



Project acronym: EOSC4CANCER
Grant Agreement Number: 101058427
Project full title: EOSC4CANCER
Call identifier: HORIZON-INFRA-2021-EOSC-01

D1.1 Data resources map

Towards an integrated cancer data catalogue network for europe and beyond

Version:	1.0
Status:	Final
Dissemination Level:	Public
Deliverable Type:	DEC – Websites, patent filings, videos, etc
Due date of deliverable:	30.11.2023
Actual submission date:	30.11.2023
Work Package:	WP 1 Federated data spaces to enable accessing, using, reusing and sharing cancer-related data
Lead partner for this deliverable:	UMCG
Partner(s) contributing:	UMCG

Main author(s):

Morris Swertz	UMCG
Gerieke Been	UMCG
Brenda Hijmans	UMCG

Other author(s)

Joeri van der Velde	UMCG
Ype Zijlstra	UMCG
Eleanor Hyde	UMCG
Marije van der Geest	UMCG



Funded by
the European Union

Revision History

Version	Date	Changes made	Author(s)
0.1	18.10.2023	Initial version	G.Been et al
0.8	9.11.2023	Final version before revisions	G.Been et al
1.0	30.11.2023	Final version	M.A. Swertz et al

Statement of originality

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

TABLE OF CONTENTS

Introduction	4
Catalogue development	5
Catalogue functions and features	7
Report on population of the catalogue	13
Next steps	15
Annexe 1: Manuals for MOLGENIS catalogue	17
Annexe 2: Catalogue metadata model	36

LIST OF FIGURES

- Figure 1:** European health research and samples catalogue
- Figure 2A:** EOSC4Cancer catalogue landing page
- Figure 2B:** The search interface for data sources
- Figure 3:** Interoperability map
- Figure 4:** Catalogue data model overview
- Figure 5:** Screenshot of the list view of cohorts (n=147) currently populating the EOSC4Cancer catalogue

Introduction

The long term vision of this deliverable is to create a sustainable European catalogue for cancer resources, that integrates the cataloguing efforts of individual cancer research projects and consortia.

The complex nature of cancer requires integration of advanced research data across national boundaries to enable progress in prevention of cancer and the development of optimal medical care. The European mission board for cancer has identified access to data, knowledge and digital services across European borders vital for the cancer mission. The goal in EOSC4Cancer is to make cancer-related genomic, imaging, clinical, environmental and socio-economics data accessible through the use and improvement of existing federated and interoperable systems. These systems will provide an infrastructure to securely identify, share, process and reuse FAIR cancer data across borders. EOSC4Cancer use-cases will cover the patient journey from cancer prevention to diagnosis and treatment, laying the foundation of data trajectories and workflows for future cancer research projects. Curated and FAIR datasets will be essential for advanced analytics and computational methods, including machine learning, to be reproducible and robust.

The overall goal of WP1 is to enhance access to cancer-related data by increasing the FAIRness of the participating resources by contributing data to the EOSC4Cancer metadata catalogue. The resources are data providers, both from within the consortium as well as data providers from institutes not involved in EOSC4Cancer, who will contribute their relevant study metadata. We have developed a queryable directory of the participating resources to allow researchers and clinicians to discover relevant resources and datasets containing variables including phenotypes of interest. We have built this directory by adopting and enhancing the MOLGENIS catalogue platform that is also used by other major EU projects and infrastructures, notably BBMRI-ERIC Directory, European Human Exposome Network (EHEN) and a half dozen large EU multi-center cohort studies, IMI projects and the European Medicine Agency. The mission of the MOLGENIS catalogue is to provide a sustainable 'software as a service' catalogue that allows projects to sustain their cataloguing results beyond project end-dates.

This deliverable aims to document the EOSC4Cancer catalogue which showcases existing cancer resource metadata for discovery (building on experience and standardised templates developed in previous projects, such as BBMRI-ERIC), request and as reference during research. Furthermore, this deliverable will support the process of data harmonisation and versioning of the variable mappings in EOSC4Cancer towards joint data analysis and cross-study comparisons (i.e. promoting the FAIR principles). We are now ready for the next phase, which is to invite other cancer research networks to open up and merge their cataloguing contents into this catalogue. For this interoperability purpose, the EOSC4Cancer catalogue is prepared for the implementation of data catalogue vocabulary (DCAT). DCAT is a resource description framework (RDF) vocabulary which has been designed to increase interoperability between web published data catalogues.

Below, we:

- Describe the development process
- Summarise the catalogue functions and features,
- Report on the catalogue data capture process thus far
- Describe next steps for further development and dissemination

Catalogue development

The EOSC4Cancer consortium has access to established cancer research networks and European research infrastructures for cancer data, including cancer genomics, imaging, clinical, environmental, socio-economics data and biobanks. Expanding on Deliverable 1.3, we established a **metadata catalogue** to enable researchers to (i) assess all these data sets, (ii) assess the suitability of the data sets to answer specific research questions, and (iii) to facilitate the deployment of WP4 on federated multi-institutional research use-cases.

The aim is to create a publicly findable metadata catalogue to make the collected cancer-related datasets findable, and provide suitable documentation and tools to improve their reusability (while promoting data harmonisation with WP2 to ease pooled data analysis). The metadata catalogue comprises relevant cancer related data sets collected both within EOSC4Cancer and beyond the project consortium. Specifically, an EU consortium metadata catalogue has been equipped and added to the European Networks Health Data and Cohort Catalogue for EOSC4Cancer.

The EOSC4Cancer data catalogue supports researchers and bioinformaticians in their collaborations, sharing their data and results and to document the data harmonisation efforts of WP2. The first major feature of the catalogue is to enable users to discover relevant data sources based on rich resource descriptions. Detailed metadata templates have been included to document metadata from cohorts, secondary use data, registries, studies, networks and organisations. Extensions to describe genomic and imaging datasets are currently in draft in collaboration with EU GDI (genome data infrastructure) and EUCAIM (EU cancer imaging initiative), respectively. The second major feature of the catalogue is to document data harmonisation efforts. In this part of the catalogue, cancer data sources show how they map their collected variables onto common standards to enable pooled analysis, e.g. using Maelstrom protocols¹. User journey and usability of the catalogue is of the utmost importance to ensure that they are able to make use of the catalogue in the most effective way. Therefore, new versions of user interfaces of the platform are underway.

Our ambition has been to connect the EOSC4Cancer data to a large resource of other EU project metadata rather than developing an isolated cancer catalogue. Such a common EU health data catalogue facilitates reuse of metadata across projects, sharing of data harmonisation protocols and common data elements and will ease sustainability beyond EOSC4Cancer. Therefore, as per project planning, the catalogue has been joined with the existing MOLGENIS based European Networks Health Data and Cohort Catalogue, which is implemented using the existing MOLGENIS data catalogue, a unified framework that enables European-wide sharing of data catalogues as described in Swertz et al (2022)². The project is open source, which allowed EOSC4Cancer to refine the catalogue data model. We have used this model in several new releases of the software and polished the user interface. Most importantly, the resources map has been populated with internal and external project data resources. Previously and simultaneously, many EOSC4Cancer partners, supported by the MOLGENIS team at University Medical Center Groningen (UMCG), have implemented the data resources map, a FAIR metadata cataloguing platform.

Interviews with various catalogue users from multiple project partners were conducted to discover required improvements to the catalogue data model and user interface. Based on these interviews a list of requirements was drafted. In the previous set-up (D1.3), the EU data catalogue, of which the EOSC4Cancer catalogue is a part, consisted of one homepage and

¹ Fortier et al (2022) Life course of retrospective harmonisation initiatives: key elements to consider. J Dev Orig Health Dis. <https://pubmed.ncbi.nlm.nih.gov/35957574/>

² Swertz et al (2022) Towards an Interoperable Ecosystem of Research Cohort and Real-world Data Catalogues Enabling Multi-center Studies. Yearb Med Inform. <https://pubmed.ncbi.nlm.nih.gov/36463884/>

multiple similar search pages per network. Next to this, there were configured search pages and detail pages for each resource type (e.g., studies or cohorts). There was also a separate page for the study variables and variable harmonisation details. In the current set-up each network has its own homepage, from which all other data catalogues can be found. The user interface (UI) of this network specific homepage is adaptable to fit the EOSC4Cancer specific branding.

Catalogue functions and features

The catalogue aims to support the data reuse life cycle (supporting FAIR principles: find relevant data, acquiring access, help to make datasets interoperable and support the actual reuse). Focus is on the 'metadata', i.e. description/documentation of the data. This complements the other EOSC4Cancer services, i.e. those servicing the 'data' (data access and analysis).

From a catalogue end user perspective, the catalogue comprises the following steps:

- An '**all networks landing page**' to navigate to individual catalogues, including EOSC4Cancer. This one is offered by MOLGENIS catalogue. See figure 1.
- The EOSC4Cancer **data discovery page**, which focuses on the 'F' and 'A' in FAIR. On this page users can find datasets based on *rich metadata*, including access conditions, the type of dataset (e.g., cohort, registry, study), population characteristics, number of participants, and the data variables; See figure 2. We consider this to be the main landing page for EOSC4Cancer.
- The EOSC4Cancer **interoperability map**, which focuses on the 'I' and 'R' in FAIR, that enables documentation of data resource contents on the 'variable level' (i.e. codebooks) helping reuse, documenting harmonised 'standard' variables and datasets and standard models thereof (e.g. OMOP), and providing *interoperability* maps showing if and how collected variables from the data sets have been mapped to these standards; See figure 3.

From a catalogue data manager perspective, the catalogue also comprises of:

- **Manuals** to explain to users how to use the catalogue. See annexe 1.
- A **catalogue metadata model** that allows networks and EU projects to capture the required details on both data source and variable level, and that is compatible with standards, such as DCAT. See annexe 2.
- **Standard operating procedures** for deployment, upgrading and maintenance of the software application and database.
- And the MOLGENIS platform with the MOLGENIS catalogue extension enabled.

Each of these elements will be introduced below.

Catalogue user journey

From an end user perspective, users will first encounter the EOSC4Cancer landing page, which is used for data discovery, from the home page they can discover the interoperability map. Depending on how users find and reach the catalogue, they might start from the 'all networks' landing page, or from a different network-specific landing page. From both the 'all networks landing page and from different network-specific landing pages they can reach the EOSC4Cancer landing page.

All networks landing page

EOSC4Cancer can be found on the MOLGENIS catalogue as one of the networks, see Figure 1. Note that we generally expect users to not use this landing page but to immediately be linked to the EU cancer catalogue, see next section. However, users can use this user interface to find if data sources are also available in other catalogues.

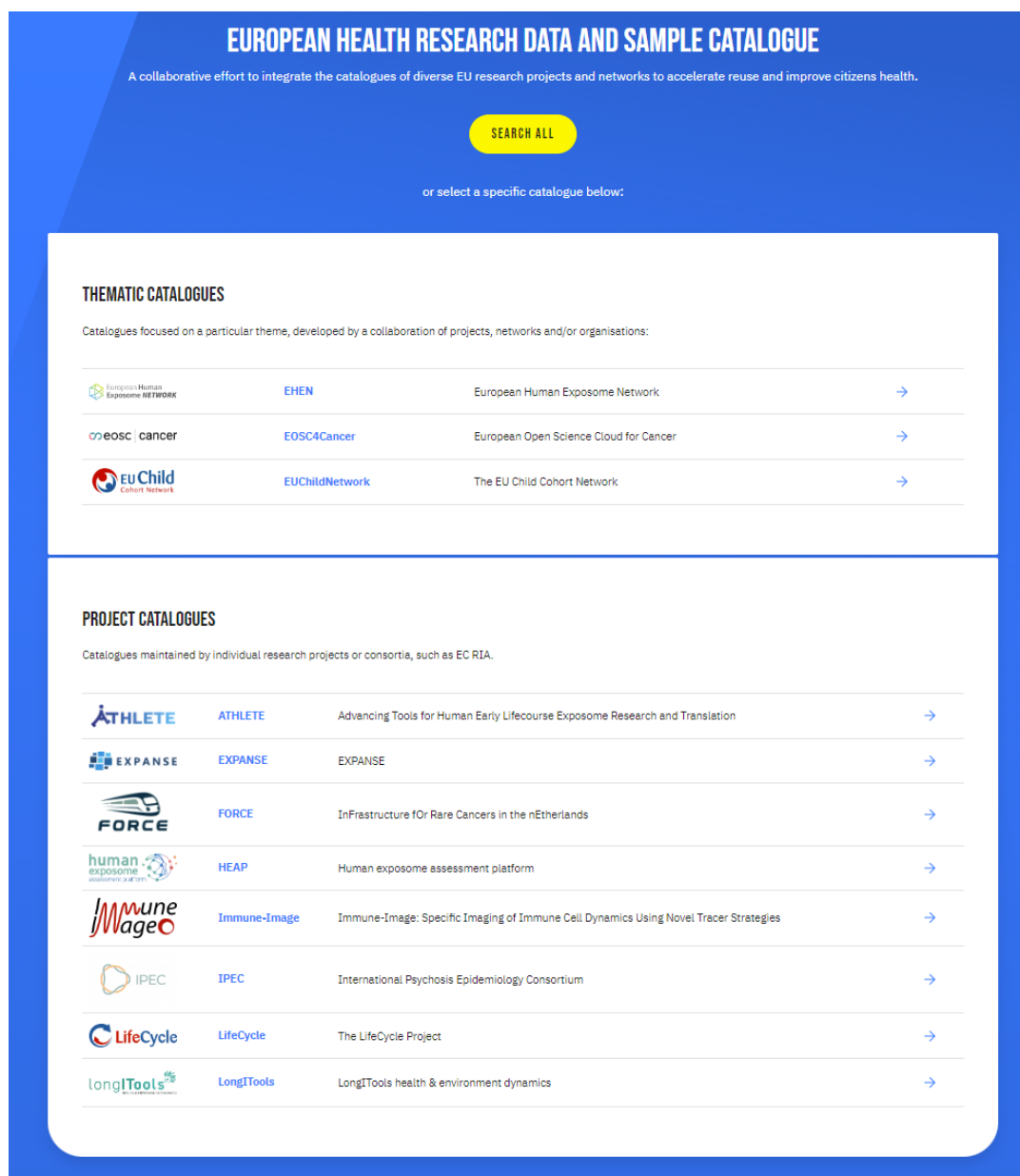


Figure 1: European health research and samples catalogue. Accessible via: <https://data-catalogue.molgenisccloud.org/catalogue/ssr-catalogue/>.

EOSC4Cancer data discovery page

The main landing page for the EOSC4Cancer catalogue is the starting point to browse cancer data resources, their datasets and the harmonised variables of the network. It is also a dashboard containing statistics for the available data resources and their datasets. Once other cancer networks join we aim to generalise the label of the EOSC4Cancer catalogue to 'EU cancer catalogue'. EOSC4Cancer can be found in the networks section, see Figure 1. In Figure 2 it can be seen that different types of data resources are navigable, and searchable (which we expect to expand with networks and studies), and also the interoperability map can be reached by clicking 'variables'. The EOSC4Cancer landingpage is publically available and accessible via:


<https://data-catalogue.molgenisccloud.org/catalogue/ssr-catalogue/EOSC4Cancer> . See **Figure 2** for a screenshot of the EOSC4Cancer landing page and a screenshot of the interface for data sources.

eosc | cancer

EOSC4CANCER COHORTS DATA SOURCES ABOUT

EUROPEAN OPEN SCIENCE CLOUD FOR CANCER

EOSC4Cancer will make diverse types of cancer data accessible: genomics, imaging, medical, clinical, environmental and socio-economic. It will use and enhance federated and interoperable systems for securely identifying, sharing, processing and reusi...




COHORTS

147

A complete overview of EOSC4Cancer cohorts and biobanks.

COHORTS




DATA SOURCES

6

EOSC4Cancer databanks and registries

DATA SOURCES




VARIABLES

46


EOSC4Cancer harmonized variables.

VARIABLES



23.916 Participants

The cumulative number of participants of all (sub)cohorts combined.



Longitudinal 1%

Percentage of longitudinal datasets. The remaining datasets are cross-sectional.

A

eosc | cancer

EOSC4CANCER COHORTS DATA SOURCES ABOUT

DATA SOURCES

Group of individuals sharing a defining demographic characteristic.

DETAILED COMPACT

FILTERS

Search in datasources

Type to search.. SEARCH

Areas of Information

Datasource Types

CNCR

Czech National Cancer Registry (CNCR in the Czech Republic)

Czech National Cancer Registry (National Oncology Registry - NOR) is register of oncological diseases, which periodically monitor them and their development in time. UZIS is in charge for data collection, verification, storage, protection and process...

Type	Participants
Population registry,Cancer registry	225466

CRASLNa3Sud

Cancer Registry ASLNa3Sud

Cancer Registry (ASLNa3Sud) is a register of oncological diseases, which periodically monitor them and their development in time. UZIS is in charge for data collection, verification, storage, protection and process...

Type	Participants
Population registry,Cancer registry	1100000

NCR

Netherlands Cancer Registry

B

Figure 2: A. The EOSC4Cancer catalogue landing page, from which you can navigate to the EOSC4Cancer cohorts, data sources and common data variables. B. The search interface for data sources, containing an overview for cancer databanks and registries.

EOSC4Cancer data interoperability map

The interoperability map will be made visible if you click 'variables' in Figure 2A. In EOSC4Cancer the mapping in WP2 is in process. Currently, a common data model has been drafted for use case 4.1. **Figure 3** shows the common data elements and an example of their mappings from the mtFIT study.

Figure 3A: EOSC4Cancer Variables Overview

The screenshot shows the 'VARIABLES' section of the EOSC4Cancer portal. It includes a search bar, a list of variables, and a 'HARMONIZATIONS' tab.

Variable	Harmonization
Age at invitation in years	→
Deepest point of scope insertion	→
Lesion location at colonoscopy	→
Data entry date	→
Resection result at colonoscopy	→
Lesion location pathology	→
ID patient	→

Figure 3B: Detailed Variable Harmonization Table

The screenshot shows a detailed table of variable harmonization. The table lists variables and their corresponding data sources, with a status column indicating completion.

Variable	MIKROTUM	MINDACT Biobank	MMC biobank	MMCI	MoBa	Molecular Investig...	MOS Biobank	MPP, MFM biobanks	MRC Brain Banks N...	MRC-UCL	MTBIO	MTCC	mtFIT study	MU_ICS	MuCo	MULCB	Multiple Sclerosis ...	Myeloma XII clinic...	Myobank-AFM	MYST Biobank	NAPKON	National Diet and N...	National Human Bi...
Age at invitation i...													✓										
Deepest point of ...													✓										
Lesion location at...													✓										
Data entry date													✓										
Resection result a...													✓										
Lesion location p...													✓										
ID patient													✓										
Date of birth													✓										
Colonoscopy ID													✓										

Figure 3: Interoperability map. A. An overview of the currently available variables in the common data models for EOSC4Cancer. These variables were made within WP2 for use case 4.1. B. Interoperability map showing mappings status per cohort for the common data elements. Accessible via:

<https://data-catalogue.molgenisccloud.org/catalogue/ssr-catalogue/EOSC4Cancer/variables>.

Data manager perspective

The catalogue only contains metadata (in simple terms data about data). The actual data within the data sets is held locally or will be stored in a centralised system (e.g. cBioPortal) and accessed by researchers independently of the catalogue. This will be reported in future deliverables (in collaboration between WP1, 2, 3, 4). Main elements from a data manager perspective are the catalogue data model, data access information, manuals and standard operating procedures, and finally the underlying MOLGENIS software, summarised below.

Catalogue data model

The 'engine' of the catalogue is the catalogue data model, see **figure 4** for an overview. The data model provides a tabular structure to document different flavours of data resources, contacts, organisations involved, and documentation of data sets, data dictionaries, and finally mappings towards common data models. The latest version of the data model is accessible via the MOLGENIS EMX2 Github repository

<https://github.com/molgenis/molgenis-emx2/blob/master/data/datacatalogue/molgenis.csv>. A complete documentation of the data model is included in **annexe 2**.

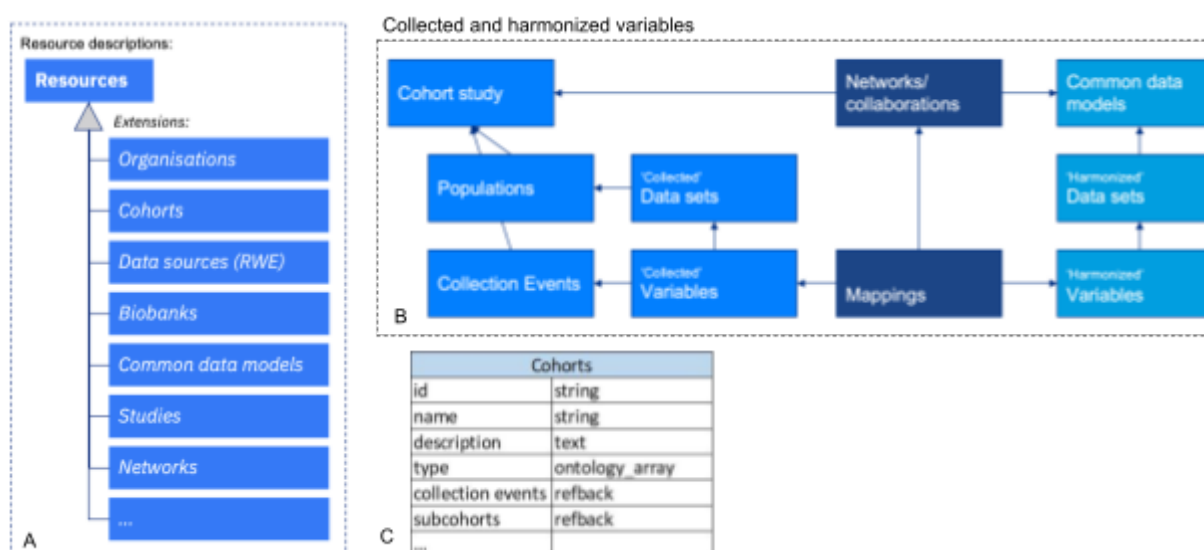


Figure 4. Catalogue data model overview. The model consists of two parts: A. The resource level metadata (e.g. describing a cohort, study or network as a whole) and B. A description of collected or standardised variables and interoperability between them. C. A subset of the data items within the cohorts table.

Manuals and standard operating procedures

To help data managers to submit information several manuals and operating procedures have been collated, see Annexe 1. In principle, every data submitter will be provided with a 'staging area' which is an empty copy of the catalogue (with links to existing data). Data submitters can fill in their details, reusing existing code systems. Once complete, the

submission can be put for review by the central data manager (currently UMCG) and if approved, merged into production.

MOLGENIS platform and MOLGENIS catalogue software

The open access web-based catalogue, building on the MOLGENIS catalogue project, can be accessed at <https://data-catalogue.molgenisccloud.org>. This catalogue software is built into the MOLGENIS FAIR data platform and is available for free as open source software for reuse by other networks at <http://github.com/molgenis/molgenis-emx2>.

Report on population of the catalogue

Next to refinement of the catalogue system to accommodate EOSC4Cancer stakeholders and use case needs, a major effort has been spent on a first collation of relevant resources.

Over the course of the first year of the project, meetings with EOSC4Cancer consortium partners have taken place with the aim of giving them information in the form of a demonstration of the catalogue and providing them the needed support to add their dataset(s) to the catalogue. First, a staging area has been set up for each partner where cohort metadata can be added. Furthermore, data managers of the EOSC4Cancer partner institutions have provided information about their data resources and data sets; this 'rich metadata' was added to the catalogue. To achieve this, we conducted a survey, followed by individual conversations, to gather the initial metadata. Data managers with appropriate permissions can fill out or edit more detailed standardised rich metadata via online forms in their data resource's staging area. After approval of the PI or data manager the data in the staging area is transferred to the live catalogue. A demo staging area can be found at <https://data-catalogue.molgeniscloud.org/testCohort/tables/#/>.

Currently, the catalogue is populated with a number of selected datasets from partners within EOSC4Cancer who initially filled out the survey and were willing to share their metadata to support the WP4 use cases. Besides the datasets that are publicly available in the metadata catalogue, several partners are in the process of adding their datasets in the staging areas and meetings have been scheduled with partners who will be onboarding next.

In addition to datasets from data providers who are participating in EOSC4Cancer, cancer-related collections from BBMRI-Directory biobanks have been added to the EOSC4Cancer catalogue, as shown in Figure 5. The BBMRI-Directory is one of the services offered by BBMRI-ERIC Common Services for IT (CS-IT) to the global biobank community and was created in collaboration with the BBMRI National Nodes and partners. The Directory provides a central listing of biobanks and their collections in the BBMRI-ERIC member states. For researchers, the Directory offers a means of finding samples and data, while for biobanks it offers a platform to share the existence of their holdings and services, and to connect with researchers interested in them.

Filters

Search in cohorts ^

Type to search.. SEARCH

Areas of information ▼

Data categories ▼

Population age groups ▼

Sample categories ▼

Cohort Types ▼

Design ▼

HOME > COHORTS

COHORTS

Group of individuals sharing a defining demographic characteristic.

DETAILED COMPACT

ACBB	Augsburg Central BioBank			→
The Augsburg Central BioBank is a centralized biobank at the University Hospital of Augsburg. It is divided into two parts, consisting of liquid and solid biosamples. For solid biosamples, we collect oncological biosamples with the main focus on lung ...				
Type	Design	Participants	Duration	
Biobank			not available	
AMC Biobank	Amsterdam UMC Biobank: Location AMC			→
Biobanks are increasingly important for conducting state-of-the-art biomedical research. AMC Biobank has been established in 2014 to support researchers in establishing biobanks that meet current ethical and quality standards. As one of the core faci ...				
Type	Design	Participants	Duration	
Biobank			not available	
ASK-Tx	Asklepios Biobank für Lungenerkrankungen - Asklepios Biobank for Lung Diseases			→
Es ist das Ziel der Biobank, Gewebe, Blut und andere Körperflüssigkeiten in hoher Qualität zu sammeln, zu lagern und für die wissenschaftliche Forschung zur Verfügung zu stellen. Sie ist Teil des „Comprehensive Pneumology Centers München“ (CPC-M) und ...				
Type	Design	Participants	Duration	
Biobank			not available	

Figure 5: Screenshot of the list view of cohorts (n=147) currently populating the EOSC4Cancer catalogue. In the screenshot the first three cohorts in the list are shown. BBMRI-ERIC biobanks are now listed as EOSC4Cancer cohorts of type 'biobank'. Accessible via: <https://data-catalogue.molgeniscloud.org/catalogue/ssr-catalogue/EOSC4Cancer/cohorts>.

Next steps

Moving forward there are three main efforts:

Support of harmonisation and use cases

A main effort of the catalogue will be to support the WP2 harmonisation work and WP4 use cases. This will include support of major metadata capture into the catalogue. For example, the project is now creating a genomic and clinical synthetic dataset based on real mutations and an EHR synthetic model in OMOP CDM format for use case 4.3, which will be documented in the catalogue. In addition, we expect to further refine the catalogue data model, e.g. to document data access facilities such as cBioPortal or EGA data access services.

Dissemination

The metadata catalogue is one of the key EOSC4Cancer outputs. Firstly, we aim to integrate the cataloguing efforts of all EU cancer related projects and we are therefore reaching out to each of them. Secondly, we are engaged in large scale standardisation efforts, notably the EUCAIM, GDI and EHDS pilot project to ensure catalogue contents can flow towards more general catalogues (DCAT-AP) and towards domain specific catalogues. Last but not least, the information about the catalogue will be widely disseminated to maximise the utilisation by both EOSC4Cancer researchers and other potential users.

Dissemination activities include:

- presentation of the catalogue to EOSC4Cancer partners at project meetings
- presentation of the catalogue during meetings of other cancer data related or data infrastructure related EU-projects, national projects and other collaborations e.g. GDI, EUCAIM, EHDS< EOSC-life, BBMRI.
- the EOSC4Cancer website features a page about WP1, which will contain information and links to the metadata catalogue³ describing what it is and its purpose and includes explanation of the terminology used for a non-scientific audience. People will be given the opportunity to subscribe to receive progress updates and contact us to collaborate.
- the catalogue is connected to other networks and EU projects via the European Networks Health Data and Cohort Catalogue¹

The catalogue and supporting materials are ready to assist the EOSC4Cancer harmonisation and federated analysis work. Driven by the needs of the researchers and data managers the catalogue system and standard operating procedures for deployment, upgrade and maintenance of the software application and database will be iteratively improved. Most effort will lie in further population of the catalogue in collaboration with WP4 and collaboration with data providers from other EU projects. It is our intention to create a sustainable catalogue, usable beyond EOSC4Cancer, that will inspire more partners and data providers beyond the project to participate.

³ <https://data-catalogue.molgeniscloud.org/catalogue/ssr-catalogue/EOSC4Cancer>

Continuous improvement

We aim to keep on improving the catalogue as a system towards the cancer mission, in particular:

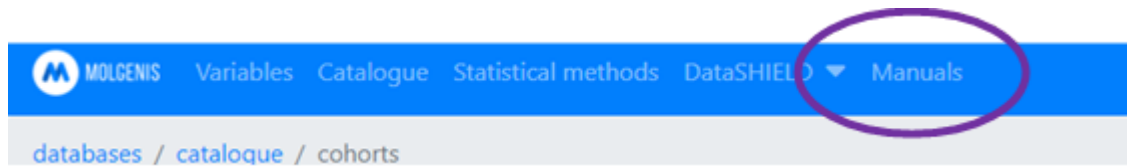
- **Populate the catalogue** with more EOSC4Cancer related resources, linking and integrating cataloguing efforts from partners and from the data sources directly. A large amount of metadata has been collected and catalogued in the cancer research networks of the past and present. In order to increase the discoverability of this data, we will be inviting other cancer research networks to open up their cataloguing content and to merge it into this catalogue.
- **Refinement of the data catalogue data model.** While the current catalogue already includes many best practices, we see options for extensions for particular domains such as imaging and genomics with the aim to provide searchable repositories. In collaboration with partners from the EUCAIM project and FAIR genomes⁴ recommendations (that are also now being developed in the GDI project. In addition, we are working with BBMRI to also include 'quantitative' data by loading pivot tables with counts that are e.g. grouped by disease to help dataset selection. With GDI we aim to pilot how 'live counting' might be possible using beacon protocol.
- **Expansion of Data Use Ontology tagging** for genetic and medical imaging datasets
- **Evaluating the functionality** of the catalogue against available use cases, in particular in close collaboration with WP2 and WP4 to work on delivery of standardised, harmonised codebook variables and a method to transform data to this format.
- **Optimise the user experience.** We have positioned EOSC4Cancer within the larger Health Data Catalogue initiative and we are in the process of developing the user interface to fit both the needs of the EOSC4Cancer use cases as well as other interested users, including researchers and clinicians. (See Figure 1).
- Landscaping and planning has started for **federated data flow and integration**, in close collaboration with WP1 and WP2 partners.
- Explore if we can also expand into **data request and access workflows** (e.g.. testing GDI data request service module and integrating with WP1-4 data access and analysis services).

The overall mission is to further streamline the process where researchers can find, select and request access to relevant data on a useful data analysis platform. The catalogue will provide sufficient metadata for reuse of cancer data, to support the EU cancer mission, accelerate data intensive cancer research and impact patients' lives.

⁴ Van der Velde et al (2022) FAIR Genomes metadata schema promoting Next Generation Sequencing data reuse in Dutch healthcare and research. <https://pubmed.ncbi.nlm.nih.gov/35418585/>

Annexe 1: Manuals for MOLGENIS catalogue

Users can navigate from within the catalogue to help in use of the catalogue:



Choosing 'Manuals' takes the user to the following page of documentation: <https://data-catalogue.molgeniscloud.org/apps/docs/#/>. Here documentation can be found for technical experts requiring information about the platform, and for researchers and data managers requiring information about submitting data to the catalogue and Armadillo, grouped according to need.

MOLGENIS Catalogue Guide

We are highly motivated to promote multi-centre human data analysis to improve people's health, linking data from many local institutions into networks, optionally by mapping data to common data models, for use in new research. To **find** the data we developed a [catalogue](#). Here you can find the metadata of the cohorts and the common data model.

We distinguish three roles when using the catalogue:

		<u>Researcher</u>	<u>Local data manager</u>	<u>Network data manager</u>
Catalogue	Data mapping/harmonisation		<u>X</u>	
Catalogue	Local metadata description		<u>X</u>	
Catalogue	Describe and upload common data model			<u>X</u>
Catalogue	Find (common) variables	<u>X</u>		
Catalogue	Find mapping specifications	<u>X</u>	<u>X</u>	
Catalogue	Request access		<u>X</u>	<u>X</u>

Researcher

Catalogue

When you have defined a research question it is time to find the relevant variables to answer it. You can use the [MOLGENIS Data Catalogue](#) to compose your dataset. The Data Catalogue only describes the variables, it does **not** contain the actual data values.

Find variables

You can use the filters and search bar to subset variables. In the example below, four filters are used, one network filter and three topic filters.

The screenshot shows the MOLGENIS Variable Explorer interface. The top navigation bar includes links for Variables, Catalogue, Statistical methods, DataSHIELD, and Help, along with Sign in and Sign up buttons. The main section is titled 'Variable Explorer' and is divided into two main panels: 'Filters' on the left and 'Variables (4)' on the right.

Filters Panel:

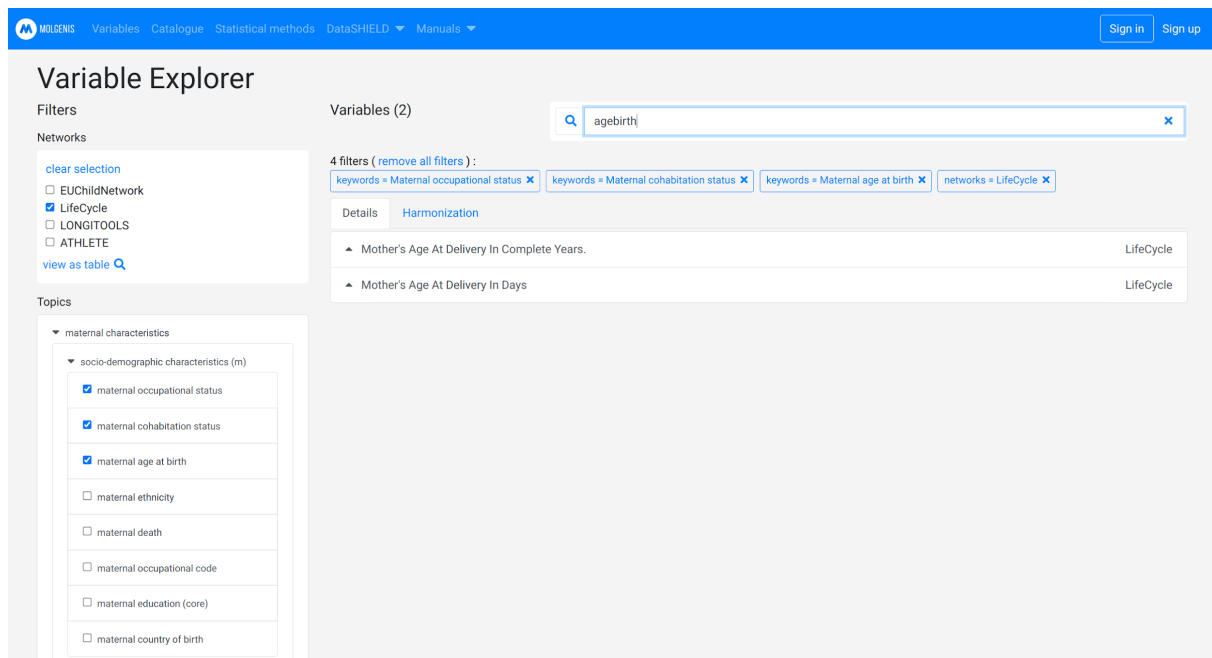
- Networks:** Includes a 'clear selection' link and checkboxes for EUChildNetwork, LifeCycle (checked), LONGITOOLS, and ATHLETE. A 'view as table' link is also present.
- Topics:** Includes a dropdown for 'maternal characteristics' which is expanded to show 'socio-demographic characteristics (m)'. Under this, checkboxes for 'maternal occupational status', 'maternal cohabitation status', and 'maternal age at birth' are checked. Other unchecked options include 'maternal ethnicity', 'maternal death', 'maternal occupational code', 'maternal education (core)', and 'maternal country of birth'. A partially visible 'lifestyle characteristics (m)' section is at the bottom.

Variables (4) Panel:

- A search bar with the placeholder 'Search variables' and a magnifying glass icon.
- A summary of 4 filters: 'keywords = Maternal occupational status', 'keywords = Maternal cohabitation status', 'keywords = Maternal age at birth', and 'networks = LifeCycle'.
- Two tabs: 'Details' and 'Harmonization'.
- A table of variables:

Variable Name	Network
▶ Mother's Age At Delivery In Days	LifeCycle
▶ Maternal Occupational Status (>-1 Year And <1 Year)	LifeCycle
▶ Cohabitation Status Of The Mother (Age ≥0 Year And <1 Year)	LifeCycle
▶ Mother's Age At Delivery In Complete Years.	LifeCycle

You can search on variables (for example: **agebirth**) with filters already set in place.



Variable Explorer

Filters

Networks

clear selection

☐ EUChildNetwork

☒ LifeCycle

☐ LONGITOOLS

☐ ATHLETE

[view as table](#)

Topics

material characteristics

socio-demographic characteristics (m)

☒ maternal occupational status

☒ maternal cohabitation status

☒ maternal age at birth

☐ maternal ethnicity

☐ maternal death

☐ maternal occupational code

☐ maternal education (core)

☐ maternal country of birth

Variables (2)

agebirth

4 filters (remove all filters):

keywords = Maternal occupational status X keywords = Maternal cohabitation status X keywords = Maternal age at birth X networks = LifeCycle X

Details Harmonization

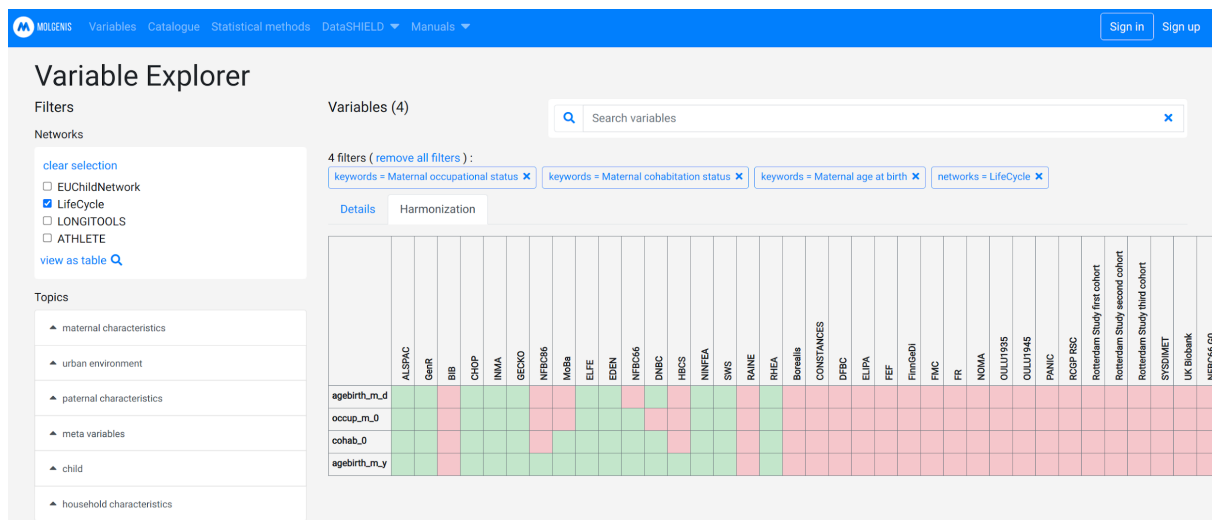
Mother's Age At Delivery In Complete Years. LifeCycle

Mother's Age At Delivery In Days. LifeCycle

In the future you will be able to use the shopping cart to select all variables of interest and create an overview.

Find harmonisation details

The harmonisation view allows you to see which cohorts have (partially) harmonised your variable of interest and thus have that variable available for analysis.



Variable Explorer

Filters

Networks

clear selection

☐ EUChildNetwork

☒ LifeCycle

☐ LONGITOOLS

☐ ATHLETE

[view as table](#)

Topics

material characteristics

urban environment

paternal characteristics

meta variables

child

household characteristics

Variables (4)

Search variables

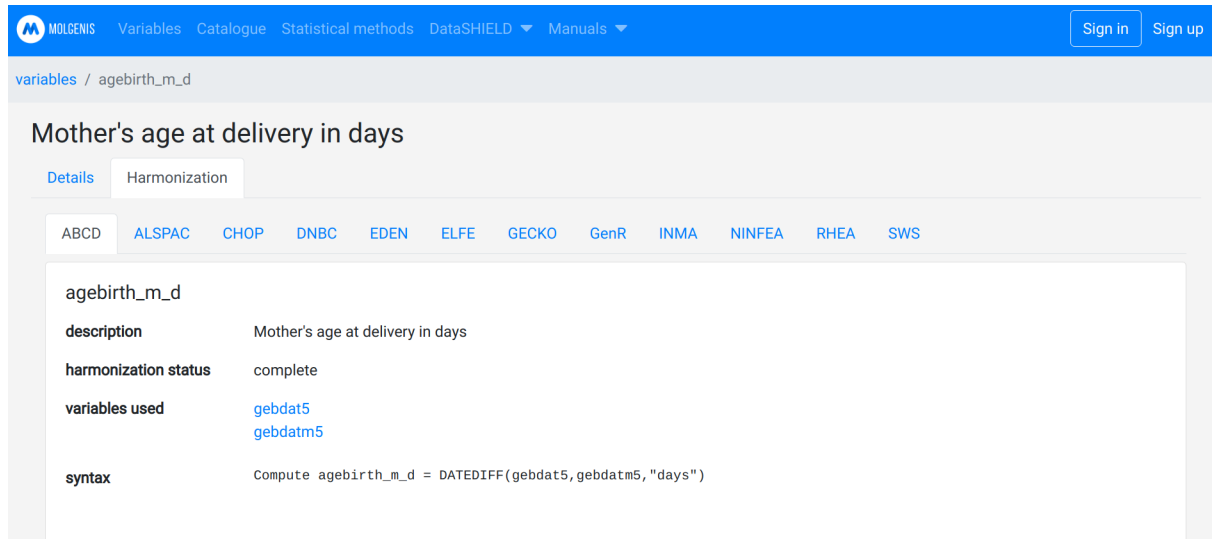
4 filters (remove all filters):

keywords = Maternal occupational status X keywords = Maternal cohabitation status X keywords = Maternal age at birth X networks = LifeCycle X

Details Harmonization

	ALSPAC	GenR	BIB	CHOP	INMA	GECKO	NFBC96	MoBa	ELFE	EDEN	NFBC56	DNBC	HBCS	NINFEA	SWS	RAINE	RHEA	Borealis	CONSTANCES	DFBC	ELIPA	FEF	FinGenD	FMC	FR	NOMA	OULLU1935	OULLU1945	PANIC	RCGP RSC	Rotterdam Study first cohort	Rotterdam Study second cohort	Rotterdam Study third cohort	SYSDIMET	UK Biobank	NFBC56 00
agebirth_m_d																																				
occup_m_0																																				
cohab_0																																				
agebirth_m_y																																				

You can view how a specific cohort has harmonised a specific variable.



The screenshot shows the MOLGENIS Data Catalogue interface. The top navigation bar includes links for Variables, Catalogue, Statistical methods, DataSHIELD, and Manuals. The main content area displays the variable 'agebirth_m_d' under the 'Mother's age at delivery in days' title. It includes tabs for 'Details' and 'Harmonization'. Below these are tabs for various data sources: ABCD, ALSPAC, CHOP, DNBC, EDEN, ELFE, GECKO, GenR, INMA, NINFEA, RHEA, and SWS. The variable details are shown in a table:

agebirth_m_d	
description	Mother's age at delivery in days
harmonization status	complete
variables used	gebdat5 gebdatm5
syntax	Compute agebirth_m_d = DATEDIFF(gebdat5, gebdatm5, "days")

Data manager of a cohort or data source

Data Catalogue

[MOLGENIS Data Catalogue](#) provides a framework to describe in detail: metadata of cohorts and of data sources; definitions of the variables collected in cohorts and data sources; and mappings to common data models. Its purpose is to facilitate pooled data analysis of multiple cohorts [Fortier et al, 2017](#) and multi-data source studies [Gini et al, 2020](#).

- The metadata of cohorts include descriptive information such as contact details, name of the cohort, and high-level summary of contents and cohort design. The metadata of data sources, of the corresponding data banks and of the organisations that provide access to them, include descriptive information such as contact details, reason for existence of the data banks, the prompt for the records in the data bank, and lag time for updating and accessing data.
- The metadata of the source variables can be considered the codebook or data dictionary of a cohort (e.g. ALSPAC) and of the tables which make up a data source's data bank(s) (e.g. the Danish Healthcare Registries).
- Similarly, the common data models (or 'target variables') can be considered the codebook for a network of organisations with access to cohorts or data sources (e.g. LifeCycle or ConcePTION)
- The mappings describe how source variables have been converted into target variables as a basis for integrated analysis.

Data harmonisation

Each organisation with access to data (which may be a cohort, or a data source composed of one or more data banks) harmonises their data according to the consortium's protocols into a common data model (CDM) format which has been centrally agreed upon. In some projects, data may be made available via [DataSHIELD](#). In these cases each resource stores the data locally in a [MOLGENIS Armadillo](#) DataSHIELD server.

Staging areas for uploads

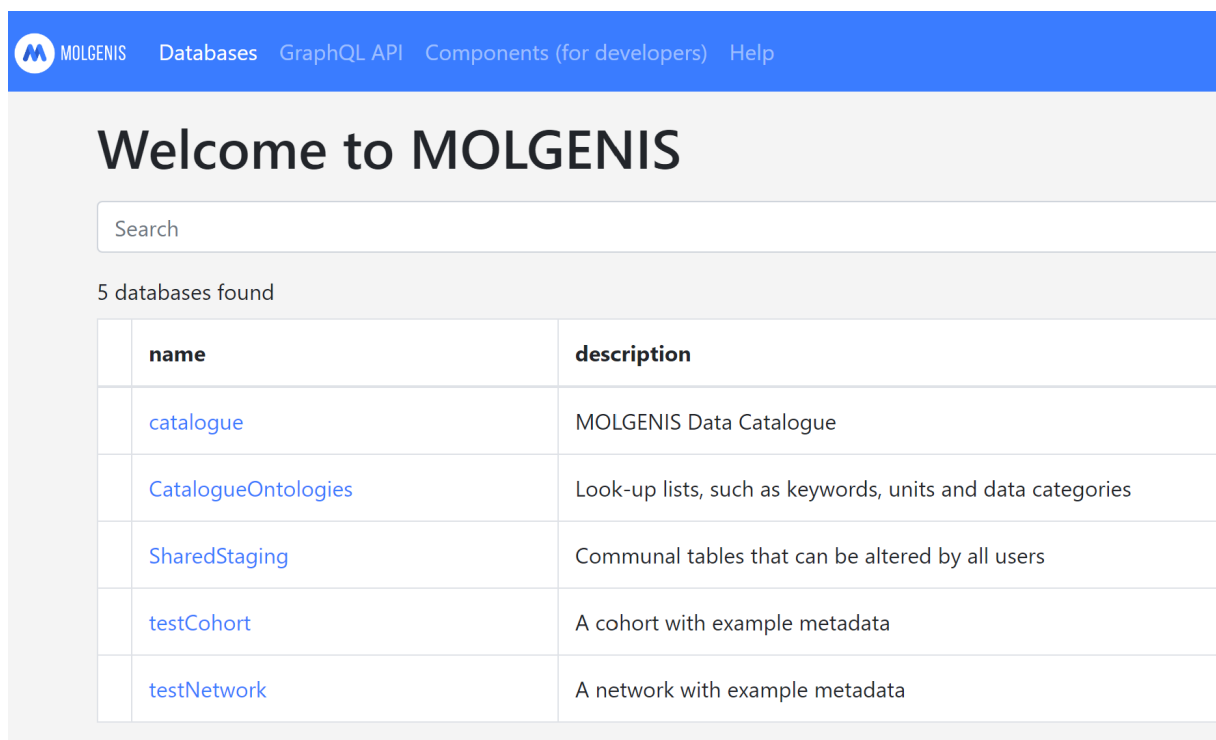
The metadata of the cohort or of the data source are first uploaded into what are called “staging areas” of the Data Catalogue. Later on the metadata are transferred to production; use of a staging area allows for review before the metadata are entered in the live catalogue.

You will need credentials to login and upload metadata.

Cohorts in projects such as ATHLETE, IPEC and LongITools use [data-catalogue-staging](#). ConcePTION uses [conception-acc](#).

When you log in, you will be able to see at least the following databases:

- **DataCatalogue:** The catalogue data, in which you can search for target variables to map to.
- **CatalogueOntologies:** This database contains the look-up list that you need for filling out some columns in the templates, e.g. format or unit. If you need to add anything to these look-up lists, contact us at molgenis-support.
- **SharedStaging:** A communal staging area in which Organisations are added and edited.
- **Your own database:** use this to upload the templates once you have filled them out.
- **Your own database :** (here: testCohort and testNetwork) Use this to fill out rich metadata and to upload the templates once you have filled them out.



MOLGENIS Databases GraphQL API Components (for developers) Help

Welcome to MOLGENIS

Search

5 databases found

name	description
catalogue	MOLGENIS Data Catalogue
CatalogueOntologies	Look-up lists, such as keywords, units and data categories
SharedStaging	Communal tables that can be altered by all users
testCohort	A cohort with example metadata
testNetwork	A network with example metadata

Figure 1. Databases in the Data Catalogue staging area.

Fill out cohort rich metadata

Open your staging area, navigate to ‘Tables’ and open the table ‘Cohorts’. Your cohort id and name are already filled out. Click on the pencil sign next to this entry to start editing your cohort rich metadata by filling out the form. ‘Subcohorts’ and ‘Collection events’ should also be filled out through this route. You can fill them out in subsections inside the ‘Cohorts’ form.

Define metadata of cohorts or data sources

This section explains how to submit the ‘source variables’ + ‘mappings from source variables to target variables’ into the Data Catalogue. Expected users of this ‘how to’ are data managers within the organisations with access to cohorts or data sources. This document assumes you have received login details for upload of your metadata. You can also watch this [instruction video](#). Note that this video used dictionary model version 2.x, which was updated to 3.x.

Define source variable metadata / source data dictionary

We use the [SourceDictionary template](#) to define variable metadata. The [SourceDictionary template](#) consists of multiple sheets. Each sheet corresponds to a table in the Data Catalogue (Figure 1). The columns in the sheet correspond to columns in the table concerned. This document describes how to fill out each of the sheets and their columns. A column with an asterisk (*) after its name is mandatory, i.e., it should contain values for the system to accept a data upload. You can download this [filled out example](#) as a reference for filling out the template. Note that there is no sheet for *AllVariables*. This table is a generic listing of all variables entered for the cohort; it shows *Variables* and *RepeatedVariables* in one table.

It is good practice to try adding a few variables to the template first and see whether your upload succeeds. To upload the metadata to the Data Catalogue see the section [Upload metadata](#).

Tables in 'testCohort'

Download all tables: [zip](#) | [excel](#) | [jsonld](#) | [ttl](#)

Data tables

Table	Description
All variables	Generic listing of all source variables. Should not be used directly, please use SourceVariables or RepeatedSourceVariables instead
Cohorts	Group of individuals sharing a defining demographic characteristic
Collection events	Definition of a data collection event for a resource
Contacts	
Data resources	Resources for data
Dataset mappings	
Datasets	Definition of a dataset within a (common) data model
Documentation	Documentation attached to a resource
Extended resources	
External identifiers	
Organisations	Research departments and research groups
Publications	Publications following bibtex format
Repeated variables	Definition of a repeated sourceVariable. Refers to another variable for its definition
Resources	Generic listing of all resources. Should not be used directly, instead use specific types such as Databanks and Studies
Subcohort counts	
Subcohorts	Subcohorts defined for this resource
Variable mappings	Mappings from collected variables to standard/harmonized variables, optionally including ETL syntax
Variable values	Listing of categorical value+label definition in case of a categorical variable
Variables	Definition of a non-repeated variable, or of the first variable from a repeated range
Version	3.2

Figure 2. Tables in a cohort's database in the Data Catalogue. Note that not all tables are filled out via the templates, some are filled via an online form, see section [Fill out cohort rich metadata](#).

Datasets sheet

The datasets/tables in a cohort or in the data banks of a data source are defined in the *Datasets* sheet. Columns with an asterisk (*) after their name are mandatory.

Column name	Description	Remarks
resource *	Cohort or data source that this table belongs to	
name *	Unique dataset or table name	

label	Dataset label	
description	Dataset description	
unit of observation	Defines what each record in this dataset describes	
number of rows	Count of the number of records in this dataset	
keywords ¹	Enables grouping of datasets into topics and helps to display variables in a tree	Find list to choose from in CatalogueOntologies Keywords
since version	Version of the data model when this dataset was introduced	e.g. 1.0.0 or 2.1
until version	Version of the data model when this dataset was deleted	e.g. 2.0.0 or 2.1

Table 1. Description of the columns that can be filled out for Datasets. * = mandatory; 1 = contact molgenis-support@umcg.nl to add Keywords

[Variables sheet](#)

The variables of the datasets specified in the *Datasets* sheet are defined in the *Variables* sheet.

Column name	Description	Remarks
resource *	Cohort or databank that this variable belongs to	Fill out your cohort or databank id
dataset *	Dataset that contains the variable.	Datasets must be predefined in the <i>Datasets</i> sheet
name *	Variable name, unique within a dataset	
label	Human readable variable label	
format	The data type of the variable	Find list to choose from in CatalogueOntologies Formats
unit ¹	Unit in case of a continuous or integer format	Find list to choose from in CatalogueOntologies Units
description	Description of the variable	

exampleValues	Examples of values in a comma separated list	Makes your data more insightful. E.g. 1,2,3 or TRUE,FALSE or 1.23,4.56,3.14
vocabularies ¹	Refer to ontologies being used	Find list to choose from in CatalogueOntologies Vocabularies
collection event.resource	Your cohort id	The collectionEvent needs to be predefined in the <i>CollectionEvents</i> sheet; e.g. y1 or y2
collection event.name	Refer to the name of a collection event	The collection event needs to be predefined in the <i>CollectionEvents</i> sheet; e.g. y1 or y2
keywords ¹	Enables grouping of variables into topics and helps to display variables in a tree	Find list to choose from in Catalogue Keywords
since version	Version of the data model when this variable was introduced	e.g. 1.0.0 or 2.1
until version	Version of the data model when this variable was deleted	e.g. 2.0.0 or 2.1

Table 2. Description of the columns that can be filled out for Variables. * = mandatory; 1 = contact molgenis-support@umcg.nl to add Vocabularies, Keywords or Units

[Variable values sheet](#)

The coding of categorical variables is defined in the *Variable values* sheet. This sheet is optional, but it is highly recommended to fill out the codes and values for your categorical variables, so that your data becomes more insightful for those that are interested.

Column name	Description	Remarks
resource *	Cohort or databank that the variable belongs to	Fill out your cohort or data bank id
variable.dataset *	Dataset that contains the variable	Datasets must be predefined in the <i>Datasets</i> sheet
variable.name *	Variable name	Variables must be predefined in the <i>Variables</i> sheet
value *	The code or value used	e.g. 1, 2 or -99
label *	The label corresponding to the value	e.g. 'yes', 'no' or 'NA'

order	The order in which the code list should appear	e.g. 1
is missing	Whether this value indicates a missing field	TRUE or FALSE
since version	Version of the data model when this variable value was introduced	e.g. 1.0.0 or 2.1
until version	Version of the data model when this variable value was deleted	e.g. 2.0.0 or 2.1

Table 3. Description of the columns that can be filled out for Variable values. * = mandatory

Repeated variables sheet

The *Repeated variables* sheet is optional, and is most often used by cohorts whose variables are observed repeatedly. Variables that are repeats of a variable defined in the sheet *Variables* are defined in the *Repeated variables* sheet. Defining your repeated variables using this sheet will limit the amount of information that has to be repeated when filling out repeated variables. This sheet is optional.

Column name	Description	Remarks
resource *	Cohort or databank that this variable belongs to	Fill out your cohort or databank id
dataset *	Dataset name.	e.g. core
name *	Variable name.	e.g. height_1
Is repeat of.dataset *	Dataset that contains the variable that is repeated	Tables must be predefined in the <i>Datasets</i> sheet; e.g. core
is repeat of.name *	Name of the variable that is repeated	Variables must be predefined in the <i>Variables</i> sheet; e.g. height_0
collection event.resource	Your cohort id	The collection event needs to be predefined via forms; e.g. y1 or y2
collection event.name	Refer to the name of a collection event	The collection event needs to be predefined via forms; e.g. y1 or y2
since version	Version of the data model when this variable was introduced	e.g. 1.0.0 or 2.1
until version	Version of the data model when this variable was deleted	e.g. 2.0.0 or 2.1

Table 4. Description of the columns that can be filled out for Repeated variables. * = mandatory

Define harmonisations

We use the [Mappings template*](#) to describe the harmonisations. The [Mappings template*](#) consists of two sheets (Dataset mappings and Variable mappings). It is used to define the mappings from source variables to target variables, or the Extraction, Transformation and Load (ETL) process from a data source to a common data model (CDM). You can download this [filled out example](#) as a reference for filling out the template.

Dataset mappings sheet

Harmonisation procedures at the table level are defined in the *Dataset mappings* sheet, irrespective of whether the table is in a cohort or in a data bank.

Column name	Description	Remarks
source *	Databank or cohort id	
source dataset *	Source dataset name	Datasets must be predefined in the <i>Datasets</i> sheet in the SourceDictionary template
target *	Name of the target common data model	e.g. LifeCycle_CDM, LongITools_CDM, see variable explorer
target dataset *	Target dataset name	Map to a dataset that is defined in a common data model
description	Description of the harmonisation	
syntax	Syntax used for this harmonisation	

Table 5. Description of the columns that can be filled out for Variable mappings. * = mandatory

Variable mappings sheet

Harmonisation procedures at the variable level are defined in the *Variable mappings* sheet.

Column name	Description	Remarks
source *	Databank or cohort id	
source dataset *	Source table name	Datasets must be predefined in the <i>Datasets</i> sheet in the SourceDictionary template

source variables	Source variable name(s)	Variables must be predefined in the <i>Variables</i> sheet in the SourceDictionary template; When multiple variables are mapped together use a comma-separated list, e.g. v1,v2,v3
source variables other datasets.dataset	Other source tables	When using variables from multiple other datasets, use a comma-separated list, e.g. dataset1,dataset2,dataset3 ¹
source variables other datasets.name	Source variable(s) from other datasets than filled out under source dataset	When using variables from multiple other datasets, use a comma-separated list, the order corresponding to the order of the datasets they are derived from specified under source variables from other datasets.dataset ¹
target *	Name of the target common data model	e.g. LifeCycle_CDM, LongITools_CDM, see variable explorer
target dataset *	Target dataset name.	Map to a dataset that is defined in a common data model
target variable *	Target variable name	Map to a variable that is defined in a common data model
match *	Whether the harmonisation is partial, complete or NA (non-existent)	Find list to choose from in CatalogueOntologies [StatusDetails]
description	Description of the harmonisation	
syntax	Syntax used for this harmonisation	

*Table 6. Description of the columns that can be filled out for Variable mappings. * = mandatory; 1 = see sheet Variable mappings in the [example template](#) for an example on how to fill this out (last line)

[Upload metadata](#)

When you have filled out the template(s) you can start uploading the metadata. When you log in to MOLGENIS Data Catalogue you will see a listing of databases that are accessible to you. Click on your cohort's database to access it. Go to 'Up/Download' in the menu. Use 'browse' to select a template and 'upload' to start uploading your metadata. After uploading, you can view your metadata under 'Tables'.

Please report any bugs or difficulties to molgenis-support.

Find harmonisations

When your data is uploaded to the Data Catalogue you can find your own harmonised variables in variable details in the [Variable Explorer] (<https://data-catalogue.molgenisccloud.org/catalogue/catalogue/#/variable-explorer/>) once they have been transferred there. Use the search bar to find your variable(s) of interest.

Variables (1)

Search: cohab

Details Harmonization

▼ Cohabitation Status Of The Mother (Age ≥0 Year And <1 Year) LifeCycle

[view details](#)

variable cohab_0

description Cohabitation status of the mother: are her and her partner living together as a couple? "Mother's partner" can be the biological partner, a new partner or a partner of the same gender. cohab_0: at birth or as near to birth as possible and within one year of birth.

unit -

format binary

n repeats 17

mapped by GECKO ALSPAC SWS NFBC86 ABCD ELFE GenR DNBC RHEA MoBa INMA CHOP NINFEA NFBC66 EDEN

Click on "Details"

Cohabitation status of the mother (age ≥0 year and <1 year)

Details Harmonization

name cohab_0

description Cohabitation status of the mother: are her and her partner living together as a couple? "Mother's partner" can be the biological partner, a new partner or a partner of the same gender. cohab_0: at birth or as near to birth as possible and within one year of birth.

unit -

format binary

n repeats 17

mapped by ABCD ALSPAC CHOP DNBC EDEN ELFE GECKO GenR INMA MoBa NFBC66 NFBC86 NINFEA RHEA SWS

	ALSPAC	GenR	BIB	CHOP	INMA	GECKO	NFBC86	MoBa	ELFE	EDEN	NFBC66	DNBC	HBCS	NINFEA	SWS	RAINE	RHEA	Borealis	CONSTANCES	DFBC	ELIPA	PEF	FimGeDi	FMC	FR	NOMA	OULU1935	OULU1945	PANIC	RCGP RSC	Rotterdam Study first cohort	Rotterdam Study second cohort	Rotterdam Study third cohort	SYSDIMET	UK Biobank	NFBC66 G0
cohab_0																																				
cohab_1																																				
cohab_2																																				
cohab_3																																				

Click on "Harmonization"

Cohabitation status of the mother (age ≥0 year and <1 year)	
Details	Harmonization
ABCD	ALSPAC CHOP DNBC EDEN ELFE GECKO GenR INMA MoBa NFBC66 NFBC86 NINFEA RHEA SWS
cohab_0	
description	Cohabitation status of the mother during pregnancy of shortly after birth: does mother live with a partner as a couple. This includes women who don't have a partner.
harmonization status	complete
variables used	QZW11
syntax	Recode QZW11 (missing = sysmis) (2=1) (1=2) (3=2) into cohab_0.
cohab_1	
description	-

[Request access \(catalogue\)](#)

If you do not have an account to upload data to the Data Catalogue yet, please email molgenis-support to apply for an account.

[CDM / Network data manager](#)

The Network data manager is responsible for uploading the common data model (CDM) to the Data Catalogue.

[Catalogue](#)

[Define CDM metadata](#)

[MOLGENIS Data Catalogue](#) (sometimes also called 'EMX2 catalogue') provides a framework to describe in detail: cohort metadata; definitions of the data variables collected (aka 'source variables'); and mappings to common data models (aka 'target variables'). Its purpose is to facilitate pooled data analysis of multiple cohorts.

- The cohort metadata provides descriptive information such as contact details, name of the cohort, and high-level summary of contents and cohort design.
- The 'source variables' can be considered as a codebook or data dictionary for a cohort (e.g. ALSPAC).
- Similarly, the common data model metadata (or 'target variables') can be considered the codebook for a network of cohorts working together (e.g. LifeCycle)
- The mappings describe how source variables have been converted into target variables as basis for integrated analysis.

This section explains how to submit the 'target variables' (also called the harmonised model or common data model) into the Data Catalogue. Expected users of this 'how to' are central data managers of networks such as LifeCycle or LongITools. You will need login details to upload metadata to MOLGENIS Data Catalogue.

[Define common data elements](#)

We use the [TargetDictionary template](#) to describe the common data model elements. The [TargetDictionary template](#) consists of multiple sheets. Each sheet corresponds to a table in the Data Catalogue. The columns in the sheet correspond to columns in the table concerned. This document describes how to fill out each of the sheets and their columns. A column with

an asterisk (*) after its name is mandatory, i.e., it should contain values for the system to accept a data upload. You can download this [filled out example](#) as a reference for filling out the template. Note that there is no sheet for *All variables*. This table is a generic listing of all variables entered for the cohort; it shows *Variables* and *Repeated variables* in one table.

It is good practice to try adding a few variables to the template first and see whether your upload succeeds. To upload the metadata to the Data Catalogue see the section [To upload the metadata to the Data Catalogue](#) see the section [Upload metadata](#) to the Data Catalogue.

Tables in 'testNetwork'

Download all tables: [zip](#) | [excel](#) | [jsonld](#) | [ttl](#)

Table	Description
All variables	Generic listing of all source variables. Should not be used directly, please use SourceVariables or RepeatedSourceVariables instead
Collection events	Definition of a data collection event for a resource
Contacts	
Datasets	Definition of a dataset within a (common) data model
Documentation	Documentation attached to a resource
Extended resources	
External identifiers	
Models	Data models
Networks	Collaborations of multiple institutions
Organisations	Research departments and research groups
Publications	Publications following bibtex format
Repeated variables	Definition of a repeated sourceVariable. Refers to another variable for its definition
Resources	Generic listing of all resources. Should not be used directly, instead use specific types such as Databanks and Studies
Subcohorts	Subcohorts defined for this resource
Variable values	Listing of categorical value+label definition in case of a categorical variable
Variables	Definition of a non-repeated variable, or of the first variable from a repeated range
Version	3.2

Figure 1. Tables in a Network's staging area in the Data Catalogue.

[Fill out network rich metadata](#)

Open your staging area, navigate to 'Tables' and open the table 'Networks'. Your network id and name are already filled out. Click on the pencil sign next to this entry to start editing your network rich metadata by filling out the form. The network's common data model should also be filled out in the same manner under 'Models'. 'Subcohorts' and 'Collection events' are filled out through the same route, by accessing the corresponding tables. Make sure to choose the right **model** id under resource when defining your 'Subcohorts' and 'Collection events', e.g. LongITools should select LongITools_CDM.

Define the common data model

Datasets sheet

The network's datasets are defined in the *Datasets* sheet. Columns with an asterisk (*) after their name are mandatory.

Column name	Description	Remarks
resource *	Id of the model .	e.g LifeCycle_CDM, LongITools_CDM or ATHLETE_CDM
name *	Unique dataset name	
label	Dataset label	
description	Dataset description	
unit of observation	Defines what each record in this dataset describes	
number of rows	Count of the number of records in this dataset	
keywords ¹	Enables grouping of datasets into topics and helps to display variables in a tree	Find list to choose from in CatalogueOntologies Keywords
since version	Version of the data model when this dataset was introduced	e.g. 1.0.0 or 2.1
until version	Version of the data model when this dataset was deleted	e.g. 2.0.0 or 2.1

Table 1. Description of the columns that can be filled out for Datasets. * = mandatory

Variables sheet

The network's variables are defined in the *Variables* sheet.

Column name	Description	Remarks
resource *	Id of the model that contains this variable	e.g LifeCycle_CDM, LongITools_CDM or ATHLETE_CDM
dataset *	Dataset that contains the variable	Datasets must be predefined in the <i>Datasets</i> sheet

name *	Variable name, unique within a dataset	
label	Human readable variable label	
format	The data type of the variable	Find list to choose from in CatalogueOntologies Formats
unit ¹	Unit in case of a continuous or integer format	Find list to choose from in CatalogueOntologies Units
description	Description of the variable	
keywords ¹	Enables grouping of variables into topics and displaying in a tree	Find list to choose from in CatalogueOntologies Keywords
example values	Examples of values in a comma separated list	Makes your data more insightful; e.g. 1,2,3 or TRUE,FALSE or 1.23,4.56,3.14
mandatory	Whether this variable is required within this collection	
vocabularies ¹	Refer to ontologies being used	Find list to choose from in CatalogueOntologies Vocabularies
collection event.resource	Refer to the resource that contains the collection event	e.g. LifeCycle
collection event.name	Refer to a collection event	e.g. y1 or y2
since version	Version of the data model when this variable was introduced	e.g. 1.0.0 or 2.1
until version	Version of the data model when this variable was deleted	e.g. 2.0.0 or 2.1

Table 2. Description of the columns that can be filled out for Variables. * = mandatory; 1 = contact molgenis-support@umcg.nl to add Vocabularies, Keywords or Units

[Variable values sheet](#)

The coding of categorical variables is defined in the *Variable values* sheet. This sheet is optional, but it is highly recommended to fill out the codes and values for your categorical variables, so that your data becomes more insightful for those that are interested.

Column name	Description	Remarks
resource *	Id of the model that contains this variable	e.g LifeCycle_CDM
variable.dataset *	Dataset that contains the variable	Datasets must be predefined in the <i>Datasets</i> sheet
variable.name *	Variable name	Variables must be predefined in the <i>Variables</i> sheet
value *	The code or value used	e.g. 1, 2 or -99
label *	The label corresponding to the value	e.g. 'yes', 'no' or 'NA'
order	The order in which the code list should appear	e.g. 1
is missing	Whether this value indicates a missing field	TRUE or FALSE
ontology term URI	Reference to an ontology term that defines this categorical value	e.g. http://purl.obolibrary.org/obo/DOID_1094
since version	Version of the data model when this variable value was introduced	e.g. 1.0.0 or 2.1
until version	Version of the data model when this variable value was deleted	e.g. 2.0.0 or 2.1

Table 3. Description of the columns that can be filled out for Variable values. * = mandatory

[Repeated variables sheet](#)

The *Repeated variables* sheet is optional. Variables that are repeats of a variable defined in the sheet *Variables* are defined in the *Repeated variables* sheet. Defining your repeated variables using this sheet will limit the amount of information that has to be repeated when filling out repeated variables. This sheet is optional.

Column name	Description	Remarks
resource *	Id of the model that contains this variable	e.g LifeCycle_CDM or ATHLETE_CDM
dataset *	Dataset name	e.g. core

name *	Variable name	e.g. height_1
label	Human readable variable label	
is repeat of.dataset *	Dataset that contains the variable that is repeated	Datasets must be predefined in the <i>Datasets</i> sheet; e.g. core
is repeat of.name *	Name of the variable that is repeated	Variables must be predefined in the <i>Variables</i> sheet; e.g. height_0
collection event.resource	Refer to the network that contains the collection event	e.g. LifeCycle
collection event.name	Refer to the name of a collection event	The collection event needs to be predefined via online forms, as described above; e.g. y1 or y2
since version	Version of the data model when this variable was introduced	e.g. 1.0.0 or 2.1
until version	Version of the data model when this variable was deleted	e.g. 2.0.0 or 2.1

Table 4. Description of the columns that can be filled out for Repeated variables. * = mandatory

[Request access](#)

Send an email to molgenis-support@umcg.nl to apply for an account to upload metadata to the Data Catalogue.

[Upload metadata](#)

When you log in to MOLGENIS Data Catalogue you will see a listing of databases that are accessible to you. Click on your network's database to access it. Go to 'Up/Download' in the menu. Use 'browse' to select a template and 'upload' to start uploading your metadata. After uploading you can view your metadata under 'Tables'.

Annexe 2: Catalogue metadata model

The catalogue consists of the following main tables:

- [Table: All variables](#)
- [Table: Collection events](#)
- [Table: Contacts](#)
- [Table: Datasets](#)
- [Table: Documentation](#)
- [Table: External identifiers](#)
- [Table: Linked resources](#)
- [Table: Publications](#)
- [Table: Quantitative information](#)
- [Table: Resources](#)
- [Table: Subcohorts](#)
- [Table: Variable mappings](#)
- [Table: Variable values](#)

Table: All variables

Overview and relationships:

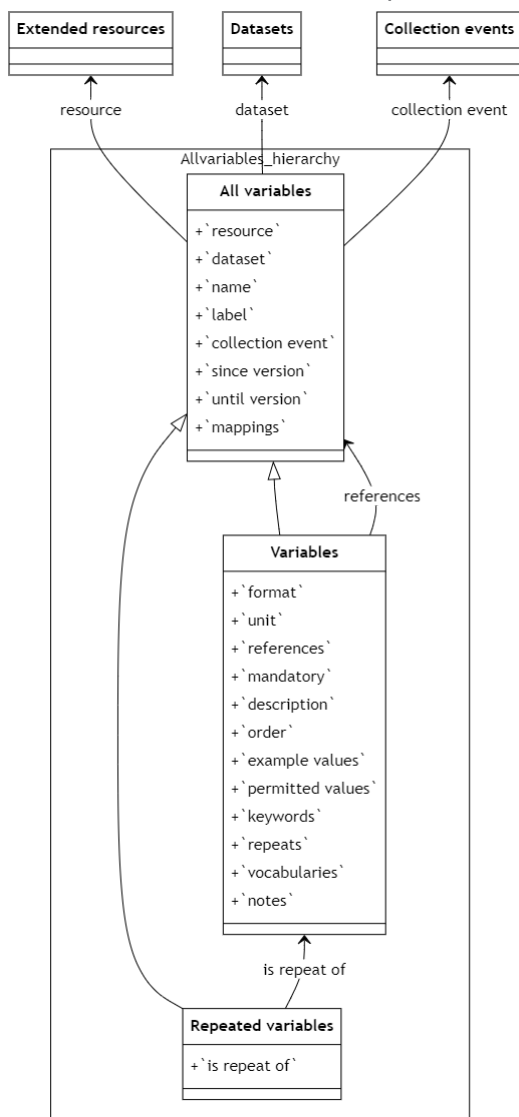


Table definition:

Generic listing of all source variables. Should not be used directly, please use SourceVariables or RepeatedSourceVariables instead

Extended table definitions:

Table 'All variables' has the following subclasses/specializations:

All variables

Generic listing of all source variables. Should not be used directly, please use SourceVariables or RepeatedSourceVariables instead

Repeated variables (extends: All variables) – *Definition of a repeated sourceVariable. Refers to another variable for its definition*

Variables (extends: All variables) – *Definition of a non-repeated variable, or of the first variable from a repeated range*

Column definitions:

resource – *Data source that this variable was collected in*

domain: All variables

constraints: ref(Extended resources) required

dataset – *Dataset this variable is part of*

domain: All variables

constraints: ref(Datasets, refLink=resource) required

name – *name of the variable, unique within a table*

domain: All variables

constraints: string required

label – *Human friendly longer name, if applicable*

domain: All variables

constraints: string

collection event – *in case of protocolised data collection this defines the moment in time this variable is collected on*

domain: All variables

constraints: ref(Collection events)

since version – *When this variable was introduced*

domain: All variables

constraints: string

until version – *When this variable was removed if applicable*

domain: All variables

constraints: string

format – *Data type, e.g. string,int,decimal,date,datetime etc*

domain: Variables

constraints: ontology(CatalogueOntologies.Formats)

unit

domain: Variables

constraints: ontology(CatalogueOntologies.Units)

references – *to define foreign key relationships between variables within or across tables*

domain: Variables

constraints: ref(All variables, refLink=resource)

mandatory – *whether this variable is required within this collection*

domain: Variables

constraints: bool

description

domain: Variables
constraints: text

order – *to sort variables you can optionally add an order value*

domain: Variables
constraints: int

example values

domain: Variables
constraints: string_array

permitted values

domain: Variables
constraints: reback(Variable values, refBack=variable)

keywords

domain: Variables
constraints: ontology_array(CatalogueOntologies.Keywords)

repeats – *listing of all repeated variables defined for this variable*

domain: Variables
constraints: reback(Repeated variables, refBack=is repeat of)

vocabularies

domain: Variables
constraints: ontology_array(CatalogueOntologies.Vocabularies)

notes – *Any other information on this variable*

domain: Variables
constraints: text

mappings – *in case of protocolised data collection this defines the moment in time this variable is collected on*

domain: All variables
constraints: reback(Variable mappings, refBack=target variable)

is repeat of – *reference to the definition of the sourceVariable that is being repeated*

domain: Repeated variables
constraints: ref(Variables, refLink=resource) required

Table: Collection events

Overview and relationships:

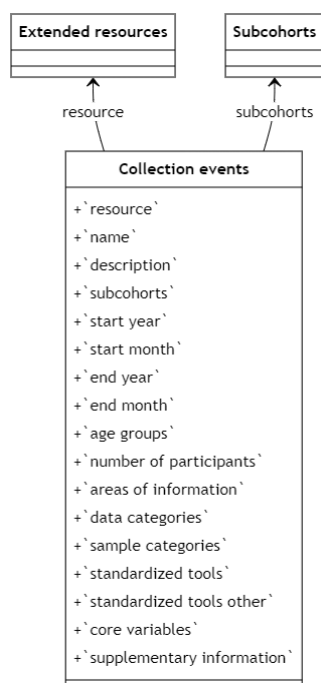


Table definition:

Definition of a data collection event for a resource

Column definitions:

resource – *Resource this collection event is part of*
constraints: ref(Extended resources) required

name – *Name of the collection event*
constraints: string required

description – *Description of the collection event*
constraints: text

subcohorts – *Subcohorts that are targetted by this collection event*
constraints: ref_array(Subcohorts, refLink=resource)

start year – *Start year of data collection*
constraints: ontology(CatalogueOntologies.Years)

start month – *Start month of data collection*
constraints: ontology(CatalogueOntologies.Months)

end year – *End year of data collection. Leave empty if collection is ongoing*
constraints: ontology(CatalogueOntologies.Years)

end month – *End month of data collection. Leave empty if collection is ongoing*

constraints: ontology(CatalogueOntologies.Months)

age groups – *Age groups included in this data collection event*

constraints: ontology_array(CatalogueOntologies.Age groups)

number of participants – *Number of participants sampled in this data collection event*

constraints: int

areas of information – *Areas of information that were extracted in this data collection event*

constraints: ontology_array(CatalogueOntologies.Areas of information cohorts)

data categories – *Methods of data collection used in this collection event*

constraints: ontology_array(CatalogueOntologies.Data categories)

sample categories – *Samples that were collected in this collection event*

constraints: ontology_array(CatalogueOntologies.Sample categories)

standardized tools – *Standardized tools, e.g. surveys, questionnaires, instruments used to collect data for this collection event*

constraints: ontology_array(CatalogueOntologies.Standardized tools)

standardized tools other – *If 'other,' please specify*

constraints: string

core variables – *Name 10-20 relevant variables that were collected in this collection event*

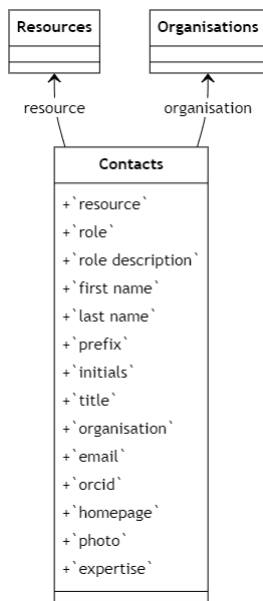
constraints: string_array

supplementary information – *Any other information that needs to be disclosed for this collection event*

constraints: text

Table: Contacts

Overview and relationships:



Column definitions:

resource – *Resource the contact is affiliated with*
constraints: ref(Resources) required

role – *Type(s) of contribution or role in the resource*
constraints: ontology_array(CatalogueOntologies.Contribution types)

role description – *Description of the role*
constraints: text

first name – *First name of the contact person*
constraints: string required

last name – *Last name of the contact person*
constraints: string required

prefix – *Surname prefix, if applicable*
constraints: string

initials – *Initials of the contact person*
constraints: string

title – *Title of the contact person*
constraints: ontology(CatalogueOntologies.Titles)

organisation – *Affiliated organisation of the contact person*
constraints: ref(Organisations)

email – *Contact's email address*
constraints: string

orcid – *Orcid of the contact person*
constraints: string

homepage – *Link to contact's homepage*
constraints: string

photo – *Contact's photograph*
constraints: file

expertise – *Description of contact's expertise*
constraints: string

Table: Datasets

Overview and relationships:

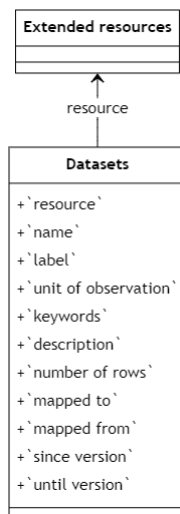


Table definition:

Definition of a dataset within a (common) data model

Column definitions:

resource – *resources that these variables are part of*
constraints: ref(Extended resources) required

name – *unique dataset name in the model*
constraints: string required

label – *short human readable description*
constraints: string

unit of observation – *defines what each record in this table describes*
constraints: ontology(CatalogueOntologies.Observation targets)

keywords – *enables grouping of table list into topic and to display tables in a tree*
constraints: ontology_array(CatalogueOntologies.Keywords)

description – *description of the role/function of this table*
constraints: text

number of rows – *count of the number of records in this table*
constraints: int

mapped to – *common dataset models this dataset has been mapped into*
constraints: refback(Dataset mappings, refBack=source dataset)

mapped from – *source datasets that have been mapped to this harmonized dataset*
constraints: refback(Dataset mappings, refBack=target dataset)

since version – *When this dataset was introduced*
constraints: string

until version – *When this dataset was removed if applicable*
constraints: string

Table: Documentation

Overview and relationships:

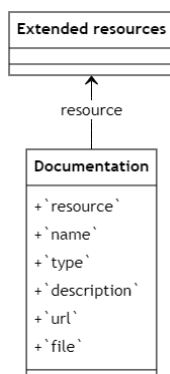


Table definition:

Documentation attached to a resource

Column definitions:

resource – *The resource this documentation is for*
constraints: ref(Extended resources) required

name – *Document name*
constraints: string required

type – *Type of documentation*
constraints: ontology(CatalogueOntologies.Document types)

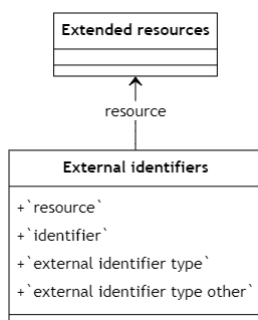
description – *Description of the document*
constraints: text

url – *Hyperlink to the source of the documentation*
constraints: string

file – *Optional file attachment containing the documentation*
constraints: file

Table: External identifiers

Overview and relationships:



Column definitions:

resource – *Resource that this external identifier belongs to*
constraints: ref(Extended resources) required

identifier – *External identifier*
constraints: text required

external identifier type – *External identifier type*
constraints: ontology(CatalogueOntologies.External identifier types)

external identifier type other – *If other, enter external identifier type*
constraints: text

Table: Linked resources

Overview and relationships:

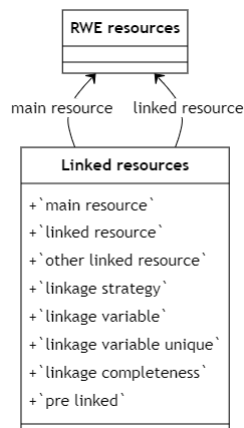


Table definition:

Links between datasource and databank

Column definitions:

main resource

constraints: ref(RWE resources) required

linked resource

constraints: ref(RWE resources) required

other linked resource – *If other linked data source, enter the name of the data source*

constraints: text

linkage strategy – *The linkage method that was used to link data banks. One entry per data bank*

constraints: ontology(CatalogueOntologies.Linkage strategies)

linkage variable – *If a single variable (or linkage key) is used to link a data bank to others, a name and description of the variable is provided. One entry per data bank*

constraints: text

linkage variable unique – *If a single variable is used to link a data bank to others, is the variable a unique identifier? One entry per data bank*

constraints: bool

linkage completeness – *Provide a high-level description of the completeness of linkages that are currently available between data banks in the data source (max 100 words)*

constraints: text

pre linked – *Does the data source constitute of linked data sources?*

constraints: bool

Table: Publications

Overview and relationships:

Publications
+ `doi`
+ `title`
+ `authors`
+ `year`
+ `journal`
+ `volume`
+ `number`
+ `pagination`
+ `publisher`
+ `school`
+ `abstract`
+ `resources`

Table definition:
Publications following bibtex format

Column definitions:

doi – *Digital object identifier*
constraints: string required

title – *Publication title*
constraints: text

authors – *List of authors, one entry per author*
constraints: string_array

year – *Year of publication (or, if unpublished, year of creation)*
constraints: int

journal – *Journal or magazine the work was published in*
constraints: string

volume – *Journal or magazine volume*
constraints: int

number – *Journal or magazine issue number*
constraints: int

pagination – *Page numbers, separated either by commas or double-hyphens*
constraints: string

publisher – *Publisher's name*
constraints: string

school – *School where the thesis was written (in case of thesis)*
constraints: string

abstract – *Publication abstract*
constraints: text

resources – *List of resources that refer to this publication*
constraints: refback(Extended resources, refBack=publications)

Table: Quantitative information

Overview and relationships:

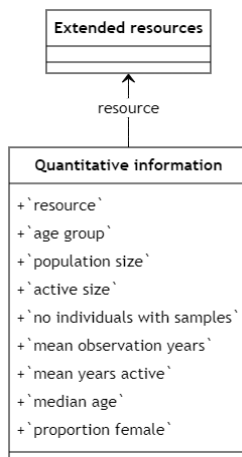


Table definition:

Quantitative information on the resource

Column definitions:

resource

constraints: ref(Extended resources) required

age group – *Select the relevant age group for this quantitative information*

constraints: ontology(CatalogueOntologies.Age groups) required

population size – *Total number of unique individuals with records captured in the data source (most recent count). In the catalogue, this will accommodate counts per year*

constraints: int

active size – *Number of unique, active, or currently registered individuals with records captured in the data source (most recent count). In the catalogue, this will accommodate counts per year*

constraints: int

no individuals with samples – *Number of unique individuals with records of biological samples (e.g., blood, urine) (most recent count). In the catalogue, this will accommodate counts per year*

constraints: int

mean observation years – *Median years for which unique individuals with records captured in the data source are observable (most recent count)*

constraints: int

mean years active – *Median time for which unique individuals with records captured in the data source are observable (most recent count)*

constraints: int

median age – *Median age of individuals within data source*

constraints: int

proportion female – *Proportion of females in the data source*

constraints: int

Table: Resources

Overview and relationships:

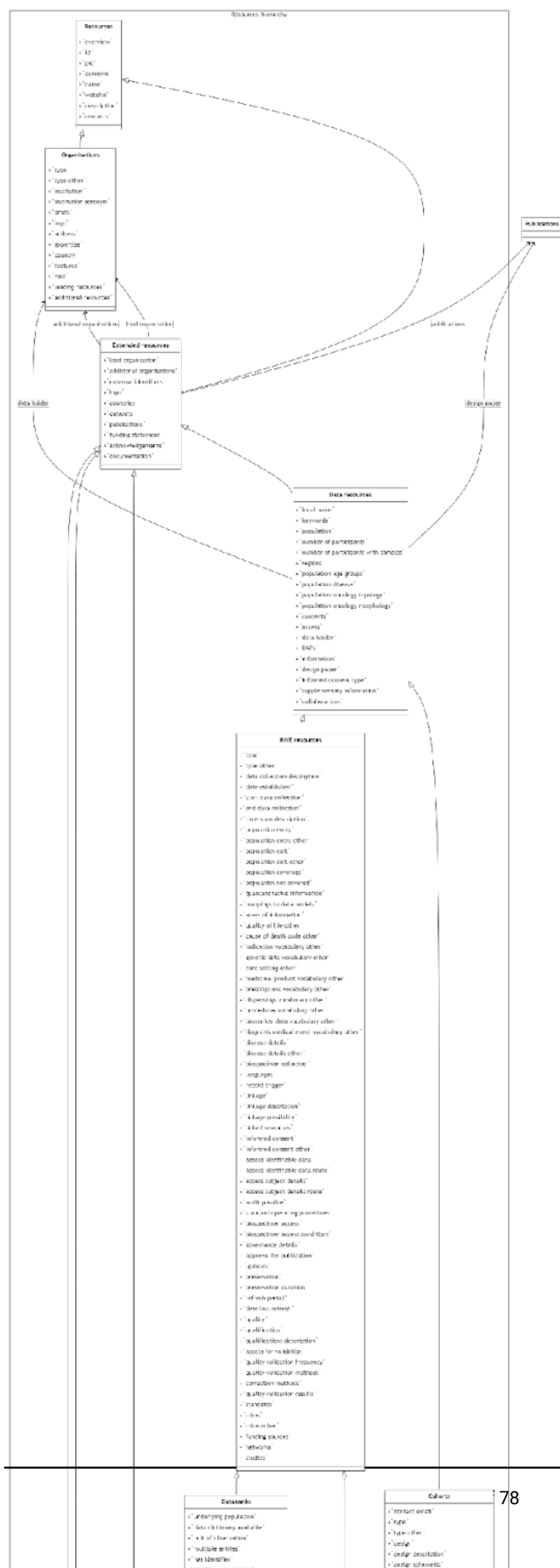


Table definition:

Generic listing of all resources. Should not be used directly, instead use specific types such as Databanks and Studies

Extended table definitions:

Table 'Resources' has the following subclasses/specializations:

Resources

Generic listing of all resources. Should not be used directly, instead use specific types such as Databanks and Studies

Extended resources (extends: Resources)

Organisations (extends: Resources) – *Research departments and research groups*

Data resources (extends: Extended resources) – *Resources for data*

Models (extends: Extended resources) – *Data models*

Networks (extends: Extended resources) – *Collaborations of multiple institutions*

Studies (extends: Extended resources) – *Collaborations of multiple institutions, addressing research questions using data sources and/or data banks*

Cohorts (extends: Data resources) – *Group of individuals sharing a defining demographic characteristic*

RWE resources (extends: Data resources) – *Real world data collections*

Data sources (extends: RWE resources) – *Collections of multiple data banks covering the same population*

Databanks (extends: RWE resources) – *Data collection from real world databases such as health records, registries*

Column definitions:

heading: overview – *General information*

id – *Internal identifier*

domain: Resources

constraints: string required

pid – *Persistent identifier*

domain: Resources

constraints: string

acronym – *Acronym if applicable*

domain: Resources

constraints: string

name – *Name used in European projects*

domain: Resources

constraints: text required

local name – *If different from above, name in the national language*

domain: Data resources

constraints: string

type – *Type of organisation; in which sector is the organisation active?*

domain: Organisations

constraints: ontology_array(CatalogueOntologies.Organisation types)

type other – *If type is 'other', a description of type of organisation*

domain: Organisations

constraints: text

institution – *University, company, medical centre or research institutes this organisation is part of*

domain: Organisations

constraints: text

institution acronym – *Short name of the organisation*

domain: Organisations

constraints: string

email – *Contact email address for person responsible for organization entry in catalogue*

domain: Organisations

constraints: email

logo – *Logo of the organisation*

domain: Organisations

constraints: file

address – *Address of the organisation*

domain: Organisations

constraints: text

expertise – *A short description of the expertise of this institution*

domain: Organisations

constraints: text

country – *Country in which the institution head office or coordinating centre is located*

domain: Organisations

constraints: ontology_array(CatalogueOntologies.Countries)

features – *Features that describe this organisation*

domain: Organisations

constraints: ontology_array(CatalogueOntologies.Organisation features)

role – *Roles of the institution in connection with data sources in the catalogue. Select one or more of the following:*

domain: Organisations

constraints: ontology_array(CatalogueOntologies.Organisation roles)

leading resources – *Listing of data sources, cohorts, studies (etc)*

domain: Organisations

constraints: refback(Extended resources, refBack=lead organisation)

additional resources – *Listing of data sources, cohorts, studies (etc)*

domain: Organisations

constraints: refback(Extended resources, refBack=additional organisations)

type – *Type of network, e.g. h2020 project*

domain: Networks

constraints: ontology_array(CatalogueOntologies.Network types)

features – *Characterizations of the network*

domain: Networks

constraints: ontology_array(CatalogueOntologies.Network features)

type – *Select 1 of the following types of study*

domain: Studies

constraints: ontology(CatalogueOntologies.Study types)

type other – *If other, describe the type of study*

domain: Studies

constraints: string

type – *Which of the following families of databanks best describe this data source*

domain: RWE resources

constraints: ontology_array(CatalogueOntologies.Datasource types)

type other – *If other, describe the type of datasource*

domain: RWE resources

constraints: text

keywords – *Keywords to increase findability of this resource. Try to use words that are not used in the description*

domain: Data resources

constraints: text

website – *Link to the website or homepage*

domain: Resources

constraints: hyperlink

lead organisation – *lead organisation (e.g. research department or group) for this resource*

domain: Extended resources

constraints: ref_array(Organisations)

additional organisations – *List the names of any additional organisations that contributed to the resource*

domain: Extended resources

constraints: ref_array(Organisations)

description – *Short description*

domain: Resources

constraints: text

data collection description – *Describe the process of collection and recording of data.*

domain: RWE resources

constraints: text

date established – *Date when the data source was first established. If the exact day of the month is not known, please enter 15. If the exact month is not known then please enter 15/06*

domain: RWE resources

constraints: date

start data collection – *The date when data started to be collected or extracted. If the exact day of the month is not known, please enter 15. If the exact month is not known then please enter 15/06*

domain: RWE resources

constraints: date

end data collection – *If data collection in the data source has ceased, on what date did new records last enter the data source?. If the exact day of the month is not known, please enter 15. If the exact month is not known then please enter 15/06*

domain: RWE resources

constraints: date

time span description – *Description of time span*

domain: RWE resources

constraints: text

external identifiers – *External identifier(s) for this resource (e.g. EUPASS number, UMCC register number)*

domain: Extended resources

constraints: refback(External identifiers, refBack=resource)

status – *Status of the study*

domain: Studies

constraints: ontology(CatalogueOntologies.Study status)

release frequency – *Refreshing rate (in months)*

domain: Models

constraints: int

contact email – *Contact e-mail address for this cohort*

domain: Cohorts

constraints: string

contacts – *Contact person(s)*

domain: Resources

constraints: refback(Contacts, refBack=resource)

type – *Type of resource, e.g. registry, cohort, biobank*

domain: Cohorts

constraints: ontology_array(CatalogueOntologies.Resource types)

type other – *If other, describe the type of resource*

domain: Cohorts

constraints: string

design – *The study design of this cohort, i.e. cross-sectional or longitudinal*

domain: Cohorts

constraints: ontology(CatalogueOntologies.Cohort designs)

design description – *Short description of the study design of this cohort*

domain: Cohorts

constraints: text

design schematic – *A schematic depiction of the study design of this cohort*

domain: Cohorts

constraints: file

collection type – *The data collection type of this cohort, i.e. retrospective or prospective; if both, select both*

domain: Cohorts

constraints: ontology_array(CatalogueOntologies.Collection types)

logo – *Logo of the resource, for use on homepages etc.*

domain: Extended resources

constraints: file

heading: population – *Description of the population that can potentially be captured in the resource*

number of participants – *Total number of individuals for which data is collected*

domain: Data resources

constraints: int

number of participants with samples – *Number of individuals for which samples are collected*

domain: Data resources

constraints: int

underlying population – *Provide a summary description of the underlying population (maximum 100 words) or URL to a description*

domain: Databanks

constraints: text

countries – *Countries where data from this resource largely originate from*

domain: Extended resources

constraints: ontology_array(CatalogueOntologies.Countries)

regions – *Geographical regions where data from this resource largely originate from*

domain: Data resources

constraints: ontology_array(CatalogueOntologies.Regions)

population age groups – *Which population age groups are captured in this resource? Select all that are relevant.*

domain: Data resources

constraints: ontology_array(CatalogueOntologies.Age groups)

inclusion criteria – *Inclusion criteria applied to the participants of this resource*

domain: Cohorts

constraints: ontology_array(CatalogueOntologies.Inclusion criteria)

other inclusion criteria – *Other inclusion criteria applied to the participants of this resource*

domain: Cohorts

constraints: text

start year – *Year when first data was collected*

domain: Cohorts

constraints: int

end year – *Year when last data was collected. Leave empty if collection is ongoing*

domain: Cohorts

constraints: int

population entry – *Select the possible causes / events that trigger the registration of a person in the data source*

domain: RWE resources

constraints: ontology_array(CatalogueOntologies.Population entry)

population entry other – *If other, specify the causes of entry to the underlying population*

domain: RWE resources

constraints: text

population exit – *Select the possible causes / events that trigger the de-registration of a person in the data source*

domain: RWE resources

constraints: ontology_array(CatalogueOntologies.Population exit)

population exit other – *If other, specify the causes of exit from the underlying population*

domain: RWE resources
constraints: text

population disease – *Does the resource collect information on a specific disease subpopulation (e.g., as in a disease-specific registry)?*

domain: Data resources
constraints: ontology_array(CatalogueOntologies.Diseases)

population oncology topology – *Does the resource collect information on specific cancer subtype(s)? If yes, select topology specifications.*

domain: Data resources
constraints: ontology_array(CatalogueOntologies.ICDO topologies)

population oncology morphology – *Does the resource collect information on specific cancer subtype(s)? If yes, select morphology specifications.*

domain: Data resources
constraints: ontology_array(CatalogueOntologies.ICDO morphologies)

population coverage – *Estimated percentage of the population covered by the data source in the catchment area. Please describe the denominator.*

domain: RWE resources
constraints: text

population not covered – *Description of the population covered by the data source in the catchment area whose data are not collected, where applicable (e.g.: people who are registered only for private care)*

domain: RWE resources
constraints: text

quantitative information – *Numerical summaries describing data bank population*

domain: RWE resources
constraints: refback(Quantitative information, refBack=resource)

subcohorts – *List of subcohorts or subpopulations for this resource*

domain: Cohorts
constraints: refback(Subcohorts, refBack=resource)

heading: contents – *Data model and contents*

datasets

domain: Extended resources
constraints: refback(Datasets, refBack=resource)

areas of information – *Areas of information that were collected*

domain: RWE resources
constraints: ontology_array(CatalogueOntologies.Areas of information ds)

data dictionary available – *Are a data dictionary and a data model available?*

domain: Databanks
constraints: bool

biospecimen collected – *If the data bank contains biospecimens, what types of specimen*

domain: RWE resources

constraints: ontology_array(CatalogueOntologies.Biospecimens)

languages – *Languages in which that the records are recorded (in ISO 639, https://en.wikipedia.org/wiki/List_of_ISO_639-1_codes)*

domain: RWE resources

constraints: ontology_array(CatalogueOntologies.Languages)

record trigger – *What triggers the creation of a record in the data bank? e.g., hospital discharge, specialist encounter, dispensation of a medicinal product, recording of a congenital anomaly*

domain: RWE resources

constraints: text

collection events – *List of collection events defined for this resource*

domain: Cohorts

constraints: refback(Collection events, refBack=resource)

unit of observation – *Based on the prompt, what is the unit of observation of a record (e.g., person, prescription)?*

domain: Databanks

constraints: text

multiple entries – *Can there be multiple entries for a single person in the data bank? For example, may a person contribute multiple records to the data bank?*

domain: Databanks

constraints: bool

heading: linkage – *Data linkage*

has identifier – *Is there a unique identifier for a person in the data bank?*

domain: Databanks

constraints: bool

identifier description – *Describe the variable that is used as a unique identifier for a person in the data bank? If the unique identifier is not at level of a person (for example hospital encounter), describe how this translated to an individual level*

domain: Databanks

constraints: text

linkage description – *Provide a high-level description of the linkages that are either: currently available between data sources in the data source (when pre-linked = yes); linkages that are possible when using the data source*

domain: RWE resources

constraints: text

linkage possibility – *Can this data source be linked to other data sources?*

domain: RWE resources

constraints: bool

linked resources – *List of resources that are linked into this main resource*

domain: RWE resources

constraints: refback(Linked resources, refBack=main resource)

release type – *Select whether this resource is a closed dataset or whether new data is released continuously or at a termly basis*

domain: Cohorts

constraints: ontology(CatalogueOntologies.Release types)

release description – *Description of the release cycle of this resource*

domain: Cohorts

constraints: text

linkage options – *Linkage options with additional data sources that are available for this resource*

domain: Cohorts

constraints: text

heading: access – *Access and validation information*

data holder – *The name of the organisation that is responsible for governance of the data bank*

domain: Data resources

constraints: ref(Organisations)

reason sustained – *Description of the reason why the data bank is sustained by the organisation (e.g., for surveillance, clinical purposes, financial or administrative purposes, research purposes)*

domain: Databanks

constraints: text

data access conditions – *Codes defining data access terms and conditions*

domain: Cohorts

constraints: ontology_array(CatalogueOntologies.Data access conditions)

data use conditions – *Codes defining data use terms and conditions*

domain: Cohorts

constraints: ontology_array(CatalogueOntologies.Data use conditions)

data access conditions description – *Description of data access terms and use conditions*

domain: Cohorts

constraints: text

data access fee – *Does a fee apply to gain access to data of this cohort?*

domain: Cohorts

constraints: bool

informed consent – *Is informed consent required for use of the data for research purposes?*

domain: RWE resources

constraints: ontology(CatalogueOntologies.Informed consents)

informed consent other – *If other, describe the conditions when informed consent is required*

domain: RWE resources
constraints: text

access identifiable data – *Can identifiable data be accessed in the data bank (including patient/practitioner name/practice name)?*

domain: RWE resources
constraints: text

access identifiable data route – *If yes above, what is the route to access or process this information? What permission is required?*

domain: RWE resources
constraints: text

access subject details – *Can individual patients/practitioners/practices be contacted in the data bank?*

domain: RWE resources
constraints: bool

access subject details route – *If yes above, what is the route to access or process this information? What permission is required?*

domain: RWE resources
constraints: text

audit possible – *Are external parties allowed to audit the data? For example, is it possible for an external party to audit the quality or validity of the data source?*

domain: RWE resources
constraints: bool

access third party – *Can (an extract of) the data bank be accessed with permission by a third party?*

domain: Databanks
constraints: bool

access third party conditions – *If above is 'yes', describe the conditions under which third-party access may be granted*

domain: Databanks
constraints: text

access non EU – *Can (an extract of) the data bank be accessed with permission by a non-EU/EEA institution?*

domain: Databanks
constraints: bool

access non EU conditions – *If yes above, describe the conditions under which non-EU/EEA access may be granted*

domain: Databanks
constraints: text

standard operating procedures – *Is there a standard operating procedure document that defines the processes and procedures for data capture and management?*

domain: RWE resources

constraints: bool

biospecimen access – *If the data bank contains biospecimens (e.g., tissue samples), can these be retrieved?*

domain: RWE resources

constraints: bool

biospecimen access conditions – *If yes above, describe the conditions under which permission to retrieve biospecimens may be granted*

domain: RWE resources

constraints: text

governance details – *If available, provide a link to documents or webpages that describe the overall governance of the data source bank (governing data access or utilisation for research purposes by existing DAPs)*

domain: RWE resources

constraints: text

approval for publication – *Is an approval needed to publish the results of a study using the data*

domain: RWE resources

constraints: bool

heading: updates – *Information on the regularity of updates and time lags*

refresh – *Average number of days between refresh of data bank with new records*

domain: Databanks

constraints: int

lag time – *How many days is the lag time after refresh before a record can be extracted? (e.g., a lag time may occur if the originator conducts quality checks)*

domain: Databanks

constraints: int

preservation – *Are records preserved in the data bank indefinitely?*

domain: RWE resources

constraints: bool

preservation duration – *If no to the above, for how long (in years) are records preserved in the data bank?*

domain: RWE resources

constraints: int

refresh period – *If data are refreshed on fixed dates (e.g., every June and December), when are the refreshes scheduled? Select all that apply from the following:*

domain: RWE resources

constraints: ontology_array(CatalogueOntologies.Refresh periods)

date last refresh – *Date of last update/refresh*

domain: RWE resources

constraints: date

heading: quality – *List of relevant studies conducted using the data bank*

qualification – *Has the data source successfully undergone a formal qualification process (e.g., from the EMA, or ISO or other certifications)?*

domain: RWE resources

constraints: bool

qualifications description – *Has the resource successfully undergone a qualification process (e.g., from the EMA)? If yes, describe the qualification(s) granted*

domain: RWE resources

constraints: text

number of records – *Total number of unique records captured in the data bank (most recent count)*

domain: Databanks

constraints: int

completeness – *Describe the completeness of the data bank (e.g., variables with more or fewer missing values)*

domain: Databanks

constraints: text

completeness over time – *Describe any changes in completeness of the data bank (e.g., variables with more or fewer missing values) that have occurred over time*

domain: Databanks

constraints: text

completeness results – *What methods or processes are applied to check completeness of the data bank?*

domain: Databanks

constraints: text

quality description – *Describe the quality of the data bank (e.g., variables with more or fewer missing values)*

domain: Databanks

constraints: text

quality over time – *Describe any changes in quality of the data bank that have occurred over time*

domain: Databanks

constraints: text

access for validation – *Can validity of the data in the data bank be verified, e.g., by review of origin medical charts?*

domain: RWE resources

constraints: bool

quality validation frequency – *How often are data quality checks and validation steps conducted on the data bank?*

domain: RWE resources

constraints: text

quality validation methods – *What methods or processes are applied for data quality checks and validation steps conducted on the data bank?*

domain: RWE resources
constraints: text

correction methods – *What methods or processes are applied to correct illogical values in the data bank?*

domain: RWE resources
constraints: text

quality validation results – *If available, provide a link to a publication of the data quality check and validation results*

domain: RWE resources
constraints: string

heading: standards – *Use of standard data models and ontologies*

ETL standard vocabularies – *Are data mapped to standardised vocabularies during ETL to the CDM? If yes, what vocabularies are used for events, such as diagnoses?*

domain: Databanks
constraints: ontology_array(CatalogueOntologies.Vocabularies)

ETL standard vocabularies other – *If other, what other vocabularies are used?*

domain: Databanks
constraints: text

heading: information – *Other information*

design paper – *Publication(s) that describe(s) the design of this resource*

domain: Data resources
constraints: ref_array(Publications)

publications – *Other publication(s) about this resource*

domain: Extended resources
constraints: ref_array(Publications)

informed consent type – *What type of informed consent was given for data collection?*

domain: Data resources
constraints: ontology(CatalogueOntologies.Informed consent types)

funding sources – *Specify the main financial support sources for the data source in the last 3 years. Select all that apply*

domain: RWE resources
constraints: ontology_array(CatalogueOntologies.Funding types)

funding scheme – *The source of funding for the study. Select all that apply*

domain: Studies
constraints: ontology_array(CatalogueOntologies.Study funding)

funding statement – *Statement listing funding that was obtained for this resource*

domain: Extended resources
constraints: text

acknowledgements – *Acknowledgement statement and citation regulation for this resource*

domain: Extended resources
constraints: text

documentation – *Descriptive document(s) available for this resource, e.g. informed consent*

domain: Extended resources
constraints: refback(Documentation, refBack=resource)

supplementary information – *Any other information that needs to be disclosed for this resource*

domain: Data resources
constraints: text

heading: collaborations – *List of relevant collaborations*

studies – *Listing of studies that used this cohort*

domain: Cohorts
constraints: refback(Studies, refBack=cohorts)

networks – *The consortia or networks that this cohort is involved in*

domain: Cohorts
constraints: refback(Networks, refBack=cohorts)

networks – *The consortia or networks that this study is part of*

domain: Studies
constraints: ref_array(Networks)

networks other – *List the names of any other networks that are not listed and this resource is involved in*

domain: Studies
constraints: text

networks – *List of networks that this datasource is associated with*

domain: RWE resources
constraints: refback(Networks, refBack=data sources)

studies – *List of studies that this datasource is associated with*

domain: RWE resources
constraints: refback(Studies, refBack=data sources)

data sources

domain: Networks
constraints: ref_array(Data sources)

databanks

domain: Networks
constraints: ref_array(Databanks)

cohorts

domain: Networks
constraints: ref_array(Cohorts)

models – *The common data model(s) used by this network*

domain: Networks
constraints: ref_array(Models)

studies

domain: Networks
constraints: refback(Studies, refBack=networks)

study requirements – *Study requirements*

domain: Studies
constraints: ontology_array(CatalogueOntologies.Study requirements)

regulatory procedure number – *Regulatory procedure number, for RMP Category 1 and 2 studies only*

domain: Studies
constraints: string

heading: data – *Data management***data sources** – *Data sources that provided data into this study*

domain: Studies
constraints: ref_array(Data sources)

data sources other – *Other not listed data sources that provided data into this study*

domain: Studies
constraints: text

databanks – *Databanks that provided data into this study*

domain: Studies
constraints: ref_array(Databanks)

databanks other – *Other not listed databanks that provided data into this study*

domain: Studies
constraints: text

cohorts – *Cohorts that provided data into this study*

domain: Studies
constraints: ref_array(Cohorts)

cdms – *Common data model(s) used in this study*

domain: Studies
constraints: refback(Mappings, refBack=source)

study features – *Data features*

domain: Studies
constraints: ontology_array(CatalogueOntologies.Study features)

data characterisation details – *Provide a summary description of the data characterisation or quality check process*

domain: Studies
constraints: text

data source types – *Types of data sources used*

domain: Studies

constraints: ontology_array(CatalogueOntologies.Study datasource types)

data source types other – *Sources of data, if other*

domain: Studies

constraints: text

quality marks – *Quality marks, such as ENCePP seal*

domain: Studies

constraints: ontology_array(CatalogueOntologies.Study quality marks)

number of data sources – *Total number of data sources included in the study*

domain: Studies

constraints: string

data extraction date – *Date on which the study data was extracted*

domain: Studies

constraints: date

heading: methods – *Methodological aspects*

study setting – *A short description of the study setting*

domain: Studies

constraints: text

analysis plan – *A brief summary of the analysis method (e.g. risk estimation, measures of risk, internal/external validity)*

domain: Studies

constraints: text

population description – *A short description of the study population*

domain: Studies

constraints: text

number of subjects – *Estimated number of subjects*

domain: Studies

constraints: int

age groups – *Which population age groups are studied*

domain: Studies

constraints: ontology_array(CatalogueOntologies.Age groups)

objectives – *A short description of the study objective*

domain: Studies

constraints: text

interventions – *A short description of the study interventions*

domain: Studies

constraints: text

comparators – *A short description of the study comparators*

domain: Studies

constraints: text

outcomes – *A short description of the study outcomes*

domain: Studies

constraints: text

study design – *A brief summary of the study design*

domain: Studies

constraints: text

results – *A brief summary of the results of the study on study completion (from abstract)*

domain: Studies

constraints: text

topic – *An initial classification of the study purpose*

domain: Studies

constraints: ontology_array(CatalogueOntologies.Study topics)

topic other – *If the study is not concerning any of the proposed categories, please specify the details*

domain: Studies

constraints: text

trial regulatory scope – *Classification of the clinical trial in relation to the medicines authorisation*

domain: Studies

constraints: ontology_array(CatalogueOntologies.Study trial regulatory scopes)

study design classification – *Study design classifications*

domain: Studies

constraints: ontology_array(CatalogueOntologies.Study design classification)

study design classification other – *Further details on design*

domain: Studies

constraints: text

study scope – *Scope of the study*

domain: Studies

constraints: ontology_array(CatalogueOntologies.Study scopes)

study scope other – *If scope 'other'*

domain: Studies

constraints: text

population of interest – *If population of interest is 'Other', please specify which other population has been studied*

domain: Studies

constraints: ontology_array(CatalogueOntologies.Population of interest)

population of interest other – *If population of interest is 'Other', please specify which other population has been studied*

domain: Studies

constraints: text

Table: Subcohorts

Overview and relationships:

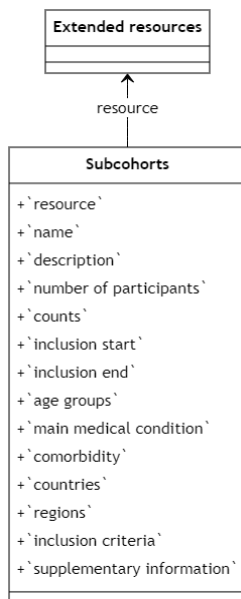


Table definition:

Subcohorts defined for this resource

Column definitions:

resource – *Resource this subcohort is part of*
constraints: ref(Extended resources) required

name – *Subcohort name, e.g. 'mothers in first trimester','newborns'*
constraints: string required

description – *Subcohort description*
constraints: text

number of participants – *Number of participants in this subcohort*
constraints: int

counts – *Total number of unique individuals per age(group), gender and year*
constraints: refbck(Subcohort counts, refBack=subcohort)

inclusion start – *Year of first included participant*
constraints: int

inclusion end – *Year of last included participant. Leave empty if collection is ongoing*
constraints: int

age groups – *Age groups within this subcohort*
constraints: ontology_array(CatalogueOntologies.Age groups)

main medical condition – *Disease groups within this subcohort, based on ICD-10 and ORPHA code classifications*

constraints: ontology_array(CatalogueOntologies.Diseases)

comorbidity – *Comorbidity within this subcohort, based on ICD-10 classification*

constraints: ontology_array(CatalogueOntologies.Diseases)

countries – *Countries where data from this subcohort largely originate from*

constraints: ontology_array(CatalogueOntologies.Countries)

regions – *Geographical regions where data from this subcohort largely originate from*

constraints: ontology_array(CatalogueOntologies.Regions)

inclusion criteria – *Inclusion criteria applied to this subcohort*

constraints: text

supplementary information – *Any other information that needs to be disclosed for this subcohort*

constraints: text

Table: Variable mappings

Overview and relationships:

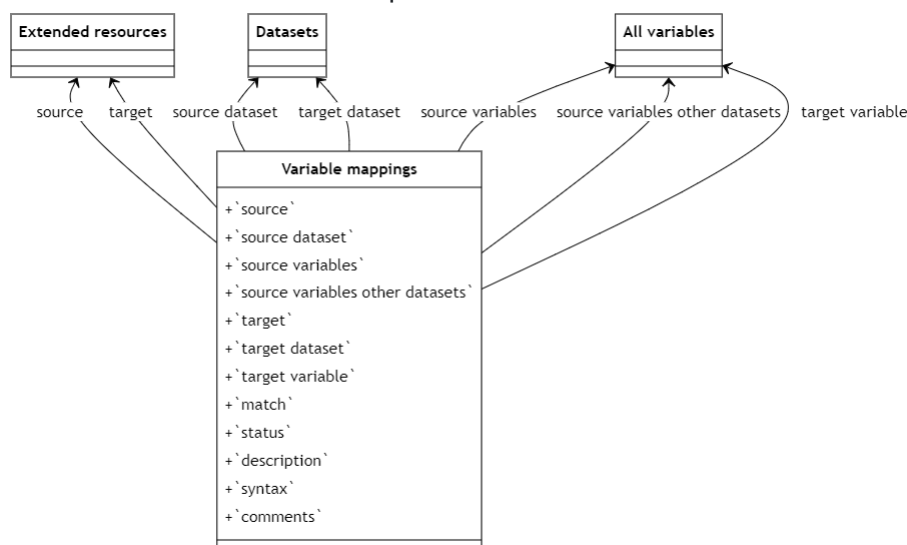


Table definition:

Mappings from collected variables to standard/harmonized variables, optionally including ETL syntax

Column definitions:

source

constraints: ref(Extended resources) required

source dataset

constraints: ref(Datasets, refLink=source) required

source variables – *Optional, source variable that was mapped from. You may also indicate that a mapping to a target variable was not done and leave this field empty (match = na)*

constraints: ref_array(All variables, refLink=source dataset)

source variables other datasets – *optional, variable from other source datasets. Initially one may only define mapping between releases*

constraints: ref_array(All variables, refLink=source)

target

constraints: ref(Extended resources) required

target dataset

constraints: ref(Datasets, refLink=target) required

target variable – *in UI this is then one lookup field. In Excel it will be two columns. Value of 'targetVariable' is filtered based on selected 'targetCollection' and together be used for fkey(collection,dataset,name) in Variable*

constraints: ref(All variables, refLink=target dataset) required

match – e.g. *'complete, partial, planned, no-match'*

constraints: ontology(CatalogueOntologies.Status details) required

status – *whether harmonisation is still draft or final*

constraints: ontology(CatalogueOntologies.Status)

description – *human readable description of the mapping*

constraints: text

syntax – *formal definition of the mapping, ideally executable code*

constraints: text

comments – *additional notes and comments*

constraints: text

Table: Variable values

Overview and relationships:

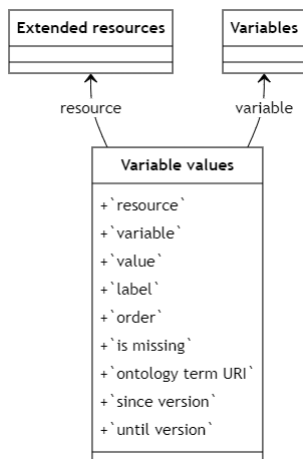


Table definition:

Listing of categorical value+label definition in case of a categorical variable

Column definitions:

resource

constraints: ref(Extended resources) required

variable – e.g. PATO

constraints: ref(Variables, refLink=resource) required

value – e.g. '1'

constraints: string required

label

constraints: string required

order

constraints: int

is missing

constraints: bool

ontology term URI – reference to ontology term that defines this categorical value

constraints: string

since version – When this variable value was introduced if applicable

constraints: string

until version – When this variable value was removed if applicable

constraints: string

