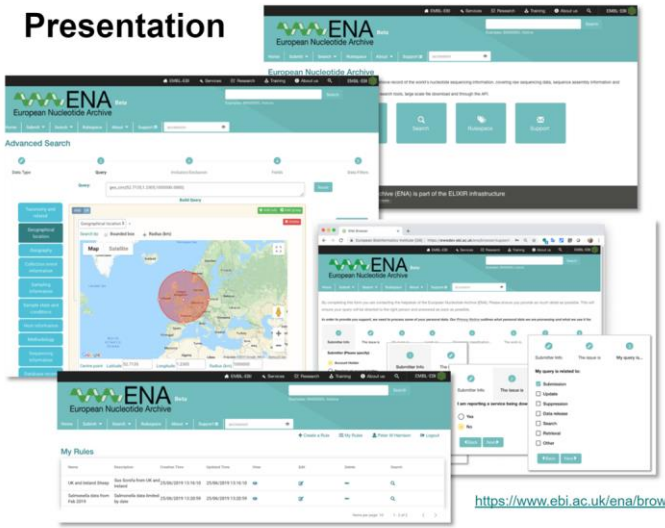






EMBL-EBI

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

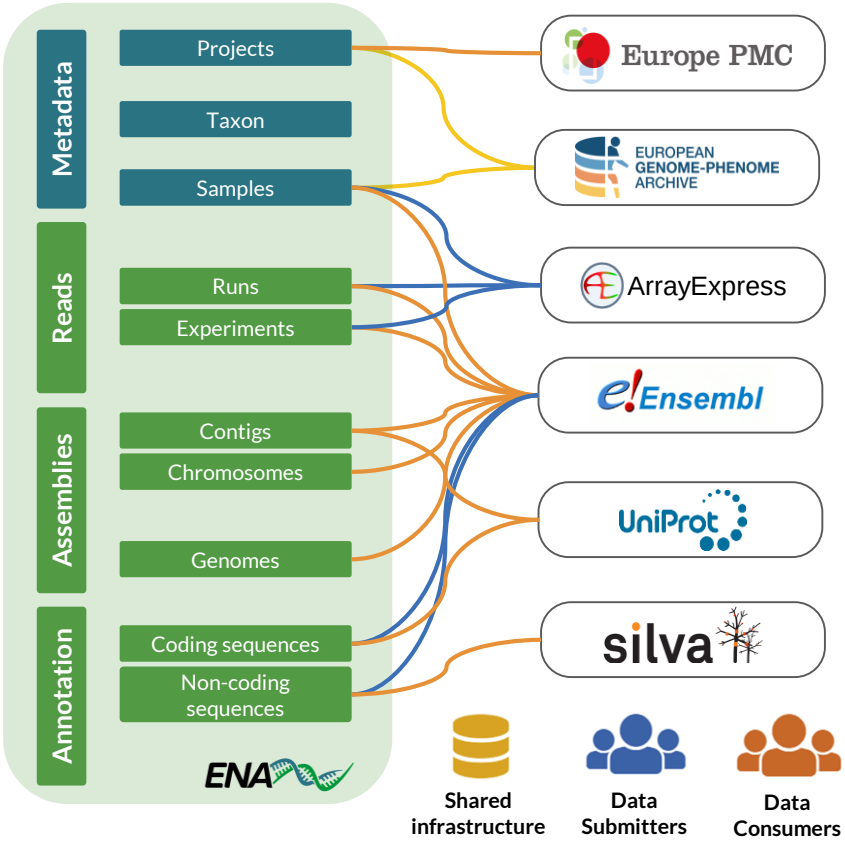
https://www.ebi.ac.uk/ena/browser/home

EMBL-EBI

ENA
European Nucleotide Archive

EMBL-EBI

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

GLOBAL CORE BIODATA RESOURCE

EMBL-EBI

International Nucleotide Sequence Database Collaboration

DDBJ
DNA Data Bank of Japan
National Institute of Genetics
INSDC
GenBank
NCBI
ENA
European Nucleotide Archive
EMBL-EBI

Values

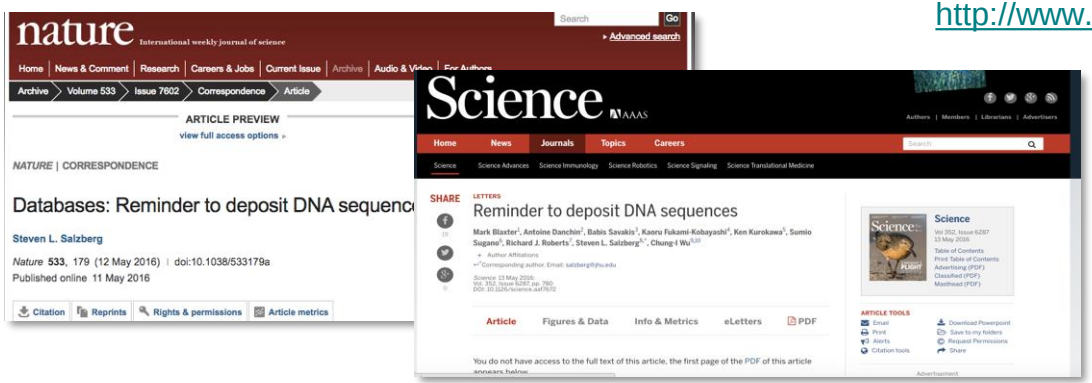
- open access for all
- globally comprehensive
- spanning life science domains
- permanent database of record
- public forum for the scientific process

Organisation

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

INSDC

http://www.insdc.org/



Instruments

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

Data coordination

European COVID-19 Data Platform

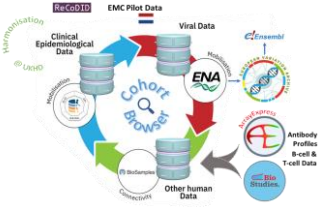
Access

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

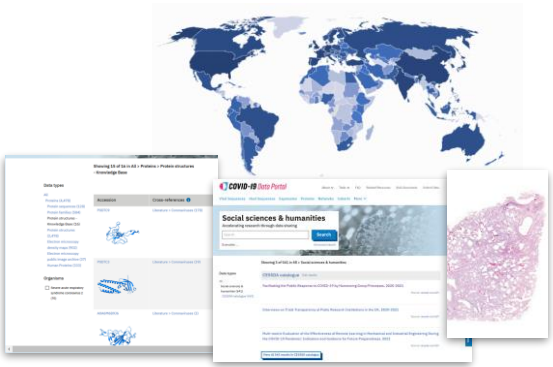
Data Hubs

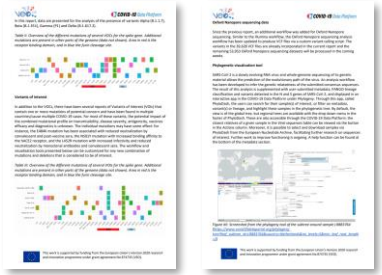
- Mobilisation (>6M) and systematic analysis (>4M) data sets from SARS-CoV-2 isolates from 121 countries
- 15 VEO public health reports

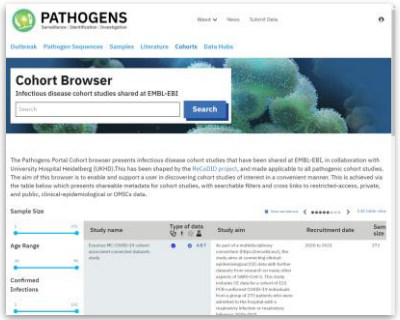
Human data



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

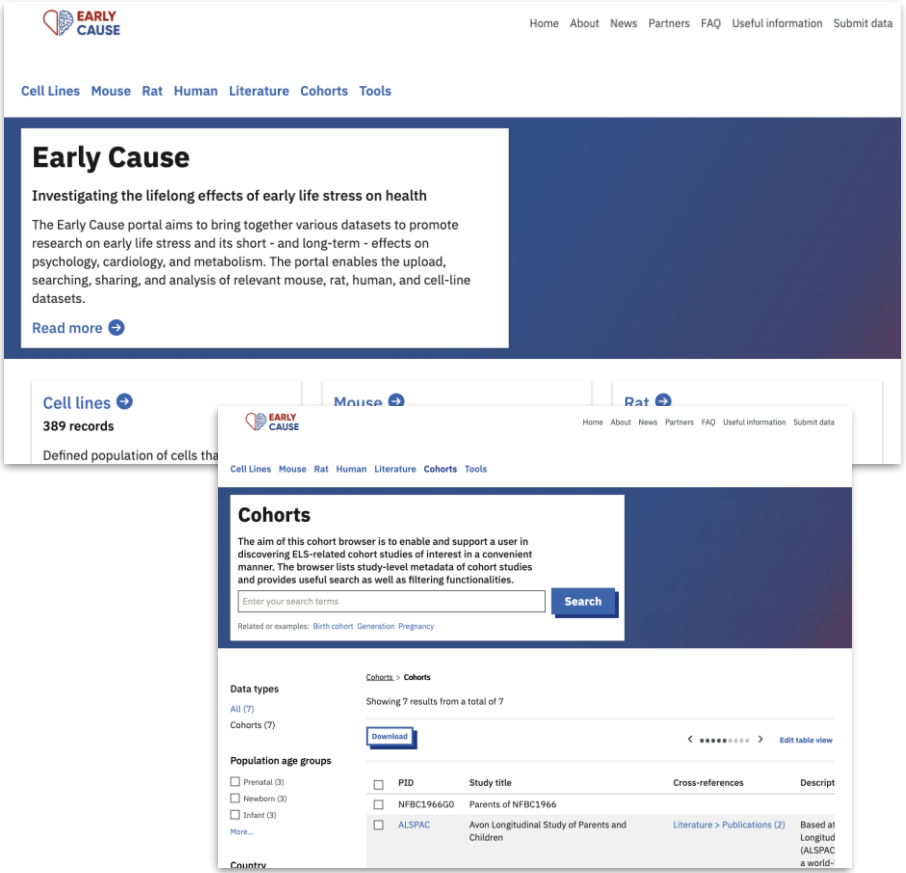






EarlyCause - early life stress

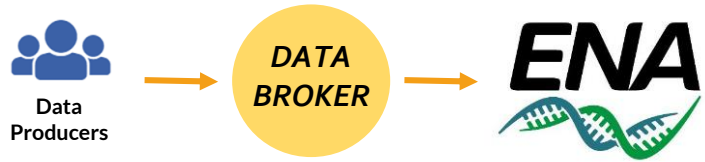
- Data portal, search functionalities
- Biodata to support investigations into lifelong effects of early-life stress
- https://portal.earlycause.eu/
- Data on organism level or data type (Mouse, Rat, Human, Cell Lines, Literature, Cohorts)
- Reusable infrastructure and framework to bring forward biodata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



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

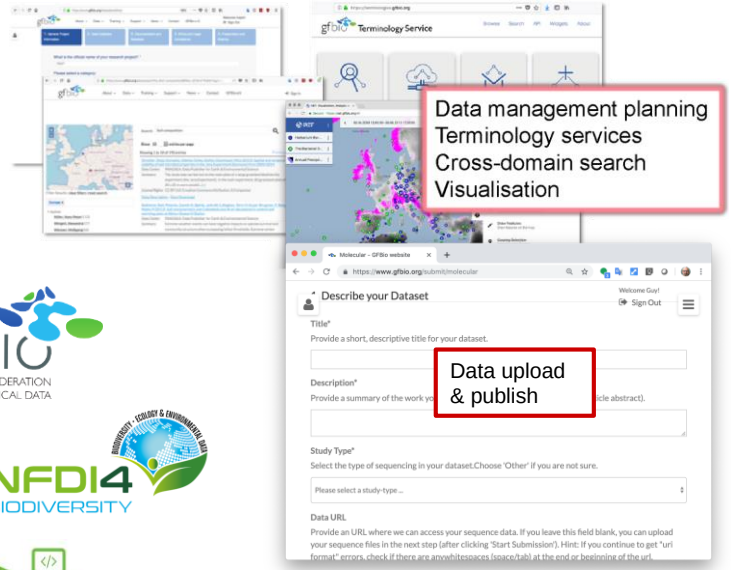


- Long history of brokering with demonstrated mutual benefits
- Can be **generalist** or have **data-type** or **domain specific** expertise

- **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



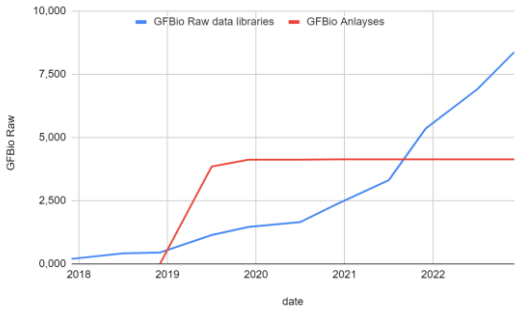
ENA
European Nucleotide Archive

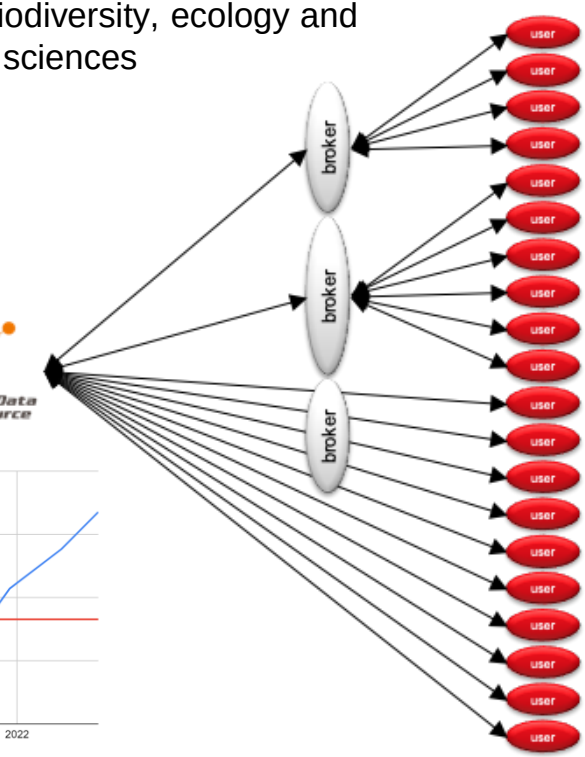
EMBL-EBI

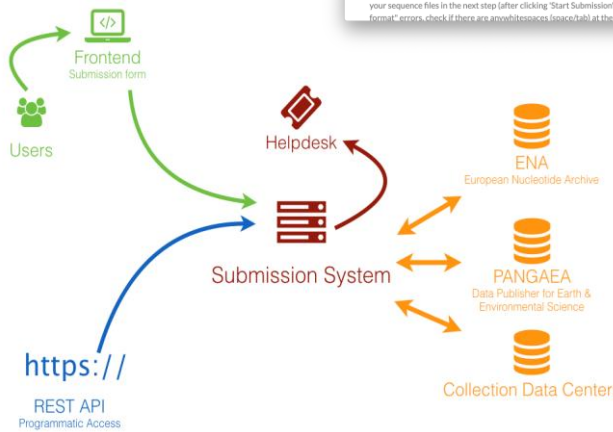
gfbio
GERMAN FEDERATION FOR BIOLOGICAL DATA

NFDI4 BIODIVERSITY

elixir Core Data Resource







EMBL-EBI

Example: Global health and SARS-CoV-2

UK Raw Data Producers

Ireland Raw Data Producers

The Netherlands Raw + Assembled Data Producers

Norway Raw Data Producers

Switzerland Raw + Assembled Data Producers

Germany Assembled Data Producers

Spain Raw Data Producers

COG-UK COVID-19 GENOMICS UK CONSORTIUM

UCD DUBLIN

teagasc AGRICULTURE AND FOOD DEVELOPMENT AUTHORITY

Erasmus MC Universitair Medisch Centrum Rotterdam

elixir NORWAY

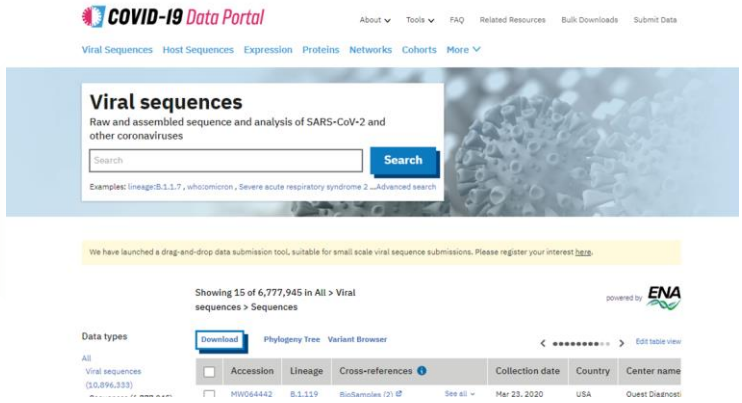
SIB Swiss Institute of Bioinformatics

ROBERT KOCH INSTITUT

SeqCOVID *

- 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)

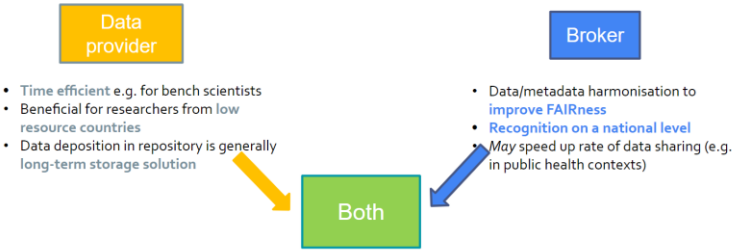
ENA



Future

Expansion of brokering network

Benefits of data brokering



Brokering data to the ENA

ENA
European Nucleotide Archive

- 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)

Some of the possible considerations of a data broker

	Aspects to consider
Data/metadata collection	• File formats • All must comply with national regulations
Data/metadata validation	• Broker-side validation • Will the broker verify whether information provided is correct?
Processing data and metadata	• Curation/masking sensitive metadata • QC of data files • Analysis of data? • Beware legal implications -> broker role changes to data processor • Or, is preference to receive data/metadata in mostly repository compliant format
Data storage	• What happens to data/metadata after submission – is there a timeframe for storage?
Helpdesk support	• In case of queries, where can users seek help – both broker + repository?
Data accreditation	• How will the data provider be credited? • What are deposition databases' policies on this?

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

BY-COVID beyond covid · elixir CONVERGE · EOSC-Life · ReCoDID · veo · EMBL-EBI

Participating in INSDC

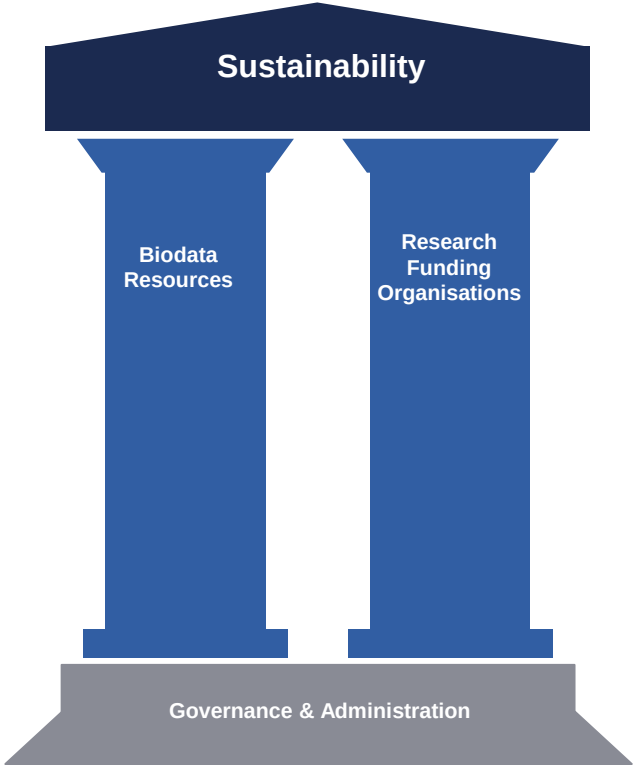
Not just ENA…

- 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/

GLOBAL BIODATA COALITION

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/

GLOBAL BIODATA COALITION

Acknowledgements

Data Coordination and Archiving (Cochrane)

Alisha Ahamed, Carla Cummins, Khadim Gueye, Vikas Gupta, Maira Ihsan, Suran Jayathilaka, Vishnukumar Balavenkataraman Kadhirvelu, 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

EMBL elixir GORDON AND BETTY MOORE FOUNDATION UK Research and Innovation wellcome Funded by the European Union

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

ENA European Nucleotide Archive

GLOBAL BIODATA COALITION