



# Secondary use of data and key RDM issues

Korbinian Bösl

Data management coordinator

ELIXIR Norway



[elixir.no](http://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<sup>1\*</sup>, Nirav Malani<sup>1</sup>, Jason Fernandes<sup>2,3</sup>, Iván D'Orso<sup>2,3</sup>, Gerard Cagney<sup>3,4,5</sup>, Tracy L. Diamond<sup>6</sup>, Honglin Zhou<sup>6</sup>, Daria J. Hazuda<sup>6</sup>, Amy S. Espeseth<sup>6</sup>, Renate König<sup>7</sup>, Sourav Bandyopadhyay<sup>8</sup>, Trey Ideker<sup>9</sup>, Stephen P. Goff<sup>9</sup>, Nevan J. Krogan<sup>3,4</sup>, Alan D. Frankel<sup>2,3</sup>, John A. T. Young<sup>10</sup>, Sumit K. Chanda<sup>7\*</sup>

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

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<sup>1</sup>, Stian Tambsrønning<sup>2</sup>, Lizzy Winstone<sup>3</sup>, Erlend L. Fjær<sup>2</sup>, Terje A. Eikemo<sup>2</sup>

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



<https://researchparasite.com/>

  
nature  
COMMUNICATIONS

ARTICLE

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

OPEN

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

Joris Deelen<sup>1</sup> et al.<sup>2\*</sup>

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

# 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

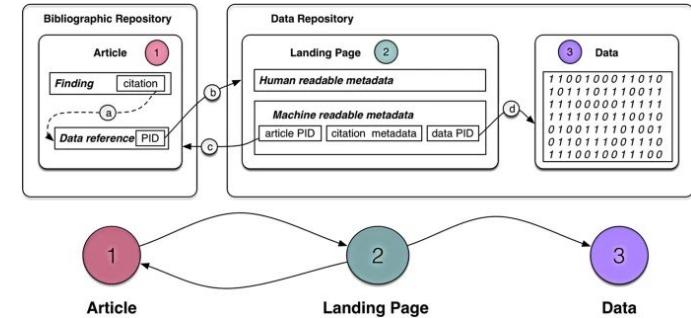
Joint Declaration of Data Citation Principles (JDDCP)

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

STAR★Methods

## Deposited Data

Raw sequencing data for CRISPR KO screens

EMBL-EBI  
ArrayExpress

E-MTAB-9638



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

The screenshot shows the Europe PMC website interface. At the top, there is a navigation bar with links for About, Tools, Developers, and Help. To the right, it says "Europe PMC plus". Below the navigation bar, a search bar displays "Search life-sciences literature (47,514,156 articles, preprints and more)". A specific article record is shown for "j.cell.2020.12.004". The article title is "Genetic Screens Identify Host Factors for SARS-CoV-2 and Common Cold Coronaviruses." by Wang R, Simoneau CR, Kulsupratkul J, Bouhaddou M, Travisano KA, Hayashi JM, Carlson-Stevermer J, Zengel JR, Richards CM, Fozouni P, Oki J, Rodriguez L, Joehnk B, Walcott K, Holden K, Sil A, Carette JE, Krogan NJ, Ott M, Puschnik AS. It was published in Cell, 184(1):106-119.e14, 09 Dec 2020. The record includes a "Free full text in Europe PMC" link. On the right side of the page, there is a sidebar titled "Data behind the article" which lists several data resources extracted from the article:

- BioStudies: supplemental material and supporting data**  
http://www.ebi.ac.uk/biostudies/studies/S-EPMC7723770?xr=true
- Functional Genomics Experiments**  
ArrayExpress - E-MTAB-9638 (1 citation)
- HPA - The Human Protein Atlas**  
HPA - HPA058342 (1 citation)
- Nucleotide Sequences**  
ENA - MN908947 (1 citation)

The URL of the page is 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)<sup>n</sup>  
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](https://re3data.org)

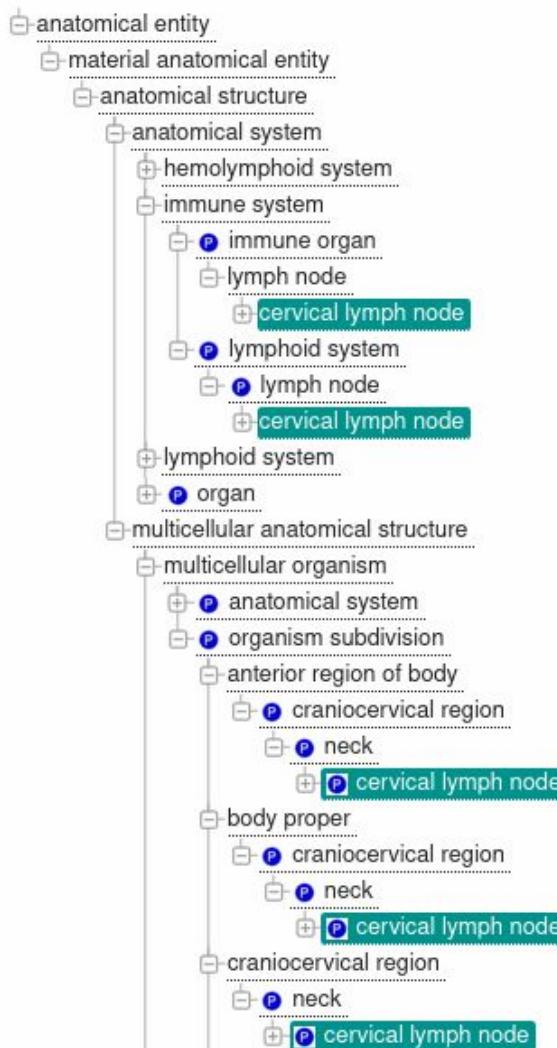


[fairsharing.org](https://fairsharing.org)

# Ontology example

# . Ü 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

Often standard (pre-)processing



[ensembl.org](http://ensembl.org)



[uniprot.org](http://uniprot.org)

# Dataindices 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



# Dataindices und portals

Non-commercial & generic

DataCite



BASE



OpenAIRE



Commercial

Google Dataset Search



Mendeley Data



WOS Data Citation Index



Discipline specific

EBSIearch



OmicsDI



Specific data portals e.g.

Pathogensportal



Biobanks - BBMRI directory



Public sector

National Health data/-registry portals

Other registries

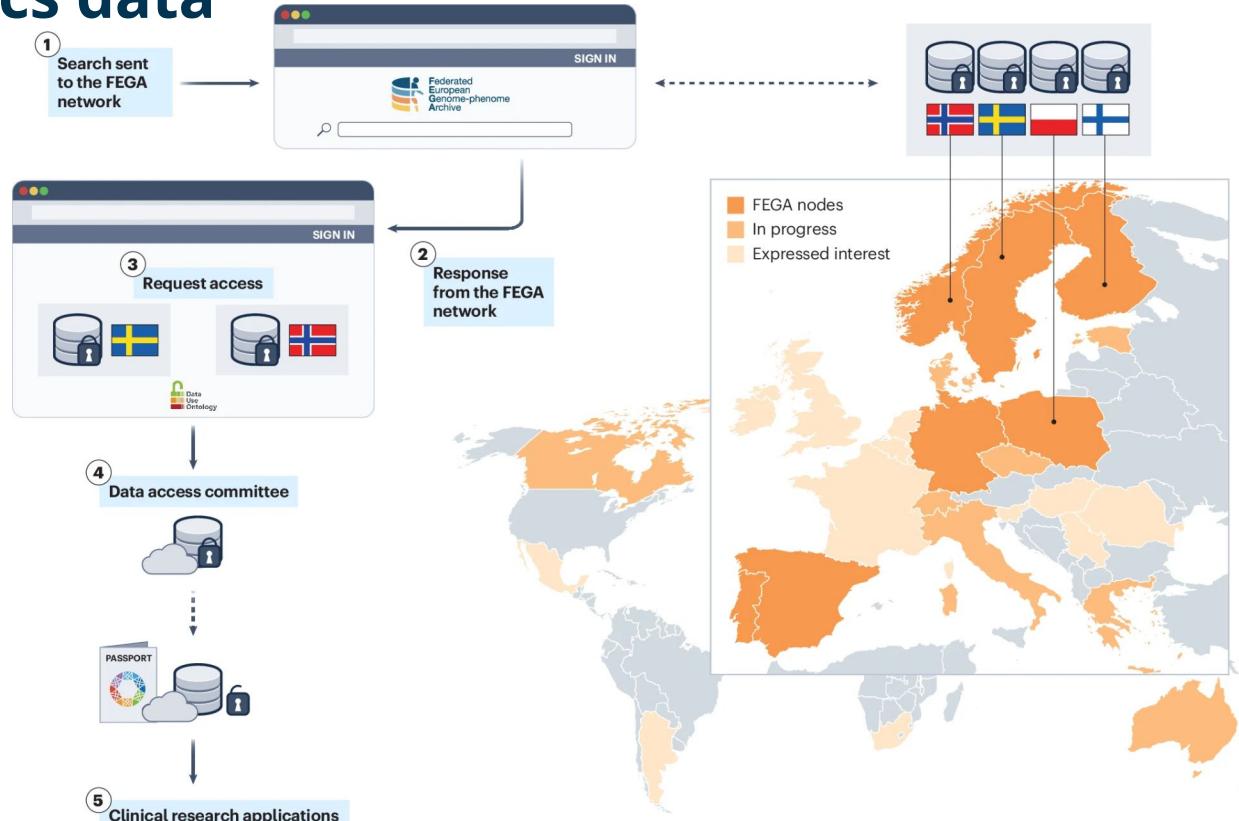


WHO data collections

data.europe.eu



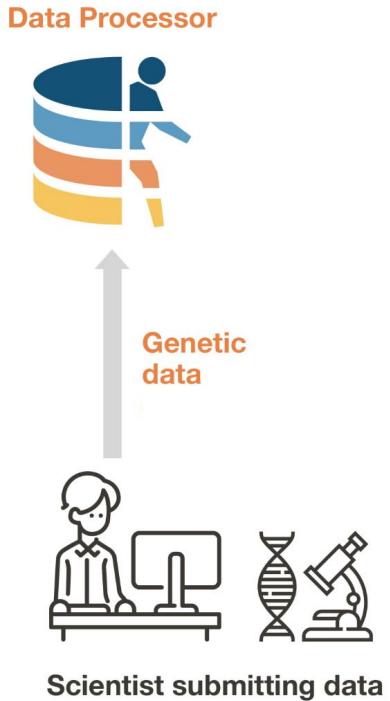
# Federated EGA as a global network of sharing human genomics data



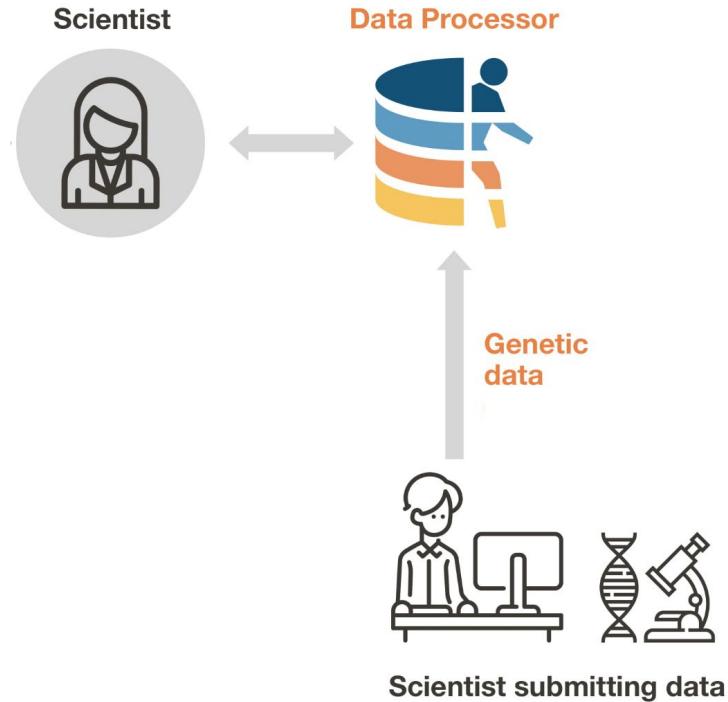
D'Altri, T., Freeberg, M.A., Curwin, A.J. et al. The Federated European Genome–Phenome Archive as a global network for sharing human genomics data. *Nat Genet* (2025).  
<https://doi.org/10.1038/s41588-025-02101-9> [link](#)

[EBI blog on FEGA paper](#)

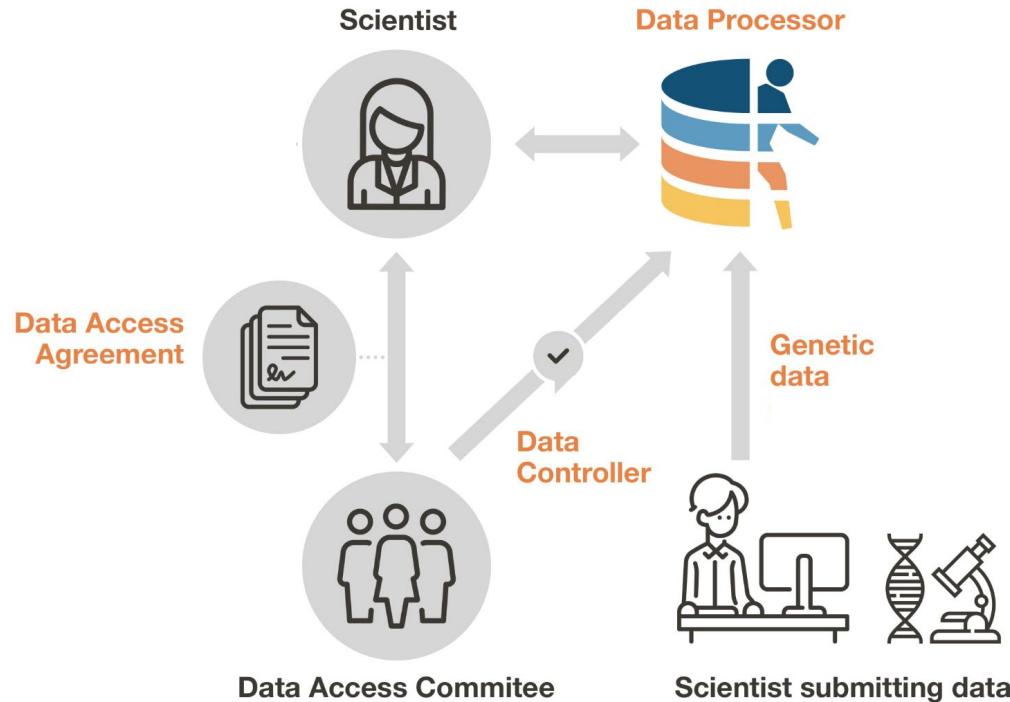
# Controlled data access



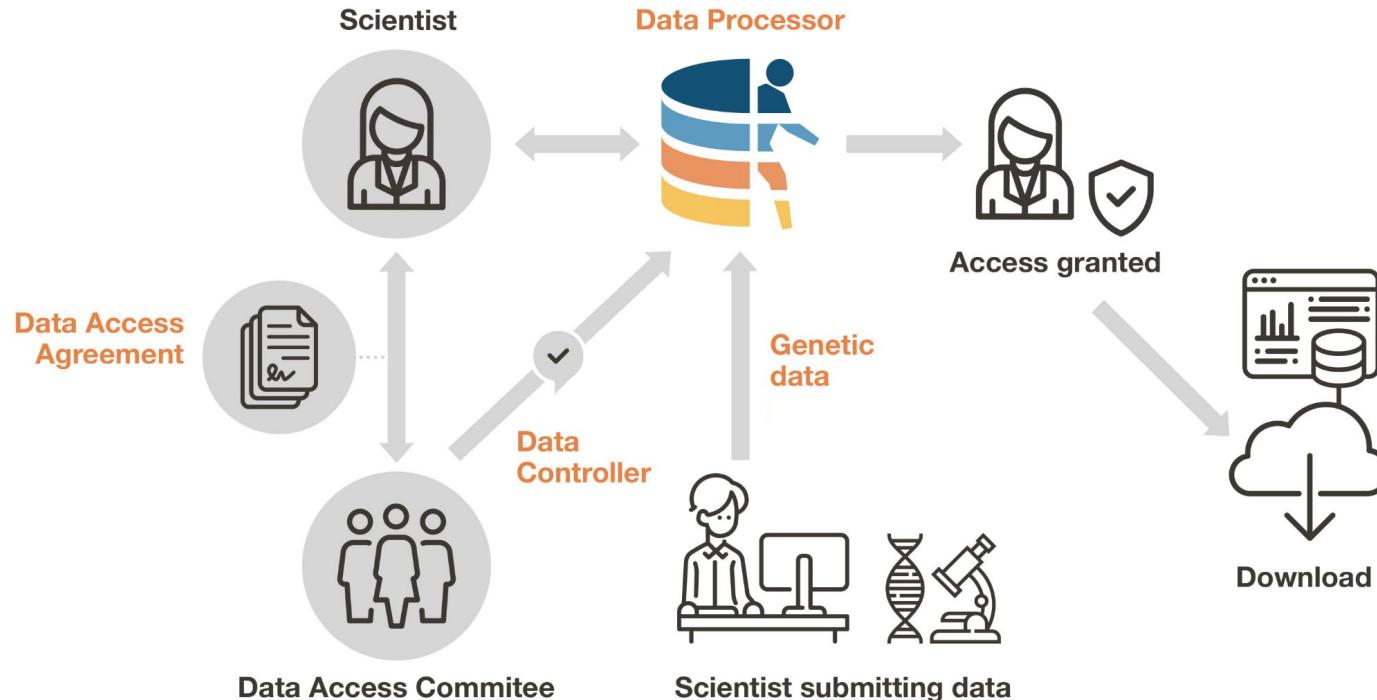
# Controlled data access



# Controlled data access



# Controlled data access



# Modelling consents - Data Use Ontology (DUO)

- data use modifier
  - clincial care use
  - collaboration required
  - ethics approval required
  - genetic studies only
  - geographical restriction
  - institution specific restriction
  - no general methods research
- + not for profit, non commercial use only
  - population origins or ancestry research prohibited
  - project specific restriction
  - publication moratorium
  - publication required
  - research specific restrictions
  - return to database or resource
  - time limit on use
  - user specific restriction
- + data use permission
- + investigation



**Global Alliance**  
for Genomics & Health  
*Collaborate. Innovate. Accelerate.*

DUO:ooooo21 ethics approval required

DUO:ooooo22 geographical restriction

# Genome beacon

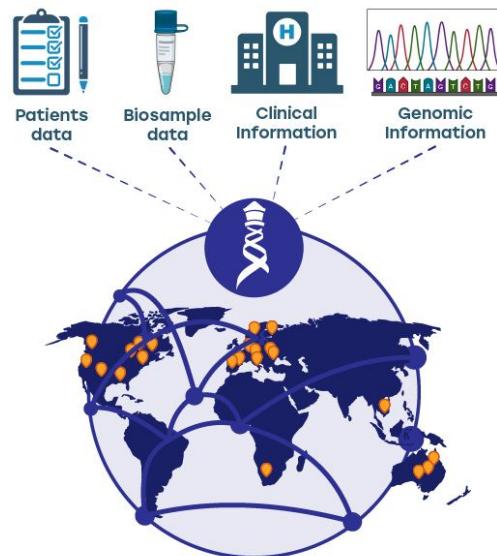
## 1. Create a query

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



## 2. Beacon API search

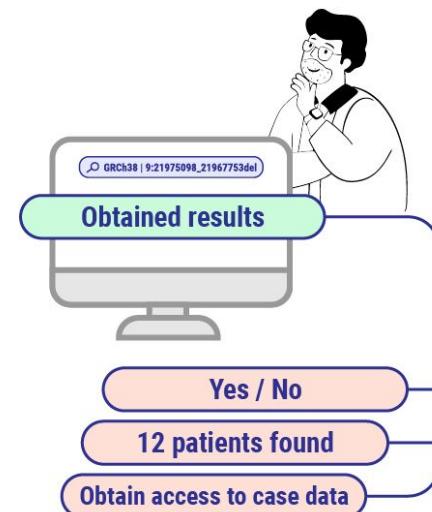
Beacon looks for the response, accommodating a wide range of data types, such as genomic and clinical data.



## 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

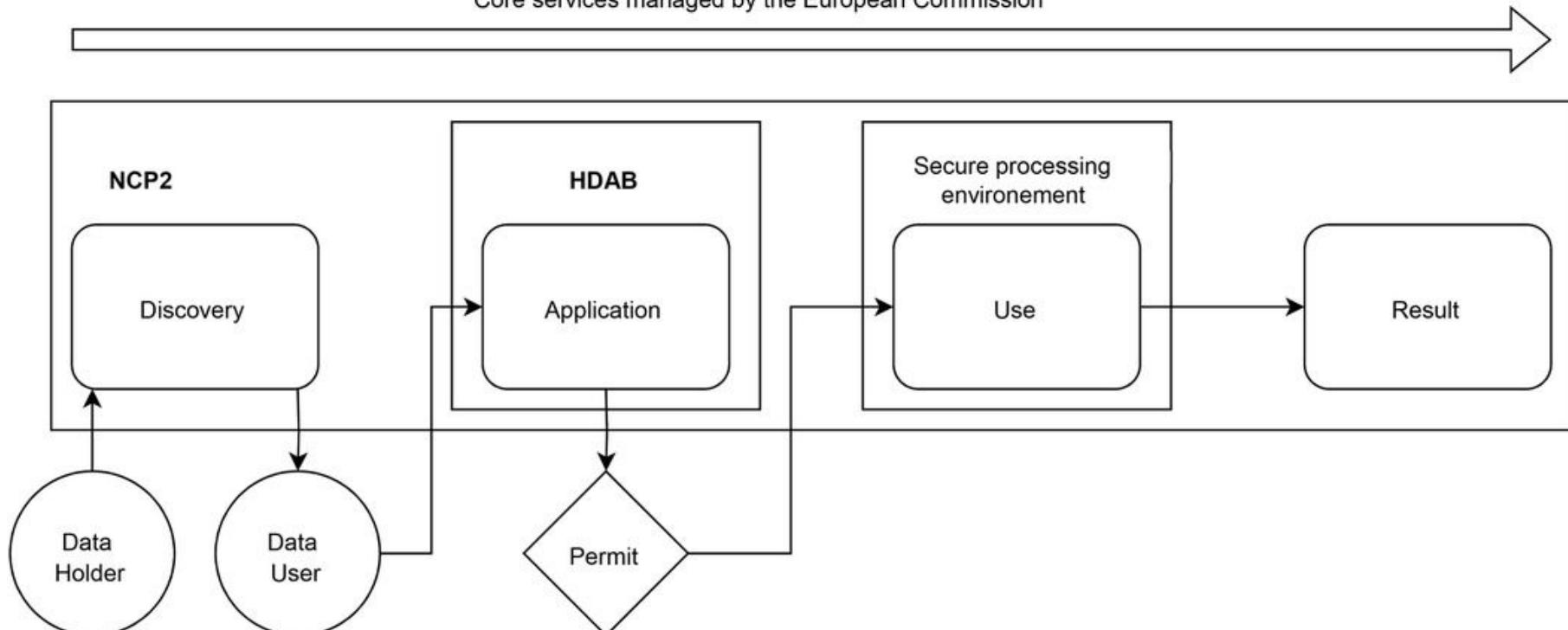
Current national implementation status varying

Compulsory from March 2029

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

# Secondary use of health data in EHDS

Core services managed by the European Commission



## Data harmonisation

Do you have to convert between data formats?

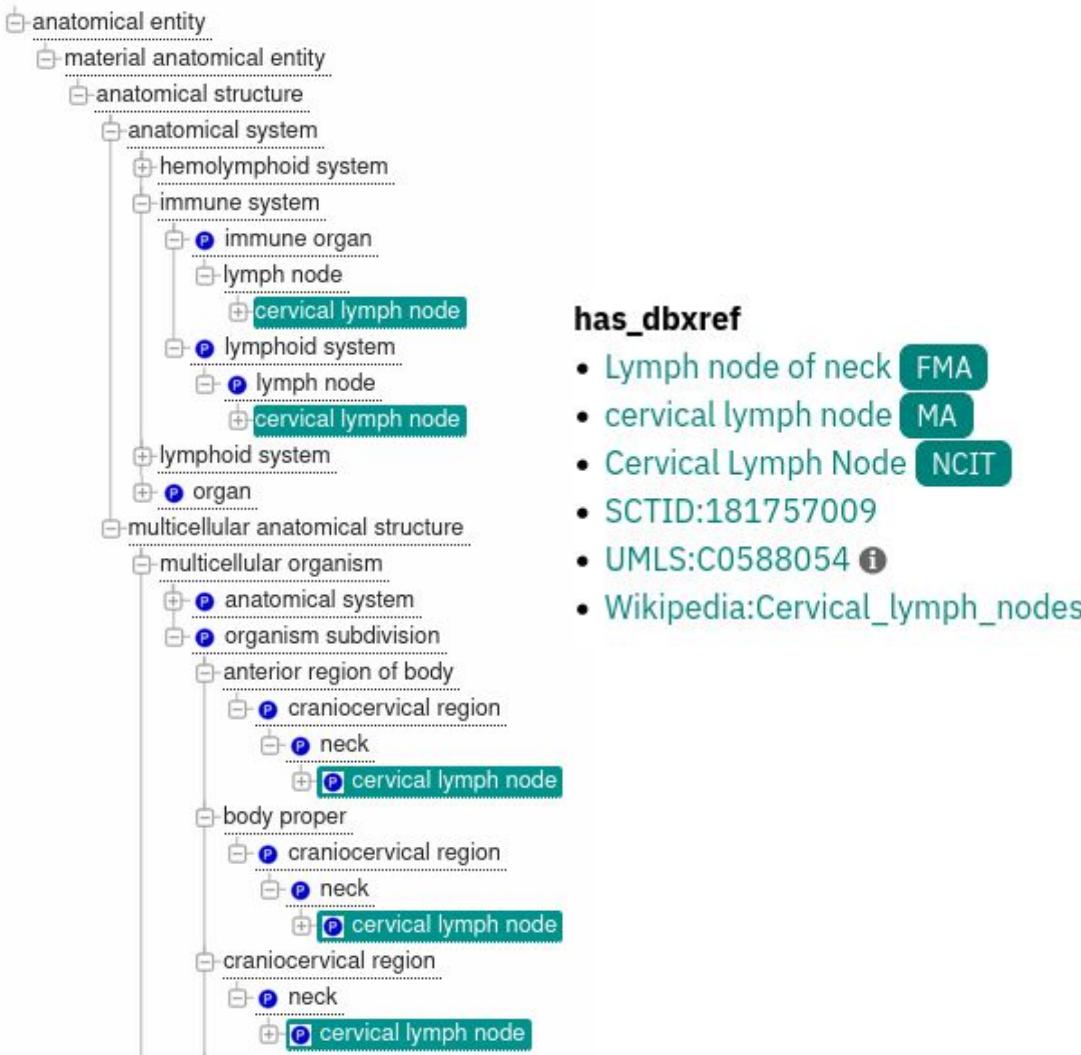
Do different datasets use different (e.g. gene-) annotations?

Do you have to re-process raw data?

Do you want to combine datasets based on ontologies?

# Data harmonisation

## . ü Uberon



## 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?



## Your tasks

In this section, information is organised around regular research data management tasks or challenges. You will find:

- Best practices and guidelines for each data management task.
- A list of all the considerations to be taken into account when performing a specific data management task.
- Links to task-specific training materials.
- Links to tool assemblies implemented by others to address specific data management challenges.
- Links to a Data Stewardship Wizard for your DMP and to step-by-step instructions to make your data FAIR.
- A summary table of tools and resources relevant for the specific task and recommended by communities.

Find your page...

Your tasks  
**Data storage**  
How to find appropriate storage solutions.

Your tasks  
**Data transfer**  
How to transfer data files.

Your tasks  
**Documentation and metadata**  
How to document and describe your data.

Your tasks  
**Ethical aspects**  
Working on aspects in the management of research data that can raise ethical issues.

Your tasks  
**Existing data**  
How to find and reuse existing data.

Your tasks  
**GDPR compliance**  
How to protect your research data, and how to make research data compliant to GDPR.

Your tasks  
**Identifiers**  
How to use identifiers for research data.

Your tasks  
**Licensing**  
How to license research data.

Your tasks  
**Data management**

- Data life cycle
- Your role
- Your domain
- Your tasks
  - Compliance monitoring
  - Costs of data management
  - Creating a data-flow diagram
  - Data analysis
  - Data brokering
  - Data deletion
  - Data discoverability
  - Data interlinking
  - Data management coordination
  - Data management plan
  - Data organisation
  - Data security
  - Data sensitivity
  - Data provenance
  - Data publication
  - Data quality
  - Data storage
  - Data transfer
  - Documentation and metadata
  - Ethical aspects

Your tasks

## Existing data

### How can you find existing data?

#### Description

Many datasets could exist that you can reuse for your project. Even if you know the literature very well, you can not assume that you know everything that is available. Datasets that you should be looking for can either be collected for the same purpose in another earlier project, but it could also have been collected for a completely different purpose and still serve your goals.

#### Considerations

- Creation of scientific data can be a costly process. For a research project to receive funding one needs to justify, in the project's data management plan, the need for data creation and why reuse is not possible. Therefore it is advised to always check first if there exists suitable data to reuse for your project.
- When the outputs of a project are to be published, the methodology of selecting a source dataset will be subjected to peer review. Following community best practice for data discovery and documenting your method will help you later in reviews.
- List the characteristics of the datasets you are looking for, e.g. format, availability, coverage, etc. This enables you to formulate the search terms. Please see [Gregory K. et al. Eleven quick tips for finding research data. PLoS Comput Biol 14\(4\): e1006038 \(2018\)](#) for more information.

#### Solutions

- Locate the repositories relevant for your field.
  - Check the bibliography on relevant publications, and check where the authors of those papers have stored their data. Note those repositories. If papers don't provide data, contact the authors.
  - Data papers provide peer-reviewed descriptions of publicly available datasets or databases and link to the data source in repositories. Data papers can be published in dedicated journals, such as [Scientific Data](#), or be a specific article type in conventional journals.
  - Search for research communities in the field, and find out whether they have policies for data submission that mention data repositories. For instance, [ELIXIR communities in Life Sciences](#).
- Locate the primary journals in the field, and find out what data repositories they endorse.

[https://rdmkit.elixir-europe.org/existing\\_data](https://rdmkit.elixir-europe.org/existing_data)



Except where otherwise noted, this work is licensed under:

<https://creativecommons.org/licenses/by/4.0/>