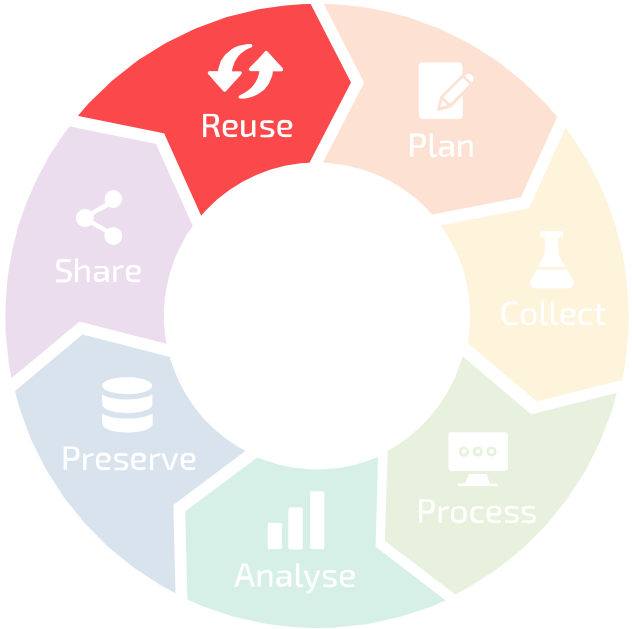
Secondary use of data and key RDM issues

Korbinian Bösl
Data management coordinator
ELIXIR Norway

elixir.no



Steps for data re-use

Motivation

Identifying and finding relevant data

Getting access to the data

- Open data
- Personal (sensitive data)
  - Controlled access through archive
  - Secondary use in EHDS

Understanding and harmonizing data

Data reuse examples

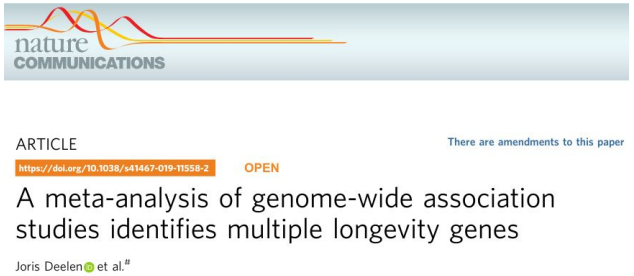

OPEN ACCESS Freely available online PLoS PATHOGENS

Review

Host Cell Factors in HIV Replication: Meta-Analysis of Genome-Wide Studies

Frederic D. Bushman[1]*, Nirav Malani[1], Jason Fernandes[2,3], Iván D'Orso[2,3], Gerard Cagney[3,4,5], Tracy L. Diamond[6], Honglin Zhou[6], Daria J. Hazuda[6], Amy S. Espeseth[6], Renate König[7], Sourav Bandyopadhyay[8], Trey Ideker[8], Stephen P. Goff[9], Nevan J. Krogan[3,4], Alan D. Frankel[2,3], John A. T. Young[10], Sumit K. Chanda[7]*

https://doi.org/10.1371/journal.ppat.1000437


nature COMMUNICATIONS

ARTICLE There are amendments to this paper

https://doi.org/10.1038/s41467-019-11558-2 OPEN

A meta-analysis of genome-wide association studies identifies multiple longevity genes

Joris Deelen et al.#

https://doi.org/10.1038/s41467-019-11558-2


European Journal of Public Health, Vol. 27, Supplement 1, 2017, 90–95

© The Author 2017. Published by Oxford University Press on behalf of the European Public Health Association. All rights reserved.

doi:10.1093/eurpub/ckw229

Informal care in Europe: findings from the European Social Survey (2014) special module on the social determinants of health

Ellen Verbakel[1], Stian Tamlagsrønning[2], Lizzy Winstone[3], Erlend L. Fjær[2], Terje A. Eikemo[2]

https://doi.org/10.1093/eurpub/ckw229

https://researchparasite.com/

Motivations for data re-use

Many datasets contain information that was not/never followed up

Allows to apply new questions/angles to a published dataset

Allows researchers to work with data they would not have the Expertise/infrastructure/resources to produce themselves

Allows to integrate data from different studies, labs, disciplines,...

Increase statistical power

Saving costs, time and effort

Data citation

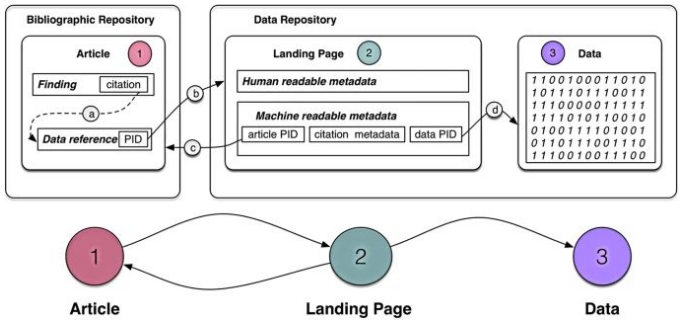
Principles: Attribution & Access

Joint Declaration of Data Citation Principles (JDDCP)

Creative Commons: TASL – Title, Author, Source, License

Many archives contain information how a dataset should be cited

Include persistent ID whenever possible



Identifying relevant datasets

1. Data underlying a scientific articles
2. (Scientific) Data in a relevant community archives
3. Dataset metasearch across archives
4. Data from public authorities

Data availability statement

CellPress

Cell

Article

Genetic Screens Identify Host Factors for SARS-CoV-2 and Common Cold Coronaviruses

Ruofan Wang,[1,13] Camille R. Simoneau,[2,3,4,5,13] Jessie Kulsuptrakul,[1] Mehdi Bouhaddou,[2,4,6,7] Katherine A. Travisano,[1] Jennifer M. Hayashi,[2,3,4] Jared Carlson-Stevermer,[8] James R. Zengel,[9] Christopher M. Richards,[9] Parinaz Fozouni,[2,3,4,5,10] Jennifer Oki,[8] Lauren Rodriguez,[11] Bastian Joehnk,[12] Keith Walcott,[12] Kevin Holden,[8] Anita Sil,[12] Jan E. Carette,[9] Nevan J. Krogan,[2,4,6,7] Melanie Ott,[2,3,4,*] and Andreas S. Puschnik[1,14,*]

STAR★Methods

Deposited Data

Raw sequencing data for CRISPR KO screens	EMBL-EBI ArrayExpress	E-MTAB-9638

ArrayExpress Functional genomics data

BIOSTUDIES / ARRAYEXPRESS / E-MTAB-9638

Release Date: 16 October 2020 • Modified: 9 March 2022 • Views: 184

[Cite] {JSON} .PageTab HTTP FTP Globus

Functional genomic screens in human cells to identify host factors for SARS-CoV-2 and common cold coronaviruses

Data underlying a scientific articles

Data on ‘reasonable request’

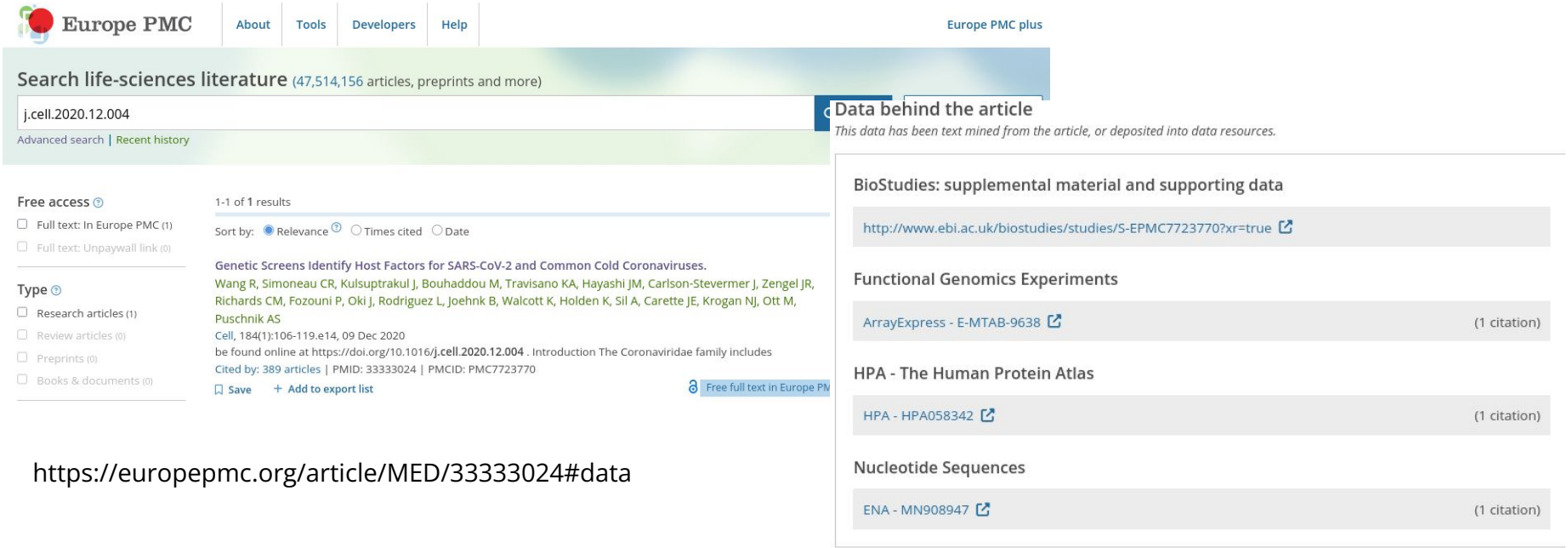
Supplemental data

Deposited data

1. Community repositories
2. Institutional repositories
3. Multidisciplinary repositories

Data underlying a scientific articles

Some literature indices link directly to the data



https://europepmc.org/article/MED/33333024#data

Data journals

new format of publication that focuses on describing a dataset, instead of/in addition to conventional research articles

scientific **data**

(GIGA)^n SCIENCE

Data in community archives

Data in a relevant community archive

Advantage: uniform format & metadata schemes, bulk queries, discipline-specific standards, sometimes curated

Where does your field publishing data?

Curated registries can help to identify data repositories:

re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES

re3data.org

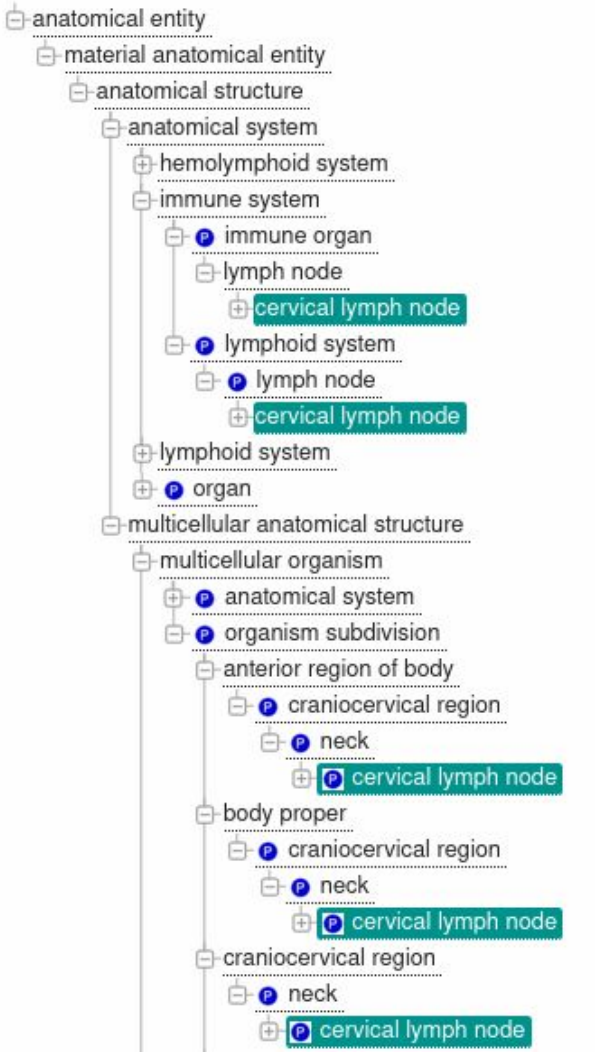
FAIRsharing.org
standards, databases, policies

fairsharing.org

Ontology example

Uberon

Ontologies
enable hierarchical searches
explicit relations
allow to make connections
synonyms (often)



Value added data & Reference data

Heavily curated and often versioned e.g.:

- Reference genomes & annotations
- Organism proteomes
- Literature extracted information
- Secondary databases
- Knowledge graphs & bases

ensembl.org

uniprot.org

Often standard (pre-)processing

Dataindicies und portals

Data metasearch engines

- Searching across disciplines
- Data in institutional archives
- Data in multidisciplinary archives

Metadata quality is critical!

Repository coverage varies

Persistent identifiers: datasets with DOI are easiest to find

Apply systematic filtering to narrow down or expand results

Dataindicies und portals

Non-commercial & generic

- DataCite
- BASE
- OpenAIRE

Commercial

- Google Dataset Search
- Mendeley Data
- WOS Data Citation Index

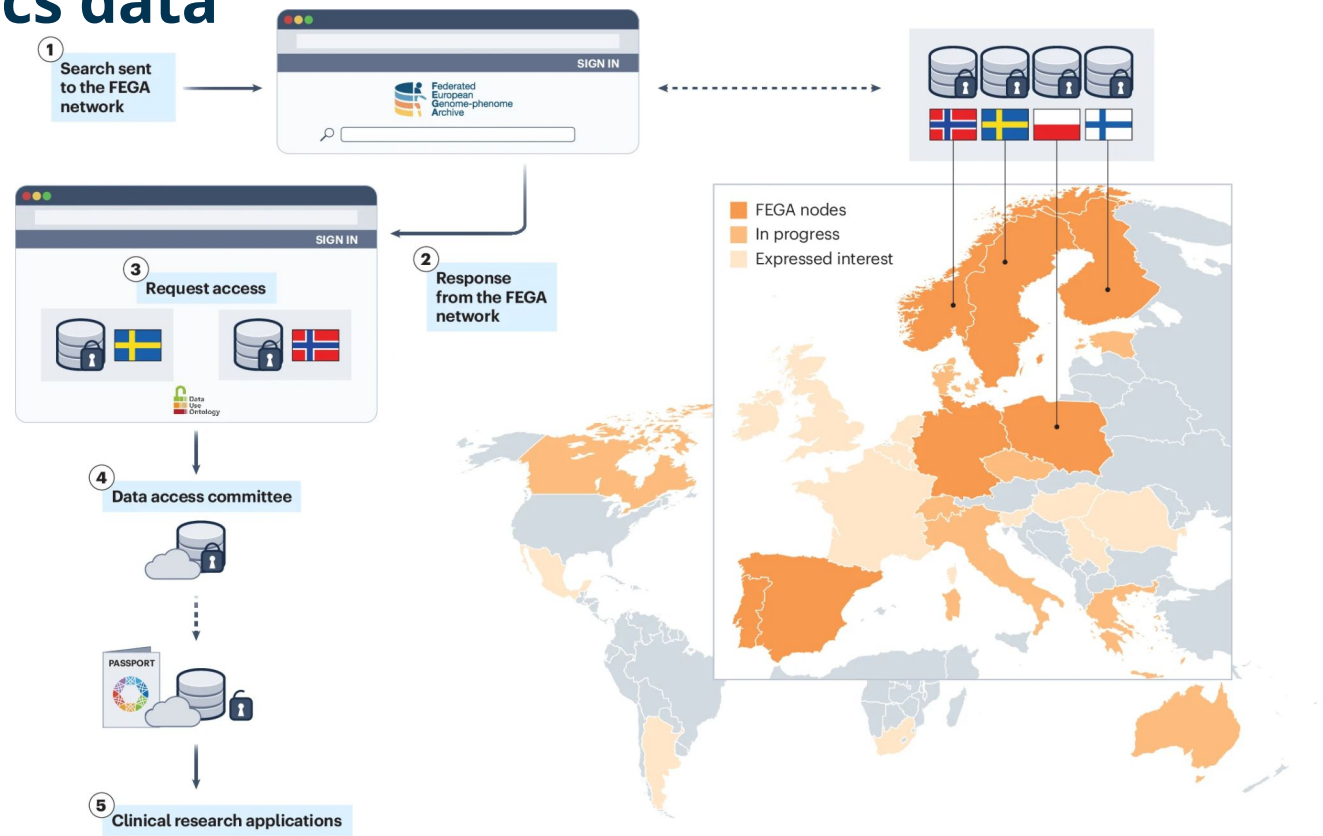
Discipline specific

- EBISearch
- OmicsDI
- Specific data portals e.g. Pathogensportal
- Biobanks - BBMRI directory

Public sector

- National Health data/-registry portals
- Other registries
- WHO data collections
- data.europe.eu

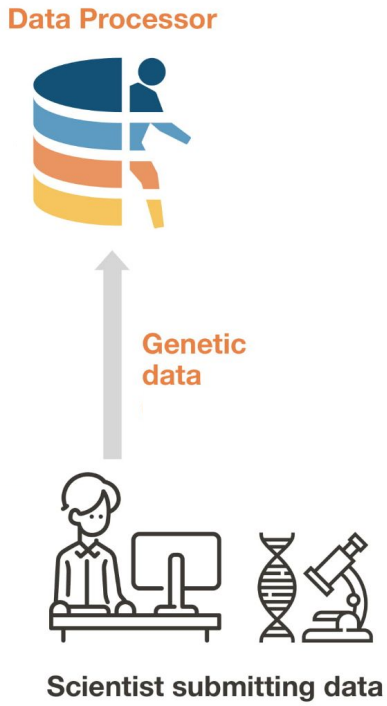
Federated EGA as a global network of sharing human genomics data



D'Altri, T., Freeberg, M.A., Curwin, A.J. *et al.* The Federated European Genome–Phenome Archive as a global network for sharing human genomics data. *Nat Genet* (2025). https://doi.org/10.1038/s41588-025-02101-9 link

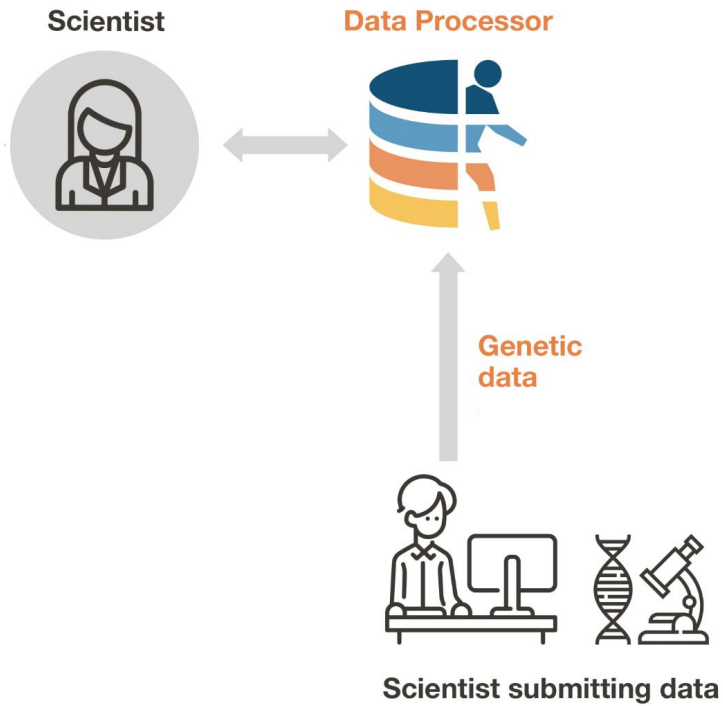
EBI blog on FEGA paper

Controlled data access



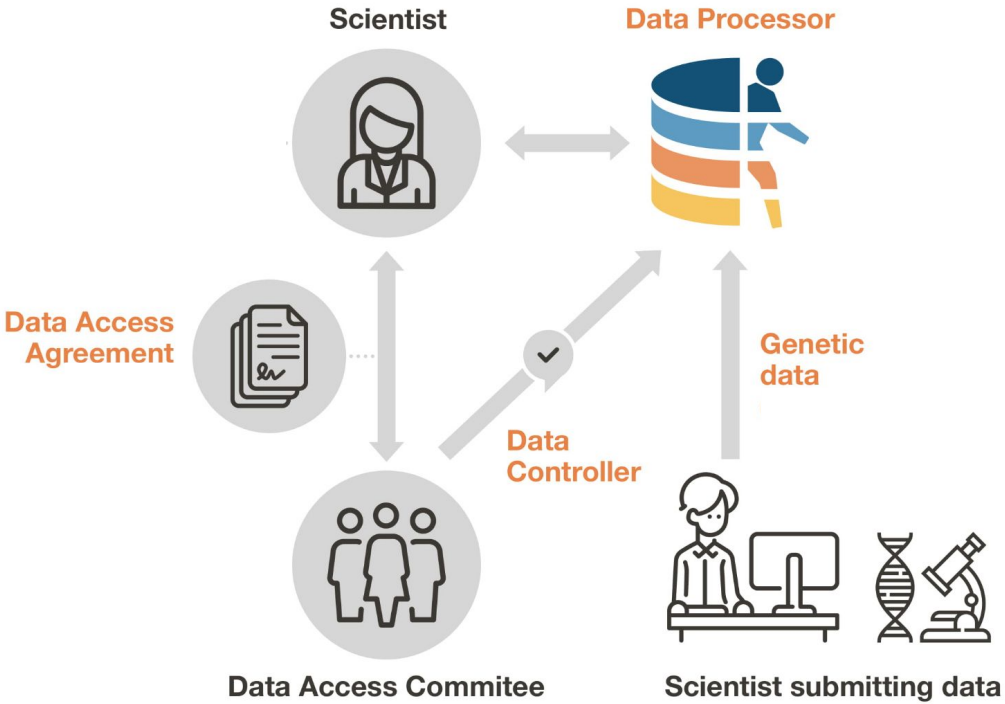
https://ega-archive.org/

Controlled data access



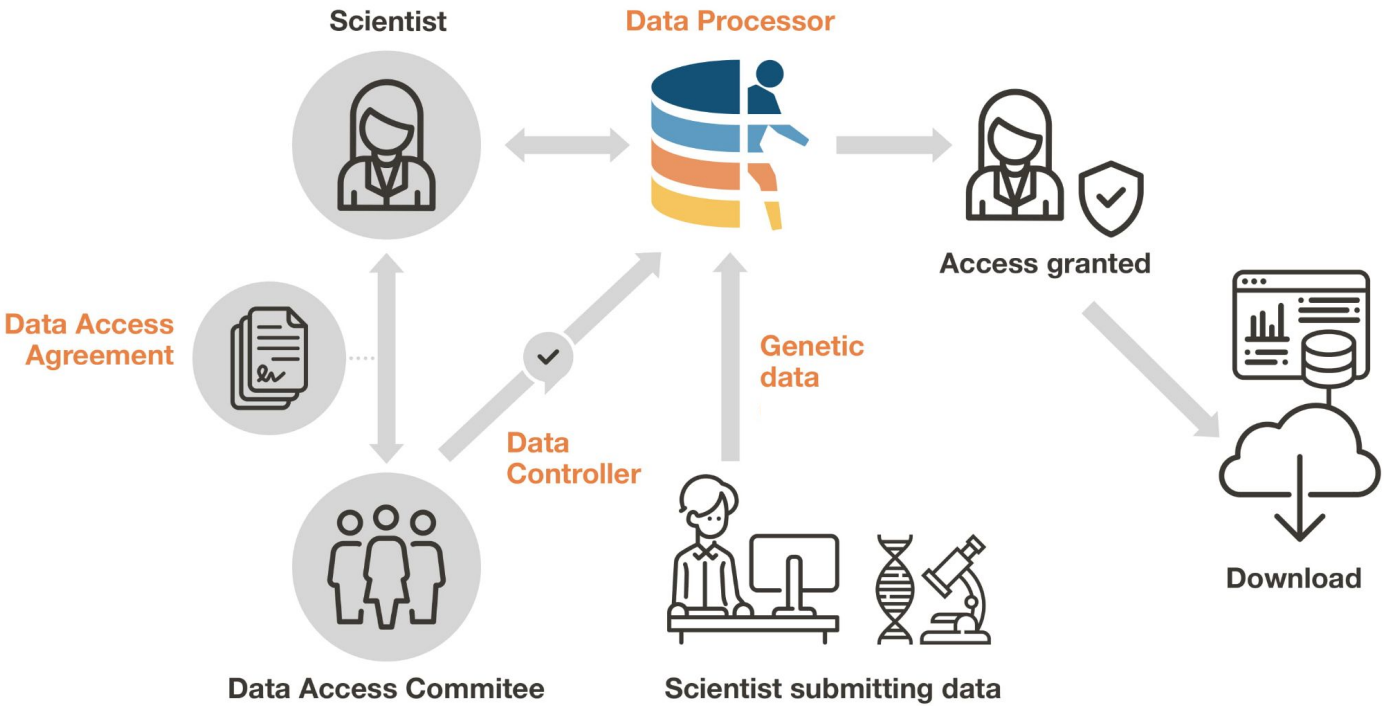
https://ega-archive.org/

Controlled data access



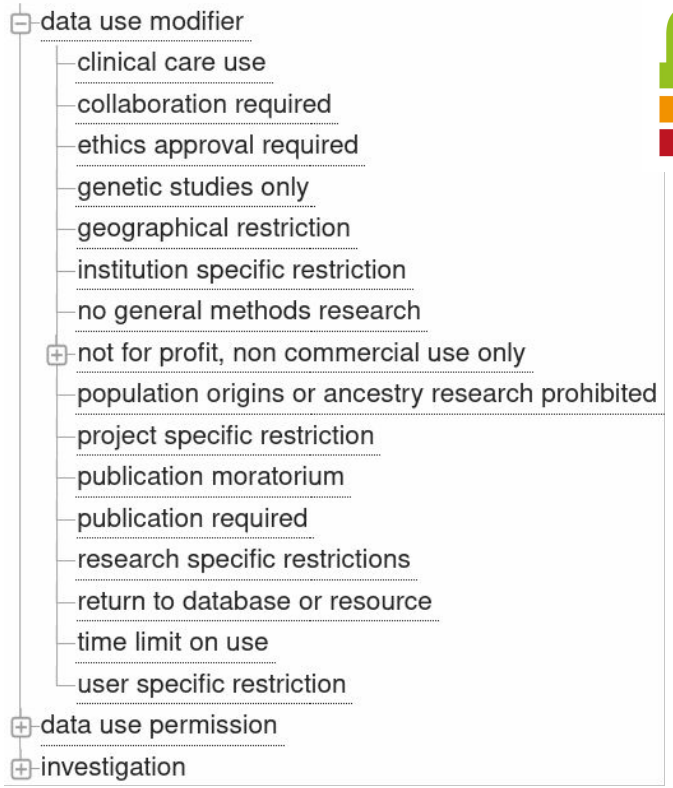
https://ega-archive.org/

Controlled data access



https://ega-archive.org/

Modelling consents - Data Use Ontology (DUO)



Data Use Ontology

Global Alliance for Genomics & Health

Collaborate. Innovate. Accelerate.

DUO:0000021 ethics approval required

DUO:0000022 geographical restriction

https://github.com/EBISPOT/DUO

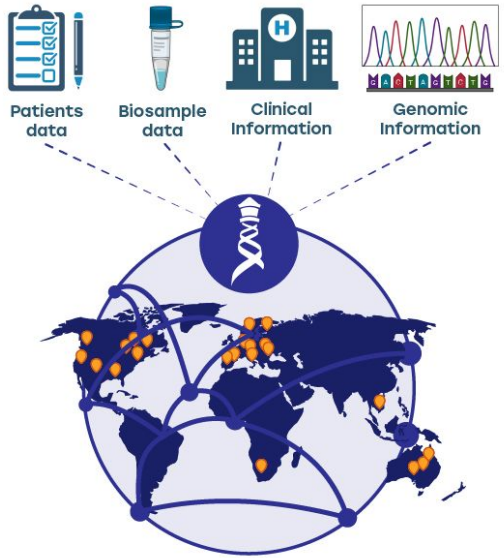
Genome beacon

1. Create a query

Check for the presence of genomic variants, or look for detailed information through structured queries.

2. Beacon API search

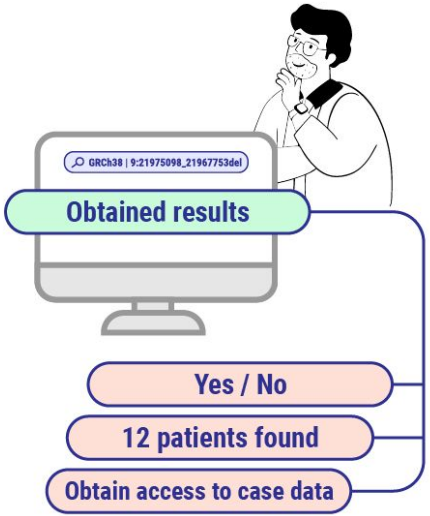
Beacon looks for the response, accommodating a wide range of data types, such as genomic and clinical data.



3. Get the response

Get the results to the query:

- Boolean (yes/no).
- Count.
- Case-level data.



Secondary use of health data in EHDS

Regulated in CHAPTER IV Article 50-65

Entered into force March 2025

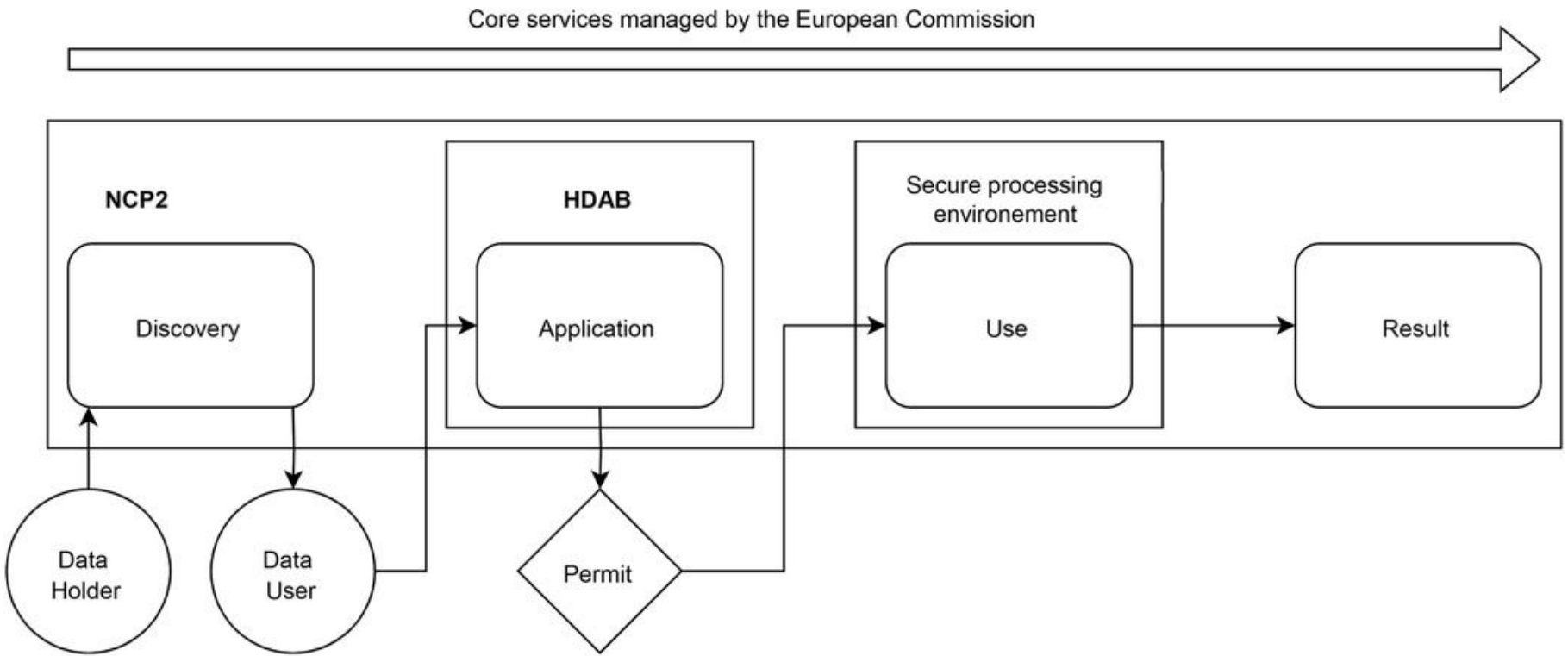
Current national implementation status varying

Compulsory from March 2029

Genetic and *omics data (Article 51 f,g) included from 2031

https://health.ec.europa.eu/ehealth-digital-health-and-care/european-health-data-space-regulation-ehds_en

Secondary use of health data in EHDS



van Drumpt S, Chawla K, Barbereau T, Spagnuelo D and van de Burgwal L (2025) Secondary use under the European Health Data Space: setting the scene and towards a research agenda on privacy-enhancing technologies. Front. Digit. Health 7:1602101. doi.org/10.3389/fdgth.2025.1602101 under CC-BY 4.0

Data harmonisation

Do you have to convert between data formats?

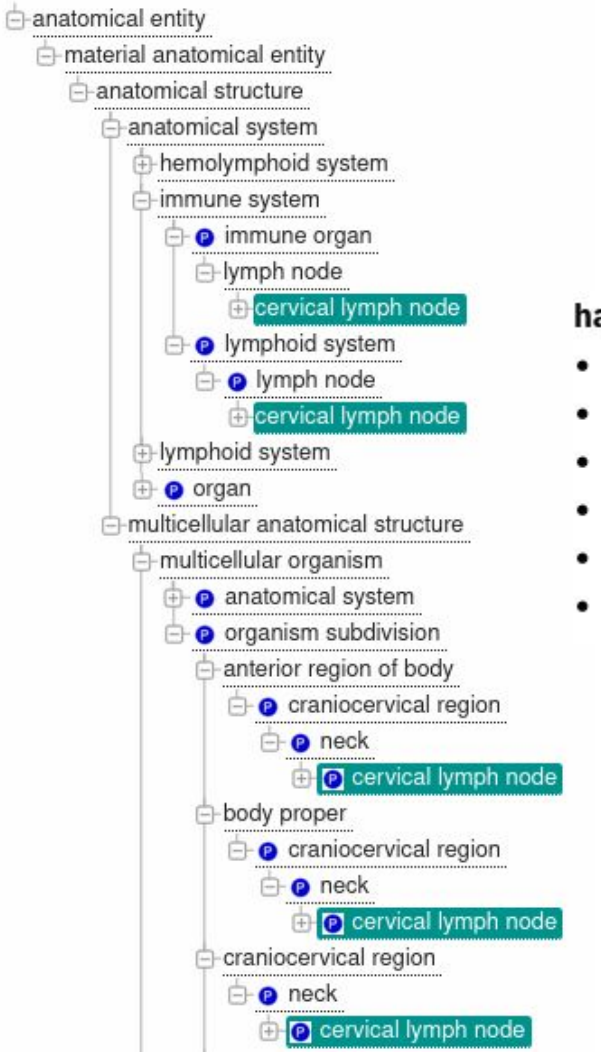
Do different datasets use different (e.g. gene-) annotations?

Do you have to re-process raw data?

Do you want to combine datasets based on ontologies?

Data harmonisation

Uberon



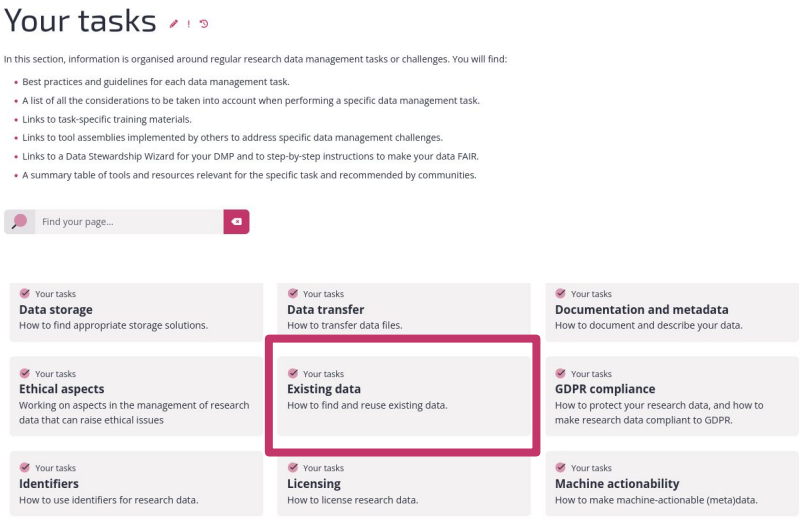
has_dbxref

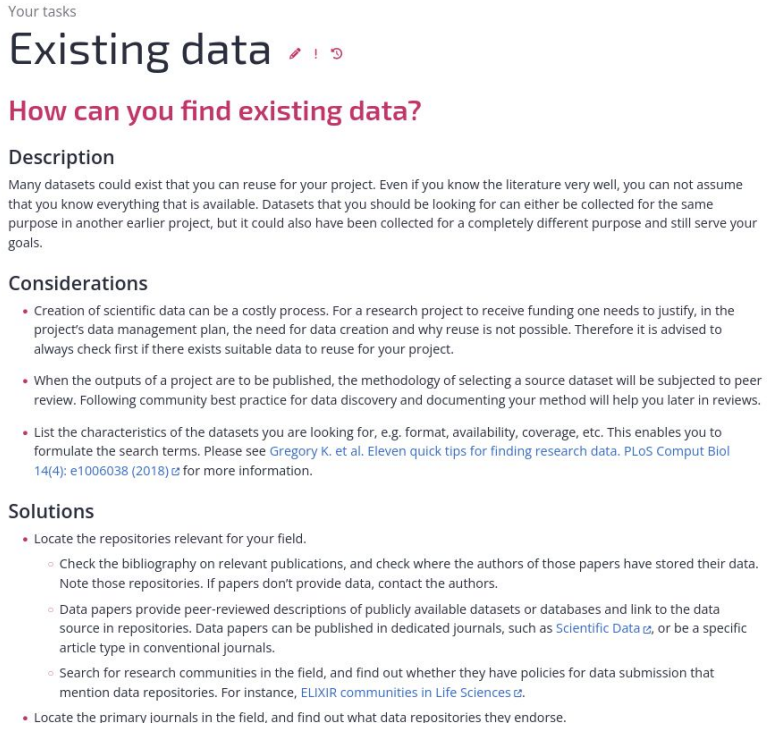
- Lymph node of neck FMA
- cervical lymph node MA
- Cervical Lymph Node NCIT
- SCTID:181757009
- UMLS:C0588054
- Wikipedia:Cervical_lymph_nodes

Check list

- ❑ How will you be affected if original source is not longer available?
- ❑ Do you have to keep a copy?
- ❑ Does the license allow (re-)use and/or combination of dataset(s)?
- ❑ Do you have a sufficient legal basis?
- ❑ Is the data (co-)controller situation clarified? DPIA?
- ❑ Is your (re-)use in line with the original consent?
- ❑ How will you deal with consent withdrawal?
- ❑ Can you use the data as is or do you have to re-format/harmonise?
- ❑ Is the (meta-)data quality sufficient?
- ❑ How will you ensure reproducibility of harmonisation/re-processing?
- ❑ Do you have resources for (secure) storage and processing?
- ❑ Can the data be made available for other re-users?

RDMkit





https://rdmkit.elixir-europe.org/existing_data