

IMI2 821520 - ConcePTION

ConcePTION

WP7 – Information and data governance, ethics, technology, data catalogue and quality support

D7.7 Prototype of FAIR data catalogue - 2nd

Lead contributor	Morris Swertz, Eleanor Hyde, Fernanda de Andrade (17 – UMCG) m.a.swertz@rug.nl
Other contributors	Rosa Gini (10 – ARS) Romin Pajouheshnia, Miriam Sturkenboom (1 – UMCU) Marianne Cunnington (41 – GSK) Petr Holub (4 – BBMRI) Helen Dolk, Maria Loane, Joanne Given (2 – ULST) David Lewis (38 – NVS) Laura Yates (13 – UKZN) Eugene van Puijenbroek, Bernke te Winkel (9 – LAREB)

Document History

Version	Date	Description	Non-contributor reviewers (if applicable)
V1.0	26 May 2021	Final	Task 7.4 team members
V0.1	8 Feb 2021	First Draft	Work Package 7
V0.2	4 Jun 2021	Second Draft	ConcePTION Managing Board
V1.0	22 Jun 2021	Final Version	IMI
V2.0	31 Jul 2021	Updated Version	

Abstract

The objective of this report is to summarize delivery of the second prototype of the FAIR data catalogue for the ConcePTION project.

The objective of this prototype of the data catalogue is to collect **summary level metadata on the CONCEPTION common data model and to map this metadata to population-based healthcare data sources and databanks** to aid ConcePTION partners in the execution of analyses to generate evidence on the safety of medicines in pregnancy (i.e. data elements and mapping catalogue). The design of this catalogue is based on the requirements analysis as delivered in D7.1 'User requirements and meta-data model for the FAIR data catalogue from WP1&2'.

In particular, the catalogue prototype consists of sections for:

- **Data access providers** - Contributors to the catalogue such as universities, companies, medical centres and research institutes
- **Data sources** - Collections of data banks covering the same population
- **Data banks** - Data collections such as registries or biobanks
- **Networks** - Collaborations of multiple institutions
- **Common Data Models** - Common Data Element models and Harmonization models
- **Studies** - Collaborations of multiple institutions, addressing research questions using data sources and/or data banks

In addition, we added supportive sections to describe the underlying data elements:

- **Releases** - defines a version of metadata on tables/variables
- **Tables** - Tables in data banks or models
- **Variables** - Variables in data banks or models
- **VariableValues** - Categorical values of variables, if applicable
- **RepeatedVariables** - Specific structure when same variable is measured at more timepoints
- **TableMappings** - Rule how common data model table should be created from source
- **VariableMappings** - Rule how a common data model variable should be created from source

In the past months we have implemented these requirements into a fully functional catalogue that is ready for user testing. The coming months will be spent thoroughly evaluating the prototype from both a data entry and catalogue (researcher) user perspective. The results of this testing period will be reported in a next deliverable D7.10: 'Test report of FAIR data catalogue 2nd'. Meanwhile all testing findings will be processed and necessary modifications made in order to accept the prototype into production state to be used by WP1, WP2 and WP7.

Table of contents

Abstract	2
Methods	4
Requirements analysis	5
Implementation using MOLGENIS framework	5
Iterative prototyping	6
Results	6
Result 1: Data catalogue database structure	6
Resources module	6
Dictionary module	7
Ontologies module	8
Result 2: Data catalogue prototype user interface	9
Home page / landing page	9
List/search screens	10
Institution screen	10
Datasource screen	11
Databank screen	13
Network screen	13
Models, Studies, Contacts, Affiliations screens	14
Table screen	14
Mappings/ETL screen	15
Result 3: Data maintenance and interoperability	16
Data entry forms for humans	16
Excel/CSV upload template	17
Programmatic and semantic interface	18
Conclusion & looking forward	20
Appendix 1: data model	21
Resources module definition	21
Data dictionary definition	27
Ontologies module definition	30

Methods

As extensively reported in the deliverable D7.1 and D7.6, the objective of the ConcePTION data catalogue is to serve the needs of ConcePTION partners to identify relevant data access providers and data sources for the conduct of studies to generate evidence on the safety of medicines in pregnancy and breastfeeding. For ConcePTION partners the catalogue will also be a repository to enable management and sharing of metadata on the individual data items, data transformation tools, scripts and decisions for harmonisations to be used in ConcePTION studies. As such, the catalogue will serve as a key enabler of transparency around the ConcePTION data pipeline.

The current (second) prototype focuses on a subset of these requested features:

Meta data model for DAPs

This feature aims to standardise the **metadata model** to capture descriptions of DAPs within Conception and to make potential data sources for pregnancy research findable and understandable for researchers. The metadata models map to key concepts representing different levels of meta data accessed through ConcePTION DAPs. These include:

- **Institutions**, organisations that play a role in the entities below
- **Data Access Providers (DAPs)**, institutions with authorisation, capacity and expertise to access and process healthcare data
- **Data Sources** which are collections of data banks having overlapping underlying (source) populations and that can be accessed by (at least) a DAP which is entitled to link them with each other at the individual level
- **Data banks** are collections of structured healthcare data which are defined and maintained by an organization, called *originator* (e.g. healthcare provider or healthcare payer), by the *prompt* that causes one record to come into existence (e.g. contact between a person and their primary care provider, versus discharge from a hospital admission, versus dispensing of a medicinal product by a pharmacy), and by its content.
- **Data tables** where the content of the data bank is stored.

These concepts have been defined after content analysis of the interviews conducted with 20 DAPs of population-based healthcare data (see Deliverable 7.5). The concepts will be expanded in the future to pregnancy report data and further mapped against additional standards such as CDISC and GDPR (controllers versus processors of data).

For a DAP or data source to be included in the catalogue it is not necessary for all levels of metadata to be available. For example, basic descriptive information on a potential DAP and/or affiliated institution may be included without more detailed information on the data sources and data banks accessed through that DAP being available.

Model and tools for common data models

This feature enables loading of complete data dictionaries for databanks/datasources and data mappings to the agreed upon common models. An important example of such common model is the **ConcePTION Common Data Model (CDM)**. The catalogue provides tools to define the **mapping** including the original data dictionaries and data models of the data banks, the Extract Transform and Load (ETL) designs and scripts that transform the original data to the ConcePTION CDM.

Future iterations of the catalogue will include additional features: The Catalogue will host a selection of the results of the Data Characterisation of data sources participating in ConcePTION studies, organised using visualisation dashboards which can be queried by researchers to understand the detailed features of the data whose description is contained in the other sections of the Catalogue.

Access to such results will be regulated via username and password; A secure negotiator service for researchers to engage DAPs in pregnancy and breastfeeding research studies.

Requirements analysis

We analysed requirements for the metadata models linking to the work to design the ConcePTION Common Data Model (CDM) based on standardised qualitative interviews with population-based healthcare DAPs within the ConcePTION consortium and on the data sources they have access to [see Deliverable 7.5].

We aimed to make the contents of the catalogue compatible with existing software, i.e., to promote findability, accessibility, interoperability and reusability (FAIRness) by enabling data exchange with existing FAIR systems and existing software. For this we made use of our network in international catalogue activities, most notably ESFRI BBMRI-ERIC (Directory of biobanks) and H2020 projects EUCAN-connect, Athlete, Cineca, LifeCycle, Longitools, which includes collaboration with Maelstrom, a leading data cataloguing and harmonisation resource from Canada. While these already can be perceived as standard, we also we analysed existing standards such as MIABIS (Merino-Martinez et al, 2016), DataCite (<https://datacite.org/>), HL7, FAIR data points (<https://www.fairdatapoint.org/>), and existing schema's used in software such as W3C data catalog vocabulary (dcat), and data dictionaries in software MOLGENIS, RedCap, Castor, OpenClinica to name a few. We combined the best practices from these projects with the needs identified in the ConcePTION consortium, acknowledging that the needs of ConcePTION are different from the needs of aforementioned consortia which are oriented to cohort data rather than large population-based healthcare datasets. The UMCG mission is to synergise and converge catalogue development in the health data space, adding the specific needs of ConcePTION into development on the international podium.

Implementation using MOLGENIS framework

We implemented the catalogue software using MOLGENIS, an open source software framework to automatically generate bespoke FAIR catalogues and web applications based on a custom defined data model. The system allows for creation of query user interfaces, a main menu, questionnaire/survey, create/edit/delete edit forms to allow curation of the data, and Excel based data import/export formats and programmatic interfaces. Manuals can be found at <http://molgenis.org> and scientific publications at <https://www.ncbi.nlm.nih.gov/pubmed/?term=molgenis>

The starting point was the conceptualisation of data bank/data source illustrated in the Introduction. The identification of such concepts was obtained in WP7 based on content analysis of the extensive interviews with the DAPS, see Deliverable 7.5. These results will also be included in a peer reviewed scientific manuscript.

In addition we analyzed background knowledge from UMCG/BBMRI from previous cataloguing efforts in this domain such as H2020 projects LifeCycle, EUCAN-connect, as well as well known catalogue efforts we collaborate with such as the Canadian Maelstrom initiative.

The results of this were operationalized into a structured data model in MOLGENIS. Subsequently we engaged in a series of prototyping activities to finetune the user interfaces. Given resourcing constraints UMCG received contributions in kind from the EUCAN-connect project, under the condition that software would be developed open source towards a joint prototype. The added value of this approach is to increase chances of long term sustainability as well as promoting interoperability (and and increases opportunities to unite catalogue efforts down the line).

Iterative prototyping

With the first draft of the prototype available, we conducted sessions with WP1, WP2, and WP7 representatives over a 4 month period to discuss the metadata model, the visualisation of that within MOLGENIS, and the query capabilities. The final iteration resulted in the prototype described in this report.

Results

Result 1: Data catalogue database structure

Using MOLGENIS, we translated the requirements into a relational data model (in collaboration with EUCAN-connect). The data model was split in three main modules:

Module	Description
Resources	Descriptions of Institutions, Data sources, Data banks, Networks and Common Data Element models
Dictionaries	Versioned definitions of Data Tables, Variables and Mappings/ETLs that are linked to Resources.
Ontologies	Lists of codes used to characterise the elements of Resource, for example 'type' or 'age' or 'country'. Importantly, we aim to link all codes used to existing code systems/vocabularies/ontologies to promote interoperability. This module can describe those.

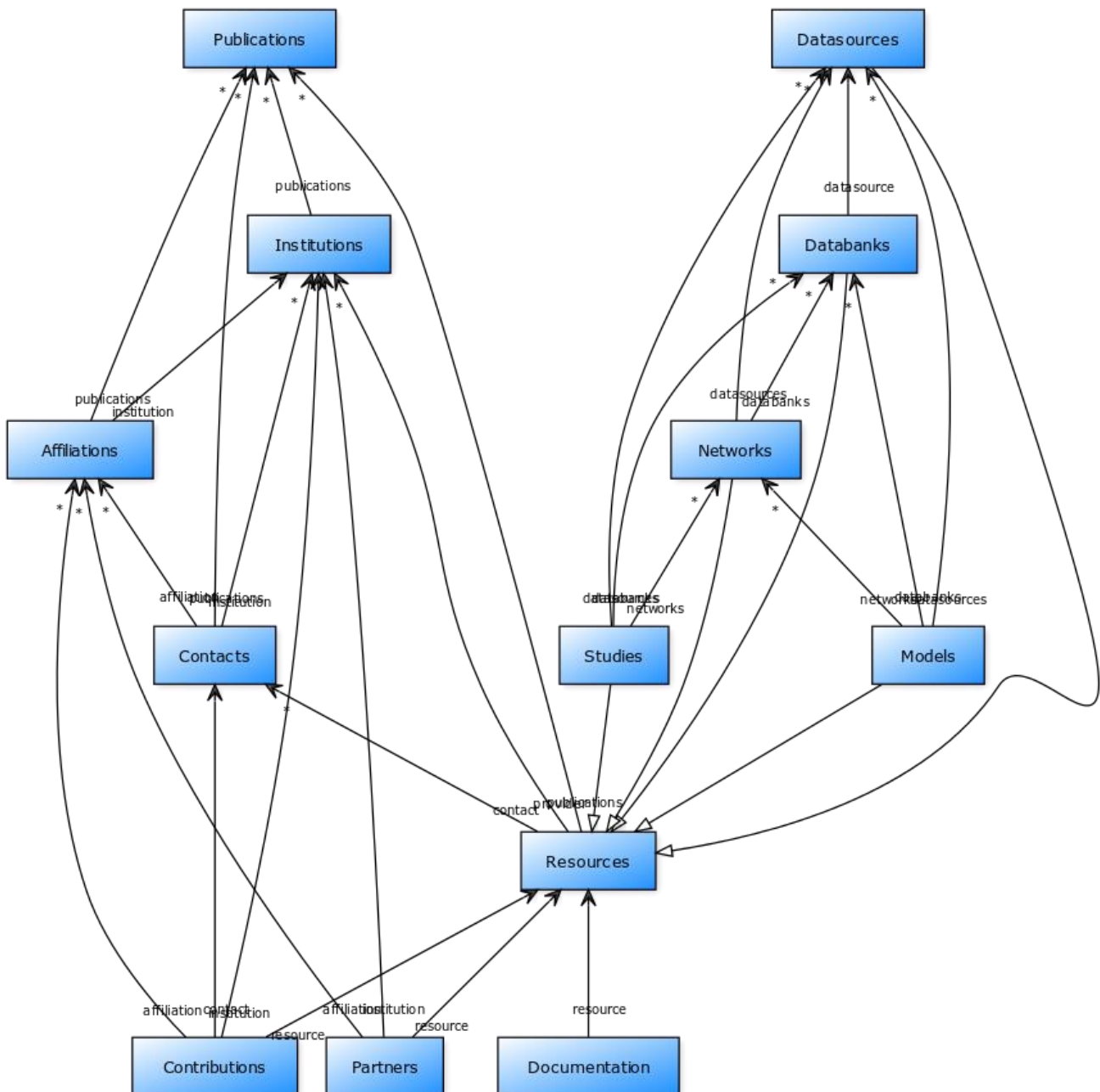
Each of the modules is summarized below. Full definition can be found in Appendix.

Resources module

This module defines the properties and relationships for describing data resources, i.e.

- Institutions - typically controller or processor of data
- Data sources - family of databanks originating from same underlying (source) population
- Data banks - collected digital data on a group of individuals such as registries, cohorts, etc
- Networks - collaborations of institutions
- Models of Common Data Elements (CDE) - describing data standards used for harmonisation across data sources
- Studies - making use of the data.

In addition this module describes the relevant contact details and affiliations of the Data Access Providers. Note that we used an abstraction, 'Resources' to define what should be commonly described, and then Datasources, Databanks, Networks, Studies and Models define additional details and relations. Full documentation in the appendix, summary image below:



CREATED WITH YUML

Dictionary module

This module defines properties and relationships for describing data dictionaries, i.e.,

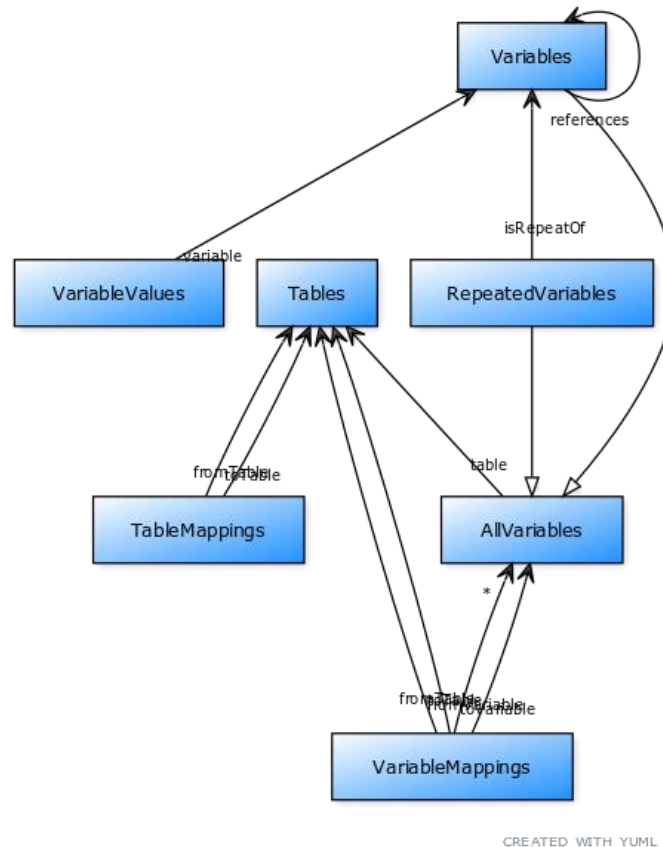
- Releases - defining a version of data dictionary
- Tables - defining data collection tables / tables defined in a standard
- Variables - defining the columns of a table, including details necessary for research use
- Variable values - codes used in case of a categorical variable
- Repeated variables - used in case a variable from Variables is collected at multiple time points, repeated variables can be used to describe those

Note that data dictionaries can apply to Data sources and Data banks, but can also be used to define Common Data (Element) Models (CD(E)M). Whether a data dictionary relates to a CDM or to a datasource/databank is made explicit.

In addition, this module defines ETL relationships between tables and variables, i.e.

- TableMappings - how records from origin table should be mapped to a CD(E)M table.
- VariableMappings - how variables should be mapped to CDE variables.

Below an overview of this module; full documentation in the appendix:



Ontologies module

Finally, this module defines the code lists we used to ensure descriptions can be made interoperable as part of data harmonisation for ConcePTION distributed studies and analyses. Each ontology follows the same structure: term, label, code, term URL (i.e. hyperlink to ontology definition source). The model contains the following ontologies/code lists:

- Topics - for codifying the contents of resources, tables and variables
- Status - for defining if a ETL is 'draft' or 'complete'
- StatusDetails - for defining the level ETL is achieved, e.g. 'partial', 'none', 'ful'
- AgeCategories - for defining ages on which data is collected
- InclusionCriteria - for defining criteria applied on population selection e.g. 'mothers'
- Regions - countries and regions
- ResourceTypes - types of Datasource, Databanks, Networks, etc, e.g. 'registry'
- Units - for characterising variables, e.g. 'cm'
- Formats - for defining type of variable, e.g. 'numeric'
- DocumentTypes - for defining type of resource documentation, e.g. 'protocol'
- InstitutionTypes - for defining type of institute, e.g. 'university'
- Vocabularies - for defining code lists used in variables, e.g. 'icd-10'
- UnitOfObservation - for defining what each row in a table is collected on, e.g. 'dispensing'
- UpdateFrequency - for defining how frequently a databank is updated, e.g. 'annually'
- ContributionTypes - for defining role of a person in a resource, e.g. 'principal investigator'
- PartnerRoles - for defining role of a institution in a resource, e.g. 'linked third party'

- Conditions - for defining conditions of use, e.g. ‘collaboration required’
- We recommend to use ontologies for each of these. E.g. ‘conditions’ could use ‘data use ontology (DUO)’.

Result 2: Data catalogue prototype user interface

Second we delivered a user interface to test the data model and design how interested users can explore the catalogue contents.

Home page / landing page

The most important result of this effort is the prototype user interface. The home page provides an overview:

HEALTH DATA CATALOGUE

Browse and manage metadata for human research data resources, such as cohorts, registries, biobanks, and multi-center studies thereof, such as EU projects and harmonisations studies. This catalogue software has been made possible by contributions from H2020 EUCAN-connect, LifeCycle, Longitools and ATHLETE as well as IMI Conception and EMA.

DATA COLLECTIONS

INSTITUTIONS ³²
Universities, Companies, Medical Centers and Research Institutes

DATASOURCES ⁷
Families of databanks collected from same population

DATABANKS ¹⁸
Such as Cohorts, Registries, Biobanks

DATA USE

NETWORKS ¹
Collaborations of multiple intitutions, datasources and/or databanks.

COMMON MODELS ³
Common Data Element models and Harmonisation models

STUDIES ⁰
Collaborations of multiple intitutions, datasources and/or databanks.

BROWSE BY INSTITUTE AND CONTACTS

CONTACTS ³²
Researchers, data managers,

AFFILIATIONS ⁴
Departments, divisions and research groups.

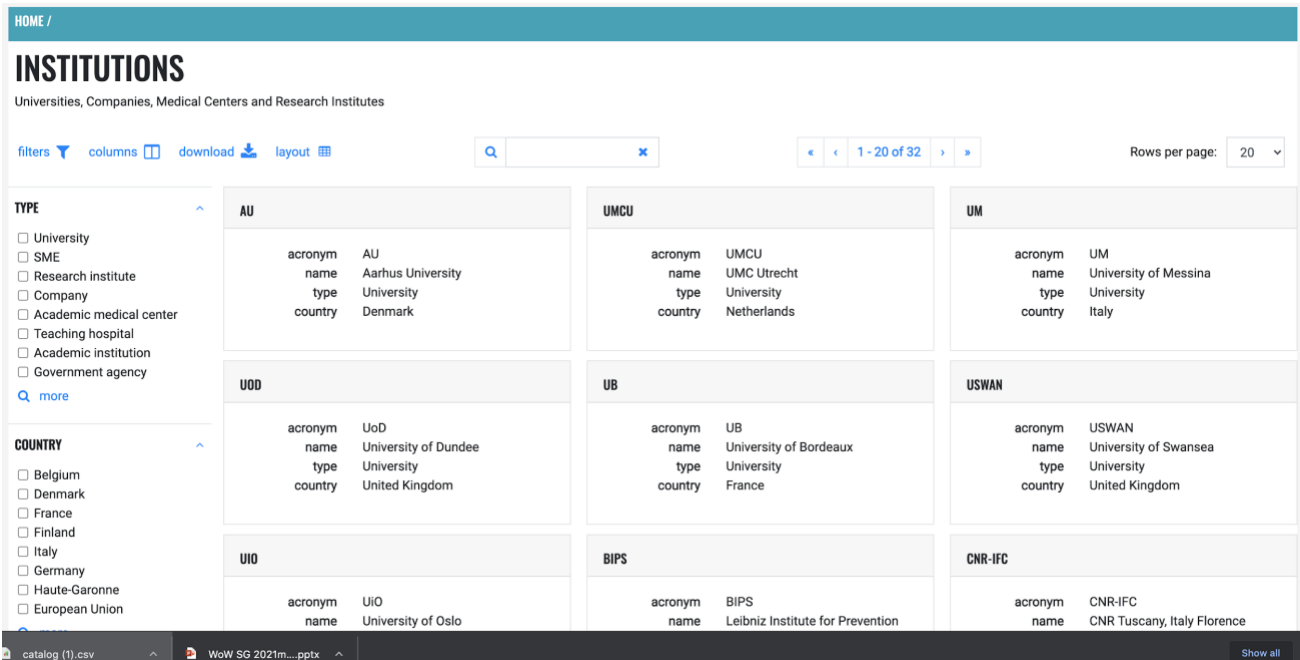
BROWSE DATA DEFINITIONS

<p>RELEASES ¹⁰</p> <p>Data releases from databanks, models or networks.</p>	<p>TABLES ⁷⁹</p> <p>Raw listing of all tables described across all releases of all databanks and common models.</p>	<p>VARIABLES ⁴⁵⁸</p> <p>Raw listing of all variables described across all releases of all databanks and common models.</p>	<p>TABLE MAPPINGS ³</p> <p>Raw listing of all mappings described between tables in databanks and those in common data models.</p>	<p>VARIABLE MAPPINGS ³⁴</p> <p>List of all mappings described variables in databanks and those in common data models.</p>
---	---	--	---	---

Below we provide examples of the main screens. All screens follow the same pattern: there is a 'listing' screen, for listing all records, and a 'detail' screen for viewing one of the records. Note how where possible each screen follows the same pattern.

List/search screens

For all screens we provide a listing screen enabling users to filter and search. They all follow the same design, where the user can add/remove filters, show/hide columns, apply search and change between 'table' and 'card' layout. One example below for 'Institutions':



HOME /

INSTITUTIONS

Universities, Companies, Medical Centers and Research Institutes

filters columns download layout

Q X

1 - 20 of 32

Rows per page: 20

TYPE	AU	UMCU	UM
<input type="checkbox"/> University <input type="checkbox"/> SME <input type="checkbox"/> Research institute <input type="checkbox"/> Company <input type="checkbox"/> Academic medical center <input type="checkbox"/> Teaching hospital <input type="checkbox"/> Academic institution <input type="checkbox"/> Government agency more	acronym AU name Aarhus University type University country Denmark	acronym UMCU name UMC Utrecht type University country Netherlands	acronym UM name University of Messina type University country Italy
COUNTRY	UOD	UB	USWAN
<input type="checkbox"/> Belgium <input type="checkbox"/> Denmark <input type="checkbox"/> France <input type="checkbox"/> Finland <input type="checkbox"/> Italy <input type="checkbox"/> Germany <input type="checkbox"/> Haute-Garonne <input type="checkbox"/> European Union	acronym UoD name University of Dundee type University country United Kingdom	acronym UB name University of Bordeaux type University country France	acronym USWAN name University of Swansea type University country United Kingdom
	UIO	BIPS	CNR-IFC
	acronym UIO name University of Oslo	acronym BIPS name Leibniz Institute for Prevention	acronym CNR-IFC name CNR Tuscany, Italy Florence

catalog (1).csv WoW SG 2021m...pptx Show all

Institution screen

The institution screen shows a listing of the organisations that provide access to resources described. E.g.

HOME / INSTITUTIONS /

UIO

UNIVERSITY OF OSLO

<https://www.uio.no/>

Description: N/A



COUNTRY

Norway

CONTACTS

N/A

PROVIDER OF:

DATASOURCES

- [NIPH - Registries of the Norwegian Institute of Public Health](#)

DATABANKS

- [NorPD - The Norwegian Prescription Database](#)
- [MBRN - The Medical Birth Registry of Norway](#)
- [NPR - The Norwegian Patient Registry](#)

NETWORKS

N/A

PARTNER IN:

- [CONCEPTION - Continuum of Evidence from Pregnancy Exposures, Reproductive Toxicology and Breastfeeding to Improve Outcomes Now \(Beneficiary\)](#)

Datasource screen

The datasources screen enables browsing of resources that are families of databanks.

[HOME / DATASOURCES /](#)

NIPH

REGISTRIES OF THE NORWEGIAN INSTITUTE OF PUBLIC HEALTH

<https://www.fhi.no>

Norway has a universal public health care system, consisting of primary health care services and specialist health care services. Many population-based health registries were established in the 1960s, with use of unique personal identifiers facilitating linkage between registries. The mandatory national health registries were established to maintain national functions

POPULATION

Norway

INCLUSION CRITERIA

Mothers and fathers and their infants

DATABANKS

- [NorPD - The Norwegian Prescription Database](#)
- [MBRN - The Medical Birth Registry of Norway](#)
- [NPR - The Norwegian Patient Registry](#)

SUMMARY STATISTICS

N/A

PROVIDER

- [UiO - University of Oslo](#)

CONTACT

N/A

DATA RELEASES

- [1.0.0](#)

DOCUMENTATION

N/A

NETWORKS

- [CONCEPTION - Continuum of Evidence from Pregnancy Exposures, Reproductive Toxicology and Breastfeeding to Improve Outcomes Now](#)

PARTNERS

N/A

CONTRIBUTORS

N/A

PUBLICATIONS

N/A

Databank screen

HOME / DATABANKS /

REGISTRY

NORPD

THE NORWEGIAN PRESCRIPTION DATABASE

<http://www.norpd.no/>
Description: N/A

<p>PART OF DATASOURCE</p> <ul style="list-style-type: none"> • NIPH - Registries of the Norwegian Institute of Public Health <p>POPULATION</p> <p style="background-color: #e67e22; color: white; padding: 2px 5px; display: inline-block;">Norway</p> <p>INCLUSION CRITERIA</p> <p>N/A</p> <p>TOPICS</p> <p>N/A</p> <p>NUMBER OF PARTICIPANTS:</p> <p>N/A</p> <hr/> <p>ORIGINATOR</p> <p>N/A</p> <p>RECORD PROMPT:</p> <p>A filled prescription in a community pharmacy or in a hospital pharmacy for outpatient use, regardless of reimbursement rights. Ambulatory use (eg infusion) does not trigger a record unless the drug is bought by the patient</p> <p>START/END YEAR</p> <p>2004 - N/A</p>	<p>PROVIDER</p> <ul style="list-style-type: none"> • UiO - University of Oslo <p>CONTACT</p> <p>N/A</p> <p>DATA RELEASES</p> <p>N/A</p> <p>DOCUMENTATION</p> <p>N/A</p> <p>NETWORKS</p> <ul style="list-style-type: none"> • CONCEPTION - Continuum of Evidence from Pregnancy Exposures, Reproductive Toxicology and Breastfeeding to Improve Outcomes Now <p>PARTNERS</p> <p>N/A</p> <p>CONTRIBUTORS</p> <p>N/A</p> <p>PUBLICATIONS</p> <p>N/A</p>
---	--

Network screen

Networks describe relations between institutions, data banks and data sources towards enabling integrated analysis:

CONCEPTION

CONTINUUM OF EVIDENCE FROM PREGNANCY EXPOSURES, REPRODUCTIVE TOXICOLOGY AND BREASTFEEDING TO IMPROVE OUTCOMES NOW

<https://www.imi-conception.eu/>

The main goal of ConcePTION is to establish a trusted ecosystem that can efficiently, systematically, and in an ethically responsible manner, generate and disseminate reliable evidence-based information regarding effects of medications used during pregnancy and breastfeeding to women and their healthcare providers. This will be achieved by generating, cataloguing, linking, collecting and analysing data from pharmacovigilance, modelling, routine healthcare, breastmilk samples through a large network.

DATASOURCES INVOLVED

- [NIPH - Registries of the Norwegian Institute of Public Health](#)
- [EFEMERIS - Evaluation chez la Femme Enceinte des MEDicaments et de leurs RISques](#)

DATABANKS INVOLVED

- [NorPD - The Norwegian Prescription Database](#)
- [MBRN - The Medical Birth Registry of Norway](#)
- [NPR - The Norwegian Patient Registry](#)

FUNDING

The ConcePTION project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No 821520. This Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA.

PROVIDER

- [UMCU - UMC Utrecht](#)
- [GSK - GlaxoSmithKline](#)

CONTACT

N/A

DATA RELEASES

- [1.0.0](#)

DOCUMENTATION

N/A

NETWORKS

N/A

PARTNERS

- [UiO - University of Oslo](#)
- [AU - Aarhus University](#)
- [UoD - University of Dundee](#)
- [ULST - Ulster University](#)
- [CHUT - Centre Hospitalier Universitaire de Toulouse](#)

Models, Studies, Contacts, Affiliations screens

Similar screens exist for Models, Studies, Contacts and Affiliations. These are not shown for brevity (i.e. many of the data sources, data banks, partners involved in ConcePTION are not shown in above screenshot).

Table screen

For each release, there is a listing of tables from the ConcePTION CDM (and in the future maybe from each local datasource), and each table has its own homepage providing a listing of its variables.

HOME / TABLES /

NETWORK: [CONCEPTION](#) / [RELEASE 1.0.0](#)

TABLE: VISIT_OCCURRENCE

Role: This table contains a summary description of the visits during which records of EVENTS, PROCEDURES, but possibly also MEDICAL_OBSERVATIONS or VACCINES or MEDICATIONS were recorded. This serves both to collect visit-level information, and to enable grouping sets of records that were recorded concurrently

TOPICS

Routine healthcare data

UNIT OF OBSERVATION

N/A

MAPPINGS/ETLS

- From: [Datasource: ARS - Version: 1.0.0 - Table: SDO](#)
- From: [Datasource: ARS - Version: 1.0.0 - Table: SPA](#)
- From: [Datasource: ARS - Version: 1.0.0 - Table: AP](#)

VARIABLES

filters columns download layout

Rows per page: 20

TOPICS	CONCEPTION.1.0.OVISIT_OCCURRENCEVISIT_EN...	CONCEPTION.1.0.OVISIT_OCCURRENCEORIGIN_0	CONCEPTION.1.0.OVISIT_OCCURRENCEMEANING...																																
<input type="checkbox"/> Allergy <input type="checkbox"/> Air pollution <input type="checkbox"/> Allergic sensitization <input type="checkbox"/> ADHD Registry diagnosis <input type="checkbox"/> ADHD symptoms <input type="checkbox"/> ASD Registry diagnosis <input type="checkbox"/> Absolute <input type="checkbox"/> Adult health-related characteristics	<table border="1"> <tr><td>name</td><td>visit_end_date</td></tr> <tr><td>label</td><td>Visit end date</td></tr> <tr><td>format</td><td>Date</td></tr> <tr><td>topics</td><td>Routine healthcare data</td></tr> </table>	name	visit_end_date	label	Visit end date	format	Date	topics	Routine healthcare data	<table border="1"> <tr><td>name</td><td>origin_of_visit</td></tr> <tr><td>label</td><td>Visit outcome</td></tr> <tr><td>format</td><td>Categorical</td></tr> <tr><td>mandator</td><td>true</td></tr> <tr><td>y</td><td>topics</td></tr> <tr><td></td><td>Routine healthcare data</td></tr> </table>	name	origin_of_visit	label	Visit outcome	format	Categorical	mandator	true	y	topics		Routine healthcare data	<table border="1"> <tr><td>name</td><td>meaning_of_visit</td></tr> <tr><td>label</td><td>Visit meaning</td></tr> <tr><td>format</td><td>Categorical</td></tr> <tr><td>mandator</td><td>true</td></tr> <tr><td>y</td><td>topics</td></tr> <tr><td></td><td>Routine healthcare data</td></tr> </table>	name	meaning_of_visit	label	Visit meaning	format	Categorical	mandator	true	y	topics		Routine healthcare data
name	visit_end_date																																		
label	Visit end date																																		
format	Date																																		
topics	Routine healthcare data																																		
name	origin_of_visit																																		
label	Visit outcome																																		
format	Categorical																																		
mandator	true																																		
y	topics																																		
	Routine healthcare data																																		
name	meaning_of_visit																																		
label	Visit meaning																																		
format	Categorical																																		
mandator	true																																		
y	topics																																		
	Routine healthcare data																																		
	<table border="1"> <tr><td>name</td><td>person_id</td></tr> <tr><td>label</td><td>Person identifier</td></tr> <tr><td>format</td><td>Character</td></tr> </table>	name	person_id	label	Person identifier	format	Character	<table border="1"> <tr><td>name</td><td>id</td></tr> </table>	name	id	<table border="1"> <tr><td>name</td><td>visit_start_date</td></tr> </table>	name	visit_start_date																						
name	person_id																																		
label	Person identifier																																		
format	Character																																		
name	id																																		
name	visit_start_date																																		

Mappings/ETL screen

Finally, one of the most exciting views in the catalogue, that enables to see how tables from data banks have been mapped to common data elements. The format of such a *specification table* was specified in Appendix 6 of Deliverable 7.5.

HOME / TABLEMAPPINGS /

TABLE MAPPING

TARGET TABLE:	VISIT_OCCURRENCE within release CONCEPTION 1.0.0			
ORIGIN TABLE:	SDO within release ARS 1.0.0			
ACTION:	For each record Create a record of VISIT_OCCURRENCE and label the records with a unique code stored in visit_occurrence_id (primary key) Copy the values of SDO into VISIT_OCCURRENCE according to the following table			
TARGET COLUMN	ORIGIN COLUMN	RULE	SYNTAX	NOTES
id			Create a record of VISIT_OCCURRENCE and label the records with a unique code stored in visit_occurrence_id (primary key)	
person_id		Create a record of VISIT_OCCURRENCE and label the records with a unique code stored in visit_occurrence_id (primary key)	VISIT_OCCURRENCE.person_id = SDO.IDUNI	
visit_occurrence_id		Label the records with a unique code stored in visit_occurrence_id (primary key)		
visit_start_date				
visit_end_date				
specialty_of_visit			VISIT_OCCURRENCE.specialty_of_visit = SDO.REPDIM	
status_at_discharge			VISIT_OCCURRENCE.status_at_discharge = SDO.MODIM	

Result 3: Data maintenance and interoperability

In addition to the data model, efforts have been made to enable interoperability, with the aim to (a) ease data upload and (b) make the catalogue ready to participate in FAIR networks (adhering to the FAIR principles).

Data entry forms for humans

While we envision that ideally most content will be entered via data entry templates by WP7 and by DAPs, or potentially and ideally in an automated manner, we are certain that data curators will need ways to edit the data. Therefore we invested in data entry forms, for example:

UPDATE INSTITUTIONS ×

acronym (required)

AU

name

Aarhus University

type

University ▼

description

country

Denmark ▼

homepage

https://international.au.dk/

logo

× Browse

publications

▼ +

partnerIn

CONCEPTION AU ▼ +

providerOf

DHA ▼

Close update Institutions

Excel/CSV upload template

For data entry/update at large scale we provide Excel/CSV templates for each of the data tables. For example:

table	name	bits	label	format
VISIT_OCCURRENCE	meaning_of_visit	utine healthcare data		Categorical
VISIT_OCCURRENCE	person_id	utine healthcare data		Character
SURVEY_ID	record	rveillance		Character
SURVEY_ID	person_id	rveillance		Character
SURVEY_ID	survey_id	rveillance		Character
SURVEY_ID	survey_date	rveillance		Character
EVENTS	end_date_record	utine healthcare data		Date
EVENTS	event_code	utine healthcare data		Categorical
EVENTS	event_free_text	utine healthcare data		Character
EVENTS	event_id	utine healthcare data		Character
EVENTS	meaning_of_event	utine healthcare data		Categorical
EVENTS	origin_of_event	utine healthcare data		Categorical
EVENTS	person_id	utine healthcare data		Character
SURVEY_ID	survey_meaning	rveillance		Categorical
CAP2	record	rveillance		Character
VISIT_OCCURRENCE	origin_of_visit	utine healthcare data		Categorical
VISIT_OCCURRENCE	specialty_of_visit	utine healthcare data		Categorical
VISIT_OCCURRENCE	status_at_discharge	utine healthcare data		Categorical
VISIT_OCCURRENCE	visit_end_date	utine healthcare data		Date
VISIT_OCCURRENCE	visit_occurrence_id	utine healthcare data		Character
CAP2	IDUNI_F	rveillance		Character
CAP2	survey_id	rveillance		Character
CAP2	DATAPARTO_ARSNEW	rveillance		Date
CAP2	survey_meaning	rveillance		Categorical
CAP1	record	rveillance		Character
CAP1	IDUNI_M	rveillance	Mother	Character
CAP1	survey_id	rveillance		Character
CAP1	DATAPARTO_ARSNEW	rveillance		Date

Programmatic and semantic interface

Finally, we have implemented industry standard interfaces, i.e. REST/GraphQL and Linkde data formats JSON-LD and TTL (both used in the FAIR data point community). Full documentation is available in the demo.

Examples of GraphQL interface

```

type Resources {
  acronym: String
  name: String
  type(limit: Int, offset: Int, orderby: ResourceTypesorderby): [ResourceTypes]
  type_agg: ResourceTypesAggregate
  logo: MolgenisFileDownload
  homepage: String
  description: String
  topics(limit: Int, offset: Int, orderby: Topicsorderby): [Topics]
  topics_agg: TopicsAggregate
  contact(limit: Int, offset: Int, orderby: Contactsorderby): [Contacts]
  contact_agg: ContactsAggregate
  provider(
    limit: Int
    offset: Int
    orderby: Institutionsorderby
  ): [Institutions]
  provider_agg: InstitutionsAggregate
  startYear: Int
  endYear: Int
  conditions(limit: Int, offset: Int, orderby: Conditionsorderby): [Conditions]
  conditions_agg: ConditionsAggregate
  licence: String
  funding: String
  acknowledgements: String
  publications(
    limit: Int
    offset: Int
    orderby: Publicationsorderby
  ): [Publications]
  publications_agg: PublicationsAggregate
  population(limit: Int, offset: Int, orderby: Regionsorderby): [Regions]
  population_agg: RegionsAggregate
  inclusionCriteria(

```

Finally, an example of semantic mapping is shown below:

Resources		dcat:Resource	Thing
Resources	acronym	dct:identifier	identifier
Resources	name	dct:title	name
Resources	type	dc:type	additionalType
Resources	logo	foaf:logo	logo
Resources	homepage	foaf:homepage	url
Resources	description	dct:description	description
Resources	topics	dct:theme	about
Resources	contact	dct:contactPoint	contactPoint
Resources	provider	dct:publisher	provider
Resources	startYear	dct:startDate	foundingDate
Resources	endYear	dct:endDate	disolutionDate
Resources	conditions	dct:accessRights	conditionsOfAccess
Resources	licence	dct:licence	license
Resources	funding	datacite:FundingReference	
Resources	acknowledgements		citation
Resources	publications		areaServed
Resources	population	obi:population	
Resources	inclusionCriteria		
Resources	partners	dct:creator	sourceOrganisation
Resources	contributors	dct:qualifiedAttribution	creator
Resources	releases	dcat:distribution	
Resources	documentation	foaf:page	

Conclusion & looking forward

The prototype is ready to be tested by both DAPs as well as researchers, as well as data manager/data scientists responsible for programming ETL and statistical procedures. In the period leading up to the next deliverable we will organise 2 testing rounds, the results of which will be reported in D7.10: 'Test report of FAIR data catalogue 2nd'. In the first testing round, we will have one-on-one interviews with 5 test users (contact details provided by the project management office) to collect independent input that will either confirm the suitability of the current design, or will lead to requests for change. We will then process the test results into a next version of the prototype. Optionally, we will repeat this testing procedure. The aim of this phase is to produce a final version of this part of the catalogue. In addition, we aim to use follow up projects to add persistent identifiers for the metadata in the catalogue, i.e., to increase FAIRness by referring to code systems and ontologies (e.g. HL7, CDISC).

Appendix 1: data model

Below a partial export of the data structure used in the catalogue. Full model is described using MOLGENIS EMX2 data model format (with some technical details omitted).

Resources module definition

tableName	table Extends	column Name	column Type	key	refTable	description
Institutions						Universities, Companies, Medical Centers and Research Institutes
Institutions		acronym	string	1		
Institutions		name	string	2		
Institutions		type	ref		InstitutionTypes	
Institutions		description	text			
Institutions		country	ref		Regions	
Institutions		homepage	string			
Institutions		logo	file			
Institutions		publications	ref_array		Publications	
Institutions		providerOf	refback		Resources	
Institutions		partnerIn	refback		Partners	list of partner roles this institution has in various resources
Contacts						Contact details of a natural person
Contacts		name	string	1		unique name to display this person properly
Contacts		institution	ref_array		Institutions	One or more instutions this contact is employed by
Contacts		affiliation	ref_array		Affiliations	One or more organisational units within institutions this person is part of
Contacts		about	text			
Contacts		photo	file			photo to make user interface look nice
Contacts		email	string	2		email, ideally institute email
Contacts		orcid	string	3		
Contacts		twitter	string	6		

Contacts	homepage	string		
Contacts	linkedin	string	4	
Contacts	researchgate	string	5	
Contacts	contributedTo	refback		Contributions
Contacts	publications	ref_array		Publications
Affiliations				Description of an organisational unit within an Institution that people like to use
Affiliations	acronym		1	short acronym, ideally of form '<institute acronym>_<unit acronym>'
Affiliations	name	string	2	full name of this affiliation
Affiliations	institution	ref		Institutions institution that is responsible for this unit
Affiliations	homepage	string		
Affiliations	description	text		
Affiliations	partnerIn	refback		Partners list of partner roles this institution has in various resources
Affiliations	contributions	refback		Contributions list of partner roles this institution has in various resources
Affiliations	publications	ref_array		Publications
Resources				Generic listing of all resources. Should not be used directly, instead use specific types such as Databanks and Studies
Resources	acronym	string	1	Unique identifier within this catalogue
Resources	name	string	2	e.g. lifelines, lifecycle
Resources	type	ref_array		ResourceTypes e.g. 'cohort', 'network', ...
Resources	logo	file		Logo for use on homepages etc.
Resources	homepage	string		Link to the home page
Resources	description	text		General description
Resources	topics	ref_array		Topics Topics that characterise the contents of this resource
Resources	contact	ref_array		Contacts Whom to contact
Resources	provider	ref_array		Institutions Organisation providing and/or coordinating access to this resource

Resources	startYear	int		Date of first collected data
Resources	endYear	int		Date of last collected data, empty if collection is ongoing
Resources	conditions	ref_array	Conditions	describing access and use conditions
Resources	licence	text		describing license of use
Resources	funding	text		funding statement
Resources	acknowledgements	text		acknowledgement statement
Resources	publications	ref_array	Publications	Publications about this resource
Resources	population	ref_array	Regions	
Resources	inclusionCriteria	ref_array	InclusionCriteria	
Resources	partners	refback	Partners	Institutions involved in the creation of this resource
Resources	contributors	refback	Contributions	Persons involved in the creation of this resource
Resources	releases	refback	Releases	Releases available from this resource
Resources	documentation	refback	Documentation	
Datasources	Resources			Datasource is a resource that defines a family of databanks sampling the same underlying population.
Datasources	databanks	refback	Databanks	
Datasources	networks	refback	Networks	
Databanks	Resources			Databank is a kind of resource that holds collected data
Databanks	datasource	ref	Datasources	what datasource this databank is part of, if applicable
Databanks	noParticipants	int		number of individuals of which data is collected
Databanks	recordPrompt	text		what triggers data collection
Databanks	updateFrequency	ref	UpdateFrequency	how often the data is updated
Databanks	lagTime	text		how long it takes for a update to become available
Databanks	subpopulations	refback	Subpopulations	
Databanks	networks	refback	Networks	

Models		Resources			
Models	networks	ref_array		Networks	
Models	datasources	ref_array		Datasources	
Models	databanks	ref_array		Databanks	
Networks		Resources			
Networks	datasources	ref_array		Datasources	
Networks	databanks	ref_array		Databanks	
					Network is a collaboration bringing data from multiple databanks together
Studies		Resources			
Studies	networks	ref_array		Networks	
Studies	datasources	ref_array		Datasources	
Studies	databanks	ref_array		Databanks	Databanks that provided data into this study
					Study is a resource that holders information of a study done in context of databanks and projects
Partners		Resources			
Partners	resource	ref	1	Resources	resource institution has contributed to
Partners	institution	ref	1	Institutions	institution that contributed
Partners	affiliation	ref_array		Affiliations	optionally, the institutional unit(s) that play a role in this resource
Partners	role	ref		PartnerRoles	role in this resource
Partners	roleDescription	text			human readable description of the role in this resource
					Institutions that partnered in the creation of a resource
Contributions		Resources			
Contributions	resource	ref	1	Resources	resource person has contributed to
Contributions	contact	ref	1	Contacts	contact information of the person who contributed
Contributions	institution	ref		Institutions	type of contribution
Contributions	affiliation	ref_array		Affiliations	optionally, the unit from which the contribution was made
Contributions	contributionType	ref_array		ContributionTypes	description of the contribution
Contributions	contributionDescription	text			longer description, typically used as homepage text for a consortium
					Persons that contributed to the creation of a resource

Documentation				Documentation attached to a resource	
Documentation	resource	ref	1	Resources	The resource this documentation is for
Documentation	name	string	1		name of the document, unique within the resource
Documentation	type	ref		DocumentTypes	type of documentation, e.g. protocol
Documentation	description	text			description of the documentation
Documentation	url	string			hyperlink to the source of the documentation
Documentation	file	file			optional file attachment containing the documentation
Publications				publications following bibtex format	
Publications	doi	string	1		digital object identifier
Publications	title	string			The title of the work
Publications	authors	string_array			List of authors, one string per author
Publications	year	int			The year of publication (or, if unpublished, the year of creation)
Publications	journal	string			The journal or magazine the work was published in
Publications	volume	int			The volume of a journal or multi-volume book
Publications	number	int			The "(issue) number" of a journal, magazine, or tech-report, if applicable. Note that this is not the "article number" assigned by some journals.
Publications	pagination	string			Page numbers, separated either by commas or double-hyphens.
Publications	publisher	string			The publisher's name
Publications	school	string			(in case of thesis) The school where the thesis was written
Publications	abstract	text			
Publications	resources	refback		Resources	list of resources that refer to this publication
Releases				Definition of a data release, in case of Model this will not include data	
Releases	resource	ref	1	Resources	Link to the resource of which contents has been released
Releases	version	string	1		version of the release
Releases	includedModels	ref_array		Releases	existing data models that are used to produce this release

Releases	includesDatabanks	ref_array		Databanks	in case of a network/study, it will only contain data of particular databanks involved
Releases	date	date			date of the release
Releases	description	text			notes specific to this release
CollectionEvents					Definition of an action of data collection for a resource
CollectionEvents	resource	ref	1	Resources	
CollectionEvents	name	string	1		
CollectionEvents	description	string			
CollectionEvents	startYear	int			period of collection start
CollectionEvents	endYear	int			period of collection end
CollectionEvents	ageMin	ref		AgeCategories	minimum ages included, if applicable
CollectionEvents	ageMax	ref		AgeCategories	maximum ages included, if applicable
CollectionEvents	noParticipants	int			number of participants sampled in this event
CollectionEvents	populations	ref_array		Subpopulations	(sub)populations that are targetted with this collection event
CollectionEvents	supplementary Information	text			any other information
Subpopulations					Subpopulations defined in this resource
Subpopulations	resource	ref	1	Resources	E.g. 'Mothers in first trimester','newborns'
Subpopulations	name	string	1		E.g. 'Mothers in first trimester','newborns'
Subpopulations	noParticipants	int			
Subpopulations	description	text			
Subpopulations	InclusionCriteria	ref_array		InclusionCriteria	
Subpopulations	geographicRegion	ref_array		Regions	e.g. province
Subpopulations	supplementary Information	text			

Data dictionary definition

tableName	tableExtends	column Name	column Type	key	refTable	description
Tables						Definition of a table within a data release
Tables		release	ref	1	Releases	resource + version this table is defined for
Tables		name	string	1		unique name in the release
Tables		label	string			short human readable description
Tables		unitOfObservation	ref		ObservationTargets	defines what each record in this table describes
Tables		topics	ref_array		Topics	enables grouping of table list into topic and to display tables in a tree
Tables		description	text			description of the role/function of this table
Tables		numberOfRows	int			count of the number of records in this table
Tables		mappings	refback		TableMappings	list of mappings between this table and standard/harmonized tables
Tables		mappingsTo	refback		TableMappings	
AllVariables						Generic listing of all variables. Should not be used directly, please use Variables or RepeatedVariables instead
AllVariables		release	ref	1	Releases	release this table definition is part of
AllVariables		table	ref	1	Tables	table this variable is part of
AllVariables		name	string	1		name of the variable, unique within a table
AllVariables		collectionEvent	ref		CollectionEvents	in case of protocolised data collection this defines the moment in time this variable is collected on
AllVariables		mappings	refback		VariableMappings	listing of the VariableMappings defined between this variable and standard/harmonized variables
Variables	AllVariables					Definition of a non-repeated variable, or of the first variable from a repeated range
Variables		topics	ref_array		Topics	
Variables		label	string			
Variables		format	ref		Formats	string,int,decimal,date,datetime etc

Variables	unit	ref		Units	unit ontology
Variables	topics	ref_array		Topics	
Variables	references	ref		Variables	to define foreign key relationships between variables within or across tables
Variables	mandatory	bool			whether this variable is required within this collection
Variables	description	text			
Variables	order	int			to sort variables you can optionally add an order value
Variables	exampleValues	string_array			
Variables	permittedValues	refback		VariableValues	
Variables	vocabularies	ref_array		Vocabularies	
Variables	repeats	refback		RepeatedVariables	listing of all repeated variables defined for this variable
RepeatedVariables	AllVariables				Definition of a repeated variable. Refers to another variable for its definition.
RepeatedVariables	isRepeatOf	ref		Variables	reference to the definition of the variable that is being repeated
VariableValues					Listing of categorical value+label definition in case of a categorical variable
VariableValues	release	ref	1	Releases	
VariableValues	variable	ref	1	Variables	e.g. PATO
VariableValues	value	string	1		e.g. '1'
VariableValues	label	string			
VariableValues	order	int			
VariableValues	isMissing	bool			
VariableValues	ontologyTermIRI	string			reference to ontology term that defines this categorical value
VariableMappings					Mappings from collected variables to standard/harmonized variables, optionally including ETL syntax.
VariableMappings	fromRelease	ref	1	Releases	
VariableMappings	fromTable	ref	1	Tables	

VariableMappings	fromVariable	ref	1	AllVariables	optional, variable. Initially one may only define mapping between releases
VariableMappings	toRelease	ref	1	Releases	
VariableMappings	toTable	ref	1	Tables	
VariableMappings	toVariable	ref	1	AllVariables	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.
VariableMappings	match	ref		StatusDetails	e.g. 'complete, partial, planned, no-match'
VariableMappings	status	ref		Status	whether harmonisation is still draft or final
VariableMappings	description	text			human readable description of the mapping
VariableMappings	comments	text			additional notes and comments
VariableMappings	syntax	text			formal definition of the mapping, ideally executable code
TableMappings					
TableMappings	fromRelease	ref	1	Releases	release being mapped from, i.e. fromRelease.resource + fromRelease.version
TableMappings	fromTable	ref	1	Tables	name of the table being mapped from
TableMappings	toRelease	ref	1	Releases	release being mapped to, i.e. toRelease.resource + toRelease.version
TableMappings	toTable	ref	1	Tables	name of the table being mapped to
TableMappings	order	int			Order in which table ETLs should be executed for this source-target combination
TableMappings	description	text			human readable description of the mapping
TableMappings	syntax	text			formal definition of the mapping, ideally executable code

Ontologies module definition

tableName	tableExtends	column Name	column Type	key	refTable	description
OntologyTerms						OntologyTerm table is superclass for all stuff that links to ontology terms
OntologyTerms		name	string	1		
OntologyTerms		code	string	2		identifier used for this code with the ontology
OntologyTerms		order	int			
OntologyTerms		definition	text			
OntologyTerms		comments	text			
OntologyTerms		parent	ref		OntologyTerms	link to a more broad term
OntologyTerms		children	refback		OntologyTerms	link to more specific terms
OntologyTerms		ontologyTerm URI	string	3		
Topics	OntologyTerms					Used to generate the tree on the left; we might want to make multiple trees?
Status	OntologyTerms					
StatusDetails	OntologyTerms					
AgeCategories	OntologyTerms					e.g. '8 week'
InclusionCriteria	OntologyTerms					
Regions	OntologyTerms					Countries, states, provinces and other geographic areas (e.g. using ISO_3166)
ResourceTypes	OntologyTerms					
Units	OntologyTerms					
Formats	OntologyTerms					
DocumentTypes	OntologyTerms					
InstitutionTypes	OntologyTerms					

Vocabularies	OntologyTerms	
ObservationTargets	OntologyTerms	
UpdateFrequency	OntologyTerms	may want to use SNOMED-CT codes
ContributionTypes	OntologyTerms	
PartnerRoles	OntologyTerms	
Conditions	OntologyTerms	