28 June 2021
EMA/104855/2021
Data Analytics and Methods Task Force

Final set of metadata and definitions, process, and catalogue tool

Version 1.1

EMA2017/09/PE/16; EUPAS39322

Official address Domenico Scarlattilaan 6 ● 1083 HS Amsterdam ● The Netherlands
Address for visits and deliveries Refer to www.ema.europa.eu/how-to-find-us
Send us a question Go to www.ema.europa.eu/contact **Telephone** +31 (0)88 781 6000 An agency of the European Union

Disclaimer:

This document is a draft for consultation purposes only.
It should not be interpreted as representing the formal position of EMA or HMA.

28 June 2021

Prepared for:

European Medicines Agency

Katerina-Christina Deli, Stefania Simou
Scientific Administrator
Healthcare Data, Data Analytics and Methods Task Force

Lead authors:

Romin Pajouheshnia, Eleanor Hyde, Morris Swertz, Lia Gutierrez, Susana Perez-Gutthann, Miguel Gil García, Karin Gembert, Nicolas Thurin, Anna Schultze, Marloes Bazelier, Magdalena Gamba, Ella Janssen, Josine Kuiper, Rosa Gini

Co-authors (co-lead teams):

UU, ARS, UMCG, UMCU, BPE, RTI-HS

On behalf of the MINERVA Consortium

Name	Key Roles Organisation	Organisation, Country
Susana Perez-Gutthann, Lia Gutierrez, Carla Franzoni, Alejandro Arana, Joan Fortuny, Estel Plana	**Proposal and Project Coordination** **Lead Tasks 1-3, 5, 10** *Co-lead SIGMA office*	RTI-HS, Barcelona, Spain
Miriam Sturkenboom, Daniel Weibel, Vjola Hoxhaj	**Lead Task 7** *President VAC4EU* *PI ConcePTION*	UMCU, Netherlands
Olaf Klungel, Satu Johanna Siiskonen, Romin Pajouheshnia, Helga Gardarsdottir, Patrick Souverein, Marloes Bazelier, Magdalena Gamba	**Lead Task 4** *EU PE&PV research network lead and management* **DAP, CPRD Aurum, UK**	UU, Netherlands
Rosa Gini, Giuseppe Roberto	**Lead Tasks 6+9, co-lead Task 4** **DAP, Tuscany, Italy**	ARS Toscana (ARS), Italy
Morris Swertz, Eleanor Hyde	**Lead Task 8, Co-lead Task 4**	UMCG, Netherlands
Cécile Droz-Perroteau, Nicolas Thurin, Régis Lassalle	**DAP, SNDS, France**	BPE, University of Bordeaux, France
Vera Ehrenstein	**DAP, Denmark**	AU-DCE, Aarhus, Denmark
Ron Herings, Josine Kuiper, Ella Janssen	**DAP, PHARMO; Netherlands** *Co-lead SIGMA office*	PHARMO, Netherlands
Ulrike Haug, Bianca Kollhorst	**DAP, GePaRD, Germany**	BIPS, Bremen, Germany
Helle Kieler, Karin Gembert	**DAP, Swedish registers**	CPE KI, Stockholm, Sweden
Ian Douglas, Anna Schultze	**DAP, CPRD GOLD, UK**	LSHTM, London, United Kingdom
Miguel Gil García, Miguel Ángel Maciá	**DAP, BIFAP, Spain**	AEMPS, Madrid, Spain
Gabriel Sanfélix-Gimeno, Aníbal Garcia Sempere, Isabel Hurtado, Clara Rodríguez-Bernal, Francisco Sanchez-Saez	**DAP, Valencia, Spain**	FISABIO, Valencia, Spain
Beatriz Poblador-Plou, Jonás Carmona-Pírez, Alexandra Prados-Torres, Antonio Gimeno-Miguel, Antonio Poncel-Falcó	**DAP, EpiChron, Aragon, Spain**	IACS, Zaragoza, Spain
Mitja Kos, Igor Locatelli	**DAP, Slovenia**	UL FFA, Ljubljana, Slovenia
Manuel Barreiro, Eugeni Domènech, Yamile Zabana	**DAP, ENEIDA patient registry, Spain**	GETECCU, Spain
Bas Middelkoop, Eoin McGrath, Andreu Gusi, Silvia Zaccagnino, Maria Paula Busto	**DAP, EBMT patient registry, Europe**	EBMT, Europe
Andres Metspalu, Steven Smit, Merit Kreitsberg Kadri Raav	**DAP, Estonian Biobank**	University of Tartu, Institute of Genomics, Estonia

DAP = data access provider; EU = European Union; PE&PV = Pharmacoepidemiology and Pharmacovigilance; PI = principal investigator; UK = United Kingdom.

Note: Full organisation names are included in the abbreviation section.

Contents

ABBREVIATIONS 6

GLOSSARY 8

1. Background 10

2. Framework of the metadata list and proof-of-concept catalogue 12

2.1. Initial use cases for the metadata list and proof-of-concept catalogue 12

2.2. Definition of metadata 12

2.3. Conceptual framework 13

2.4. Proposed proof-of-concept catalogue design 16

2.5. Proposed process for collection 17

3. Identification of metadata variables and definitions 19

3.1. Search and extraction of existing metadata variables and definitions 19

3.2. Structured interviews 19

4. Final list of metadata and definitions 20

4.1. Overview of proof-of-concept catalogue tables 20

5. Plans for further development of the metadata list and catalogue 25

6. References 27

7. Appendices 29

7.1. Appendix 1: List of articles and grey literature reviewed for extraction of metadata concepts and variables 29

7.2. Appendix 2: Technical description of MOLGENIS 32

ABBREVIATIONS

AEMPS	*Agencia Española de Medicamentos y Productos Sanitarios* (Spain)
ARS	*Agenzia Regionale di Sanità della Toscana* (Italy)
AsPEN	Asian Pharmacoepidemiology Network
AU-DCE	Aarhus University, Department of Clinical Epidemiology (Denmark)
BBMRI	Biobanking and BioMolecular Resources Research Infrastructure
BIFAP	*Base de datos para la Investigación Farmacoepidemiológica en Atención Primaria* (Spain)
BIPS	Leibniz Institute for Prevention Research and Epidemiology (Germany)
BPE	Bordeaux PharmacoEpi (France)
CDISC	Clinical Data Interchange Standards Consortium
CDM	common data model
CINECA	Common Infrastructure for National Cohorts in Europe, Canada, and Africa
CNODES	Canadian Network for Observational Drug Effect Studies
COVID-19	coronavirus disease 2019
CPE KI	Centre for Pharmacoepidemiology, Karolinska Institutet
CPRD	Clinical Practice Research Datalink
DAP	data access provider
DCAT	Data Catalog Vocabulary
DOI	Digital Object Identifier
EBMT	The European Society for Blood & Marrow Transplantation
EHDEN	European Health Data & Evidence Network
EJP-RD	European Joint Programme on Rare Diseases
ELIXIR	an intergovernmental organisation that brings together life science resources from across Europe
EMA	European Medicines Agency
ENCePP	European Network of Centres for Pharmacoepidemiology and Pharmacovigilance
ENEIDA	*Estudio Nacional en Enfermedad Inflamatoria intestinal sobre Determinantes genéticos y Ambientales*
EpiChron	a data source for studying chronic diseases (Spain)
ETL	extract, transform, load
EU	European Union
FAIR	findable, accessible, interoperable, and reusable
FDA	Food and Drug Administration
FISABIO	Foundation for the Promotion of Health and Biomedical Research of the Valencia Region (Spain)
GA4GH	Global Alliance for Genomics and Health
GDPR	General Data Protection Regulation (EU)
GePaRD	German Pharmacoepidemiological Research Database
GETECCU	Spanish Working Group on Crohn’s Disease and Ulcerative Colitis
GOLD	general practitioner online data source of the CPRD
HES	Hospital Episode Statistics, a data bank of the CPRD
HMA	Heads of Medicines Agencies
IACS	Instituto Aragonés de Ciencias de la Salud (Spain)
IMI	Innovative Medicines Initiative
LOINC	Logical Observation Identifiers Names and Codes
LSHTM	London School of Hygiene and Tropical Medicine
MIABIS	Minimum Information About BIobank Data Sharing

MINERVA	Metadata for data dIscoverability aNd study rEplicability in obseRVAtional studies
MOLGENIS	the software platform that will be used for the metadata catalogue
NCBO	National Center for Biomedical Ontology (United States National Institutes of Health)
Nictiz	National Competence Centre for Electronic Exchange of Health and Care Information (Netherlands)
OMOP	Observational Medical Outcomes Partnership
PHARMO	PHARMO Institute for Drug Outcomes Research (Netherlands)
PMDA	Pharmaceuticals and Medical Devices Agency (Japan)
PRAC	Pharmacovigilance Risk Assessment Committee
RTI-HS	RTI Health Solutions
RWE	real-world evidence
SIGMA	contract-based alliance of ENCePP research centres
SNDS	*Système National des Données de Santé* (France)
UK	United Kingdom
UL FFA	University of Ljubljana, Faculty of Pharmacy (Slovenia)
UMCG	University Medical Centre Groningen (Netherlands)
UMCU	University Medical Centre Utrecht (Netherlands)
UU	Utrecht University (Netherlands)
VAC4EU	Vaccine monitoring Collaboration for Europe

GLOSSARY

- Catalogue: The set of standardised tools used to access, search, and visualise metadata to illustrate each data bank and data source.
- Catalogue domain: The catalogue is envisioned to consist of six main domains of metadata, organised in sub-tables: *Institution, Network, Data source, Data bank, Common data model, and Study*.
- Catalogue table: An underlying table in the envisioned metadata catalogue, comprising a collection of multiple metadata variables. Each domain of the catalogue consists of one or more catalogue tables.
- Common data model (CDM): Common structure and format for data that allows for an efficient execution of programs against local data.
- Contributor: An institution that contributes content to the metadata catalogue.
- Data access provider (DAP): An organisation authorised to obtain access to and/or receive extracts from one or multiple data banks (e.g., for the purpose of research or surveillance).
- Data bank: Data collections sustained by a specified organisation, which is the data originator. The data bank is defined by the underlying population that can potentially contribute records to it, the prompt that leads to creation of a record in the data bank, and the data model of the data bank. Examples of data banks and their relation to data sources and DAPs is provided in document Section 2.3.
- Database: A data structure that stores organised information. Most databases contain multiple tables, which may each include several different fields. A database can host data from one or more data banks.
- Data characterisation: The summarisation of features of a data bank or data set, including quantitative measures of completeness, frequency, and quality.
- Data controller (joint controllers): The organisation(s) that determine(s) the purposes for which and the means by which personal data are processed.
- Data originator: An organisation that sustains the collection of records in a data bank (e.g., a healthcare payer).
- Data processor: An organisation that processes personal data only on behalf of the data controller. The data processor is usually a third party, external to the data controller.
- Data source: All the data banks referring to the same or partially overlapping underlying population that a given DAP can access (or receive extracts from) and link to one another at an individual level for the purpose of a study. For example, a DAP may purchase a licence to obtain access to one or both of the CPRD primary care data banks (Aurum, GOLD), the Hospital Episode Statistics (HES) data bank, and the Death Registration data bank. The DAP is said to provide access to the "CPRD" data source, for this study, including those three or four data banks. The DAP may also have access to an extraction from the Danish National Register, including the Danish National Patient Register data bank, Danish National Prescription Register data bank, and Medical Birth Register data bank. The resulting data source will be composed of those three data banks. A data source may or may not have a designated name; in this document, data source names are represented in quotation marks. Examples of data sources and their relation to data banks and DAPs is provided in document Section 2.3.

- Extract, transform, load (ETL): A repeatable process for converting data from one format to another, such as from a source format to the CDM. In this process, mappings to the standardised dictionary are added. It is typically implemented as a set of automated scripts.
- FAIR (findable, accessible, interoperable, and reusable) principles:
  - *Findability*: Any (healthcare) database that is used for analysis should, from a scientific perspective, persist for future reference and reproducibility. A comprehensive record of the database in terms of purpose, sources, vocabularies and terms, access-control mechanisms, licence, consents, etc., should be available.
  - *Accessibility*: Data should be accessible through a standardised and well-documented method.
  - *Interoperability*: The use of a CDM, standardised dictionary, and common statistical approaches should allow healthcare data from multiple data sources to be leveraged for evidence generation.
  - *Reusability*: For data to be reusable, the data licences should explicitly allow the data to be used by others, and the data provenance (understanding how the data came into existence) needs to be specified and updated as needed.
- Instance of a data source: When data are extracted from a data source for a study, an **instance of the data source** is created, which remains frozen and can be processed for the purposes of a study.
- Institution: An organisation connected to one or more data sources—such as a DAP, a data originator, a data controller, a data processor—or an organisation with analytic expertise, which may contribute to the catalogue.
- Metadata: A set of data that describes and gives information about other data. More specifically, information describing the generation, location, and ownership of the data set; key variables; and the format (coding, structured versus not) in which the data are collected is needed to enable accurate identification and qualification of the exposure and outcome information available. Metadata also include the provenance and time span of the data, clearly documenting the input, systems, and processes that define data of interest. Finally, metadata include details on the storage, handling processes, access, and governance of data.
- Prompt: An event that prompts the generation of a record in a data bank.
- Semantic annotation: The process of attaching metadata about concepts (e.g., established ontologies, vocabularies, persistent identifiers, data standards) to a piece of data or metadata.
- Standardised dictionary: A tool containing a list of variables and their definitions to enable transparent and consistent content across disparate observational databases and allows efficient and reproducible observational research.
- Underlying population: The population of individuals who can potentially contribute information to a data source or data bank. This may be a geographic population or a disease- or condition-specific population within a geographic area, in the case of a registry.
- Vocabulary: Standardised medical terminologies; may be an international standard (e.g., International Classification of Diseases, Anatomical Therapeutic Chemical) or a country/region-specific system or modification.

1. Background

Identification of appropriate data sources is becoming an increasing need for regulatory decision making. Metadata are descriptive data that characterise other data to create a clearer understanding of their meaning and to achieve greater reliability and quality of information. Access to a standard and electronic set of complete and accurate metadata information can contribute to identifying the data sources suitable for a specific study, facilitate description of the data sources planned to be used in a study protocol or research proposal, and contribute to assessing the evidentiary value of the results of studies. Access to metadata could also prevent publication based on incomplete information and subsequent retraction of studies (Offord 2020; Schriml et al., 2020).

The Heads of Medicines Agencies–European Medicines Agency (HMA-EMA) joint Big Data Task Force recommended "to promote data discoverability through the identification of metadata" as part of its Recommendation III: "*Enable data discoverability. Identify key meta-data for regulatory decision making on the choice of data source, strengthen the current European Network of Centres for Pharmacoepidemiology and Pharmacovigilance (ENCePP) resources database to signpost to the most appropriate data, and promote the use of the FAIR principles (Findable, Accessible, Interoperable and Reusable)*" (HMA-EMA, 2020). This goal is also included in the 2020-2021 Work Plan of the HMA-EMA joint Big Data Steering Group (HMA-EMA Big Data Steering Group, 2020).

In November 2020, the project "Strengthening Use of Real-World Data in Medicines Development: Metadata for Data Discoverability and Study Replicability" (EU PAS register number EUPAS39322) was initiated. The specific objectives of this project, as listed in the EMA technical specifications document (EMA, 2021), are as follows:

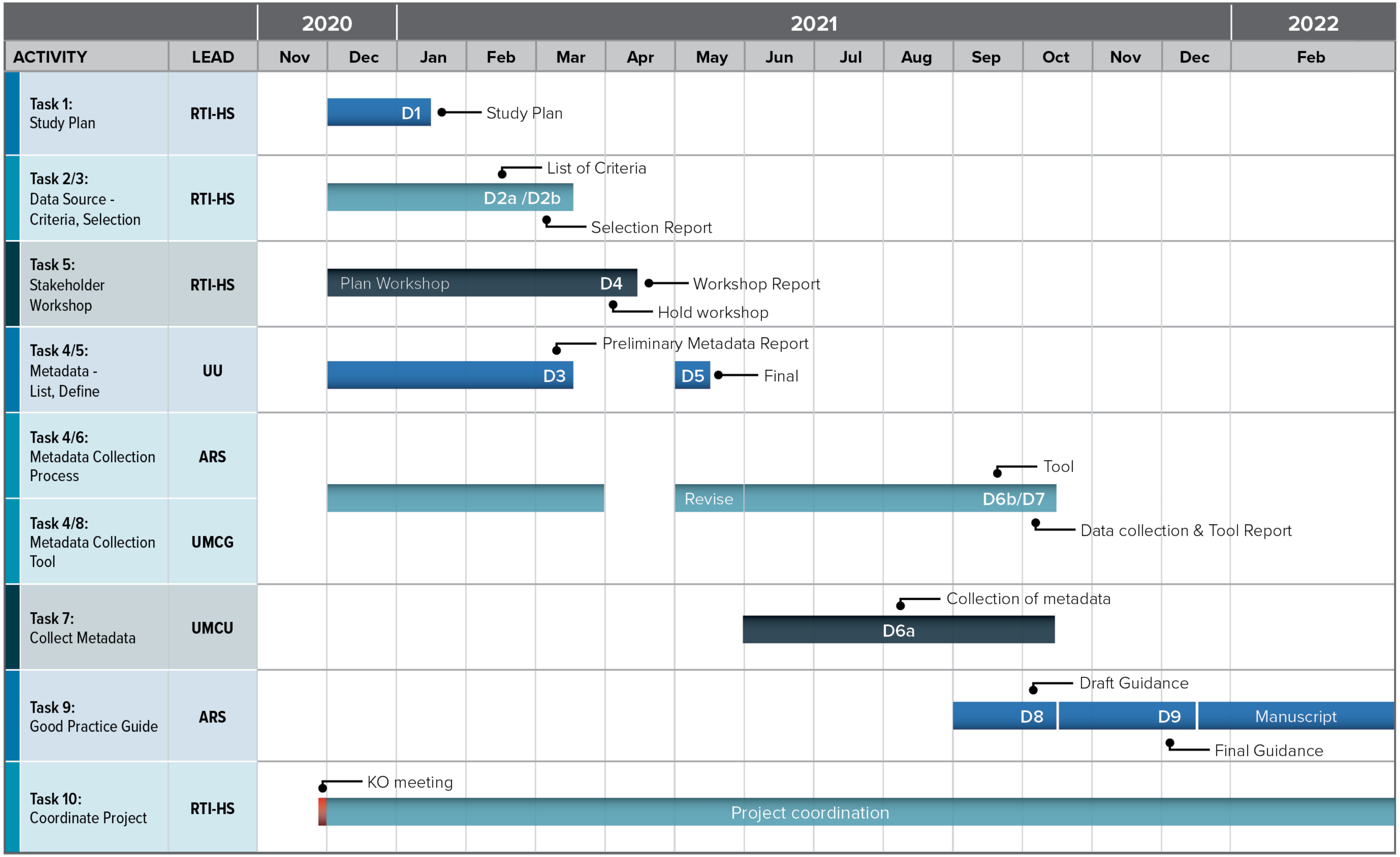
1. To define a list of criteria to identify relevant real-world data sources from which the data sources to be included in this study will be selected
2. To identify a list of minimum 10 data sources to use in the study
3. To define a set of metadata that should be collected from real-world data sources
4. To conduct an in-depth stakeholders' consultation on the metadata identified
5. To define a process to collect the set of metadata for the data sources included in the study
6. To collect the defined set of metadata for data sources included in the study
7. To develop or provide a tool enabling access to the metadata collected (e.g., through a dynamic dashboard)
8. To draft a good practice guide with the description of the metadata defined and recommendations on the use of metadata for the purpose of identifying real-world data sources for a specific study purpose

The deliverables of the project, as per the EMA technical specification document, are as follows:

1. Study plan
2. List of criteria & selection of databases
3. Preliminary set of metadata including definition
4. Organisation and conduct of stakeholders' consultation
5. Final list of metadata with definitions
6. Collection of metadata & tool
7. Report on data collection process & tool
8. Draft good practice guide
9. Good Practice Guide

An overview of proposed activities and timelines of this pilot project is provided in Figure 1.

Figure 1. Overview of the proposed activities and timelines of the pilot project



D indicates deliverable number.

This report "Final set of metadata and definitions, process and catalogue tool" corresponds to Deliverable 5 of the project. This report consists of an update of the "Preliminary set of metadata and definitions, process and catalogue tool" (Deliverable 3 of the project) based on feedback from the stakeholder consultation virtual workshop on 12 April 2021 and stakeholder feedback obtained through a survey sent prior to the workshop.

In the second half of the pilot project, a proof-of-concept metadata catalogue, with the aim of providing *living* support for investigators and evidence consumers, will be developed and piloted. The final metadata data list described in this report is designed to capture content describing the complexity of the European landscape in terms of routinely collected electronic health data and the underlying diverse healthcare systems. This will constitute the basis of the proof-of-concept catalogue, with the intention of supporting the capture of rich information at the following levels: the institutions with expertise relating to European data sources captured in the catalogue; the data sources themselves; the data banks that make up a data source (see Section 4.1 for details); tools, such as CDMs, which have an existing integration with the captured data sources; and individual study-level information, which enables the collection of rich quantitative metadata. This design builds on top of a range of H2020 and Innovative Medicines Initiative (IMI) projects, notably EUCAN Connect (2021), LifeCycle (2021), CINECA (Common Infrastructure for National Cohorts in Europe, Canada, and Africa) (2021), Athlete (2021), LongITools (2021), and ConcePTION (Dodd et al., 2020, ConcePTION 2021), which in turn build upon the infrastructure of the MOLGENIS catalogue (Swertz et al., 2010).

2. Framework of the metadata list and proof-of-concept catalogue

2.1. *Initial use cases for the metadata list and proof-of-concept catalogue*

The proof-of-concept catalogue is envisioned to be the basis upon which a system could be built that will make working with data sources easier and that should also allow regulators and prospective investigators to identify data sources that meet the requirements for specific research projects. The final metadata list was derived on the basis of the following use cases, all under FAIR principles.

- The catalogue tool could support the findability (discoverability) of data that are suitable to be used in a study to address a specific regulatory question.
- The catalogue tool could be used to assist investigators, data access providers (DAPs), and evidence consumers throughout the process of data source discovery and feasibility assessment, design of the protocol and statistical analysis plan, data processing, and development of the analysis script.
- The catalogue tool could support the dissemination and evaluation of the value and reproducibility of resulting evidence and strengthen the current ENCePP resources database.

2.2. Definition of metadata

Metadata are herein defined as "a set of data that describes and gives information about other data," more specifically, information describing:

- Location, ownership, governance of the data
- Processes underlying the storage, handling, and access of data
- Generation, provenance, and time span of the data, including the input, systems, and processes that define data of interest and the population that it covers

- Descriptors of variables captured in the data, e.g., descriptors of the underlying population
- Indicators of quality and completeness (data characterisation)
- Format (structure, model, coding) in which the data are collected
- Relevant publications that describe a data source and its suitability for research (e.g., applied studies, validation studies)
- Information on relevant contacts who provided the metadata and/or can provide more information

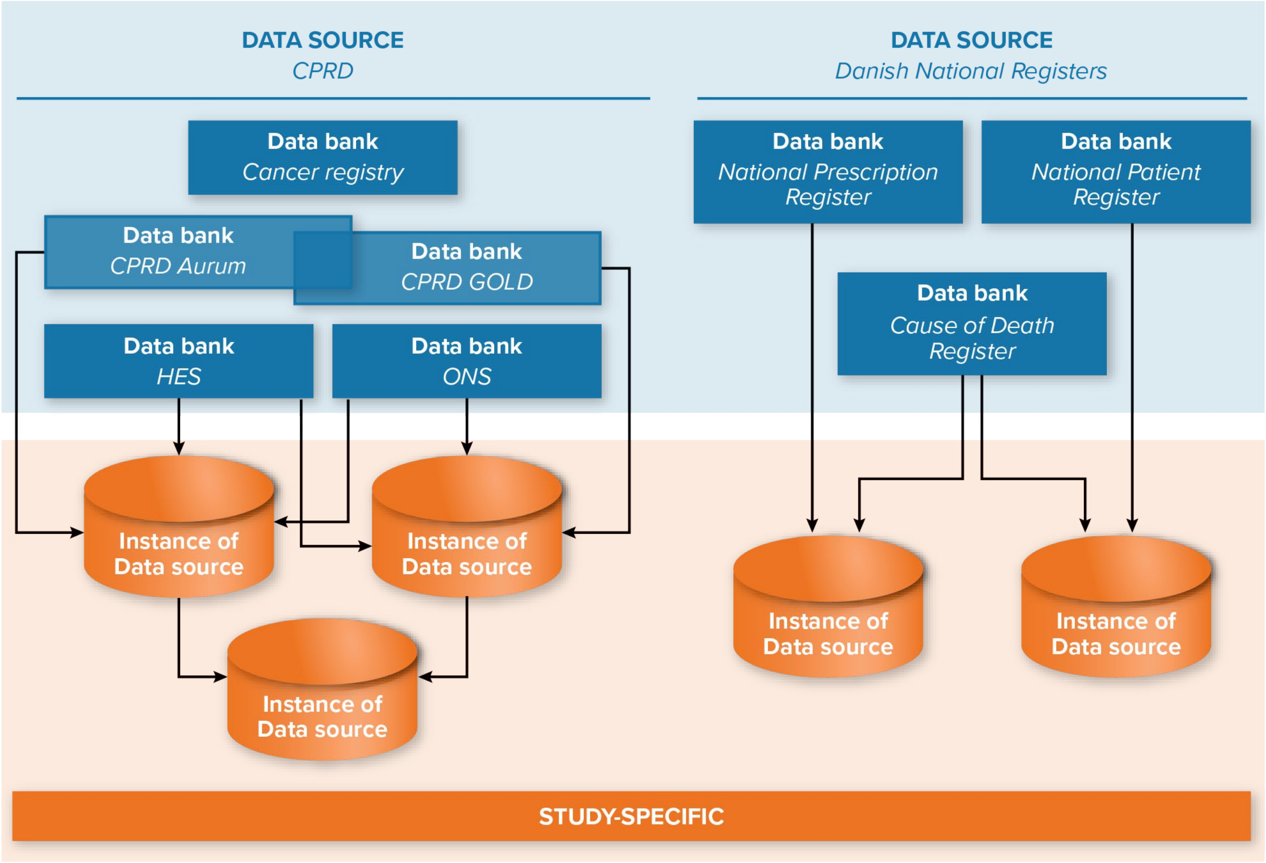
2.3. Conceptual framework

The **conceptual framework** beyond the definition of metadata stems from acknowledging that the mechanisms that put data into existence are heterogeneous across data sources and even within data sources. **Data sources** are composed of ***data banks***, which are data collections sustained by a specified organisation (e.g., a payer in a healthcare system, such as an insurance company or government; a network of clinicians; a public health or another government institution). Each data bank comes with the following:

- A specific class of events that prompt the creation of a record of an event (e.g., access to an emergency room; a visit to a primary care centre; the dispensing of a medication)
- A specific population (e.g., the persons entitled to receive healthcare assistance funded by a specific payer; persons assisted by a specific network of primary care centres; legal inhabitants of a region or country) whose events prompt the creation of a record in the data bank
- A data model and data dictionary

A ***DAP*** is an organisation that can obtain access to one or multiple **data banks**, typically subject to authorisation for use and/or data transmission of a subset of the data banks, based on a study protocol. All the data banks referring to the same study population that the DAP can access and link to each other at an individual level form a **data source**. When data are extracted from a data source for a study, an **instance of the data source** is created, which remains frozen and can be processed for the purposes of a study. The relationship between data sources and data banks is illustrated in Figure 2.

Figure 2. Illustration of the conceptual framework underlying the catalogue design



CPRD = Clinical Practice Research Datalink; CPRD GOLD = General practitioner online database of the CPRD; HES = Hospital Episode Statistics; ONS = Office for National Statistics.

The core of the metadata catalogue will be a set of standardised tools to illustrate each data bank, based on metadata associated with the above conceptual framework. This will enable a deep comprehension of the strengths and limitations of each data source in addressing specific research questions.

2.3.1. Example 1: CPRD (Clinical Practice Research Datalink) (United Kingdom)

DAP: Any institution with licence or permission to access CPRD (GOLD or Aurum) data in a study, e.g., LSHTM (London School of Hygiene and Tropical Medicine), UU (Utrecht University), etc.

Data banks: CPRD (GOLD or Aurum), primary care data; Hospital Episode Statistics (HES); Death Registration (Office for National Statistics. ONS); cancer registration data, etc.

Data source: "CPRD"

LSHTM may purchase a licence to obtain access to a CPRD primary care data bank (GOLD or Aurum) and may additionally request an extraction from the HES data bank and the Death Registration data bank from the ONS. The DAP is said to provide access to the "CPRD" data source, including those three data banks. A second DAP, UU, may purchase a licence to obtain access to only the primary care data bank and an extract from the HES data bank. In this case, this DAP is said to provide access to the "CPRD" data source, including those two data banks.

2.3.2. Example 2: Danish National Registers (Denmark)

DAP: An institution that applies for and receives time- and population-limited access to a protocol-based extraction from selected Danish National Registers following approval of a request.

Data banks: Danish National Patient Register, Danish National Prescription Register, Cause of Death Register, Danish Medical Birth Register, etc.

Data source: "Danish National Registers"

A DAP (e.g., Aarhus University or other research organisations) who fulfils all legal and data protection requirements to access an extraction (e.g., from the Danish National Patient Register data bank, Danish National Prescription Register data bank, and Danish Medical Birth Register data bank) is said to provide access to the "Danish National Register" data source, composed of those three data banks. Within a given study, a DAP for the Danish National Registers shares only aggregated data externally.

2.3.3. Example 3: ARS Toscana (Italy)

DAP: ARS Toscana

Data banks (*translated from Italian*): Registration with Healthcare System, Hospital Discharge Records, Exemptions from co-payment, Diagnostic Tests or Procedures Reimbursement, Pharmacy Dispensation Records, etc.

Data source: "ARS Toscana"

ARS Toscana has permission to obtain access to the Registration with Healthcare System, Hospital Discharge Records, Exemptions from co-payment, Diagnostic Tests or Procedures Reimbursement, and Pharmacy Dispensation Records data banks. The DAP is said to provide access to the "ARS" data source, composed of those four data banks.

2.3.4. Example 4: The European Society for Blood and Marrow Transplantation (EBMT) Registry (Multinational)

DAP: A working party within the EBMT, such as the EBMT Cellular Therapy & Immunobiology Working Party

Data banks: EBMT registry

Data source: "EBMT registry"

The working party within the EBMT gains approval for a study to be implemented. This permits a study to be conducted in the "EBMT registry," which consists of a single data bank comprising data collected from multiple forms.

2.3.5. Example 5: The Estonian Biobank (Estonia)

DAP: University of Tartu

Data banks: The Estonian Biobank, population register, Estonian Causes of Death Registry, Estonian Cancer Registry, Estonian Tuberculosis Registry, Estonian Health Insurance Fund, etc.

Data source: “The Estonian Biobank”

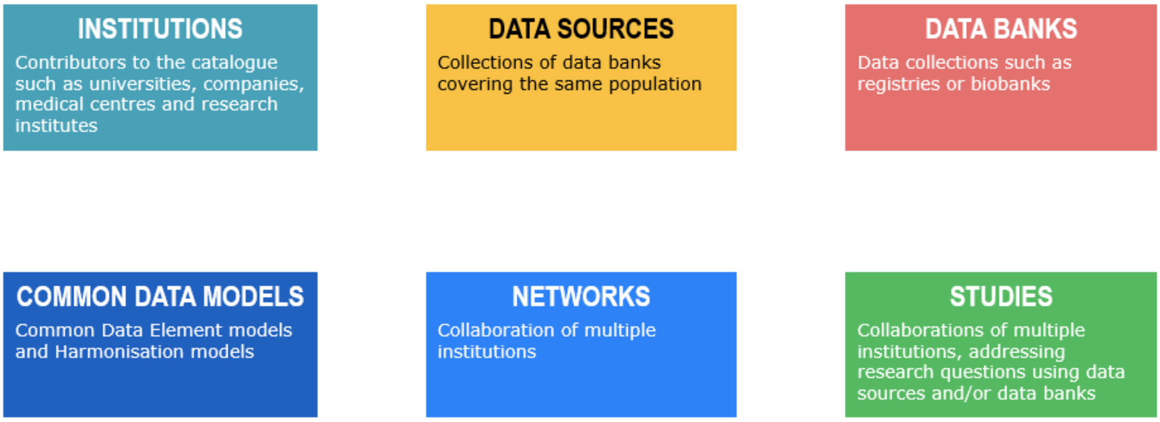
The University of Tartu has permission to obtain access to the Estonian Biobank after approval by the ethics committee. It requests and receives access to the Biobank data bank containing linked data from the population register, Estonian Causes of Death Registry, Estonian Cancer Registry, Estonian Tuberculosis Registry, and Estonian Health Insurance Fund data banks. The DAP is said to provide access to the “Estonian Biobank” data source, composed of six data banks.

2.4. Proposed proof-of-concept catalogue design

The proof-of-concept catalogue for this pilot project will be constructed based on the final list of metadata and definitions presented as the embedded Excel file in Section 4 of this document.

The proof-of-concept catalogue will consist of six domains, accessible on its home page (see Figure 3). Each tab will direct users to one of the domains, comprising multiple metadata tables, as described in Section 7.2, Appendix 2. Tools for visualisation of quantitative data are envisioned to be integrated into the platform.

Figure 3. Mock-up of catalogue home page



INSTITUTIONS

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

COMMON DATA MODELS

Common Data Element models and Harmonisation models

NETWORKS

Collaboration of multiple institutions

STUDIES

Collaborations of multiple institutions, addressing research questions using data sources and/or data banks

EMA = European Medicines Agency; EU = European Union.

The catalogue services will support the hosting of information describing the extract, transform, and load (ETL) processing of data from data sources to CDMs and details of the respective CDMs. In turn, this resource will facilitate the use of CDMs to support the execution of quality measures on the data source at the study level. The catalogue will be specifically designed to support multiple CDMs frequently used in Europe, including, but not necessarily limited to, the ConcePTION CDM, the OMOP (Observational Medical Outcomes Partnership) CDM, and the Nordic CDM used in the CARING project (Andersen et al., 2015a, Andersen et al., 2015b, But 2017, Trifirò et al., 2019).

2.4.1. Proof-of-concept catalogue implementation

To maximise interoperability of the proof-of-concept catalogue, in addition to the complete list of metadata and definitions presented herein, prior to creation of the proof-of-concept catalogue, the selected metadata items presented in this document will be compared with existing and emerging standards for common data elements and code systems/ontologies—such as MIABIS (Minimum Information About BIobank data Sharing); CDISC (Clinical Data Interchange Standards Consortium) standards, which are required for regulatory submissions to the FDA (Food and Drug Administration) (United States) and PMDA (Pharmaceuticals and Medical Devices Agency) (Japan); DCAT (Data Catalog Vocabulary), as used in the FAIR data point network; LOINC (Logical Observation Identifiers Names and Codes)—working with the standardisation efforts of existing initiatives such as GO FAIR, BBMRI (Biobanking and BioMolecular Resources Research Infrastructure), ELIXIR, GA4GH (Global Alliance for Genomics and Health), EJP-RD (European Joint Programme on Rare Diseases), and national standardisation organisations such as Nictiz (Dutch Competence Centre for Electronic Exchange of Health and Care Information). Based on the experience in these projects, the metadata will be semantically annotated, and persistent identifiers will be added inside the catalogue tool (Figure 1, Deliverable 6a). This will enable users to search the catalogue contents in combination with other catalogues and allow search robots to find the data, increasing impact and exposure of the catalogue. Semantic annotation means we will tag data inside the catalogue using established ontologies (e.g., the National Center for Biomedical Ontology BioPortal [NCBO BioPortal], 2021) that are also used by other catalogues. Persistent identifiers mean that each record has a stable URL/web address so that the record can always be found, even if the catalogue changes (e.g., digital object identifiers (DOI) for scientific publications). A recent example that demonstrates how this works in practice is the 'FAIR data points' deployed in the 'Virus Outbreak Data Network' to enable data findability to aid research in light of the coronavirus disease 2019 (COVID-19) pandemic (GO FAIR, 2021). Other examples are the 'schema.org' (Schema.org, 2021) and 'bioschemas.org' (Bioschemas.org, 2021) standards established by search engine providers that enable search results to be shown in a semantically relevant way.

The proof-of-concept catalogue tool for this pilot project will be built on MOLGENIS, an open-source data platform recommended by ELIXIR as an interoperability resource that can be flexibly configured to accommodate collected metadata. It allows for rapid prototyping and testing. Additionally, MOLGENIS is tailored to adhere to the "FAIR" principles. Therefore, it will deliver findability interfaces for human users as well as for programmatic access (including FAIR 'data points' that are now upcoming in COVID-19 research), different modes to access the data and interoperate with the catalogue tool, and support for semantic annotation to ensure metadata collected can be made interoperable with other catalogues where appropriate. A technical description of the MOLGENIS platform can be found in Section 7.2, Appendix 2.

2.5. Proposed process for collection

A proof-of-concept process for sustainable collection of metadata in the catalogue, including quality control, will be developed for this pilot project (Figure 1, Deliverable 6a). Two processes have been initially proposed and will be subject to further investigation during the rest of this pilot project.

A preliminary proposal for metadata collection and maintenance is that, when a study is conducted (e.g., on request by EMA or PRAC [Pharmacovigilance Risk Assessment Committee]), the study investigators would use the catalogue throughout the study life cycle, and catalogue updates would be an active and funded part of the study itself. The following aspects would be considered:

- The funding, which would be requested from the study funder

- Each entry describing data banks or data sources in the catalogue would be associated with a specific study or institution. Users would need to be able to view and compare multiple entries of metadata for the same data banks and data sources. At this stage, resolution of differences is not proposed.
- In many cases, the catalogue would be first populated at the time of a study, as is done in the EU PAS Registry. In the case of "emerging" data sources, which may not yet be routinely used in studies, a different route to support initial metadata entry may be required.

A second process for metadata collection and maintenance would be a centrally coordinated and funded effort to be implemented on a periodic basis, independent of a study. The following aspects would be considered:

- The strategy to determine which institution(s) would receive funding to do this effort.
- Non-native variables, e.g., the presence or absence of a disease, generally created through a study when an investigator selects and applies an algorithm to combine multiple codes (e.g., diagnostic, procedural, or treatment).
- The governance structure.

A thorough exploration of these options is scheduled for Task 6 and will be discussed in Deliverables 8 and 9 of this pilot project.

The final metadata list contains the following two types of metadata:

- Variables that need to be entered by individuals from institutions with expertise with the respective data sources and data banks
- Descriptors that are to be extracted from the data sources and data banks themselves and may support automation, including leveraging a CDM as available

The latter may take advantage of where a data source has been mapped to a CDM. Mapping of a data source to a CDM will specifically support the extraction of quantitative metadata by the use of a common syntax, making it feasible for demographic characteristics to be quantified and the results of data characterisation scripts to be visualised in dashboards in the catalogue.

The proof-of-concept catalogue will be specifically designed to support multiple CDMs used in Europe (e.g., the ConcePTION CDM, the OMOP CDM, and the Nordic CDMs). The proof-of-concept catalogue is expected to be populated mostly with data sources that have a mapping to a CDM; however, the catalogue also allows data sources without such mapping to be described. Population of the proof-of-concept catalogue will occur whenever a study is activated encompassing a data source, a process that would generally require mapping to a CDM, as per strategies C or D in Gini et al. (2020). For the purposes of this pilot project, we will develop the process and pilot it with at least one representative data source for each CDM. We will also explore a pragmatic approach for a data source that is not mapped to a supported CDM; for example, mapping the data source to the most similar, supported CDM (e.g., if the data source is composed of data banks that are Italian administrative databases, use the TheShinISS CDM); if no such CDM is available, the process would map the data source to an existing CDM, e.g., the ConcePTION CDM, limited to the information necessary for data retrieval.

3. Identification of metadata variables and definitions

3.1. Search and extraction of existing metadata variables and definitions

The preliminary list of metadata and definitions was derived from existing resources from relevant organisations that have described the metadata of health data sources, combined with information that was gathered from structured interviews with research organisations and consortia with worldwide coverage.

Literature materials were gathered by a search of the websites of key organisations and consortia considered as potential sources of information on relevant metadata. Documents and webpages providing information related to metadata were collected in a literature library. Information on metadata were extracted from the documents in the library using a data extraction template. Section 7.1, Appendix 1, contains a list of source materials included in the extraction.

3.2. Structured interviews

Next, a series of structured 60-minute interviews were conducted with representatives from seven organisations or consortia with experience in conducting pharmacoepidemiologic studies with multiple data sources or with expertise in metadata in the health domain:

- FDA Sentinel Initiative (United States Food and Drug Administration, 2021)
- CNODES (Canadian Network for Observational Drug Effects Studies) (Cnodes.ca, 2021)
- IMI EHDEN (European Health Data Evidence Network) (Ehden.eu, 2021)
- IMI ConcePTION (Imi-conception.eu, 2021)
- AsPEN (Asian Pharmacoepidemiology Network) (Aspennet.asia, 2021)
- IMI FAIRplus (Fairplus-project.eu, 2021)
- Maelstrom (Maelstrom-research.org, 2021)
- Aetion (Real-World Evidence Solution | RWE Analytics) (Aetion, 2021)

The interviews followed a standard set of questions, which were piloted in the first interview with representatives from ConcePTION. The information from the interviews was used to i) refine the scope of metadata envisioned for the catalogue, ii) identify additional key metadata variables to collect, iii) gather additional resources that could be used to inform the metadata list, iv) identify existing examples of tools that can be used to access or visualise metadata, and v) identify potential challenges or barriers to the implementation of the envisioned catalogue.

The final list of metadata and definitions presented in Section 4 is based on feedback obtained from the stakeholder consultation virtual workshop held on 12 April 2021 and stakeholder feedback obtained through a survey sent prior to the workshop. A summary of the feedback obtained during the workshop is presented in Deliverable 4 of this project (posted for the public on the EMA website). Stakeholder feedback was conflicting with respect to the length and complexity of the metadata list, with some stakeholders providing numerous suggestions for additions to the catalogue and others expressing concern that the metadata list may need simplification. To take this into account, suggested additions have been included (where agreed on by the MINERVA consortium and the Agency), and a short list of priority metadata has been maintained from the preliminary version, which identifies key metadata variables that should be prioritised during population of the catalogue by a contributor.

4. Final list of metadata and definitions

The embedded Excel file contains the final list of proposed metadata and definitions. The structure of the file and tables is described below with a high-level overview of the metadata in each table. The primary language of the catalogue will be English; entries in other languages would be permitted for names, geographic locations, or publications.

Double-click the icon to open the file in a separate window.

The metadata list is presented in the Excel file in two formats.

First, the metadata variables are organised in *tables* within six domains envisioned to be presented on the homepage of the proof-of-concept catalogue (Figure 3): Institution, Data source, Data bank, Common data model, Network, and Study. An additional catalogue-entry table is presented to represent metadata that will be collected on the metadata variables on entry to the proof-of-concept catalogue. Each domain of the proof-of-concept catalogue contains one or more tables. Each table contains the name of the metadata variables (with a hierarchical code), a description of the content, the standard allowable values, and a descriptor of how metadata will be entered in the catalogue for each variable, including variables that link tables in the catalogue and cross-populate tables.

Second, a *full list* of metadata variables is presented, labelled with catalogue domain and catalogue table.

Most metadata will be populated by catalogue contributors, either manually or, in a future implementation of the catalogue, using automated tools, which may query documentation on data sources, data banks, and studies, as well as external FAIR data catalogues (see Sections 2.4 and 2.5 for further details). In some cases, metadata entered in one table may lead to population of a variable in other tables. Persistent identifiers for institutions, data sources, data banks, CDMs, research networks, and studies would permit tables from different domains of the catalogue to be linked. As an example, when a user browses the metadata catalogue and enters the ***Study*** domain, a link between the *Study – institutions – study Institution ID* and *Institution – role – institution full name* metadata variables permits the user to navigate between the two domains of the catalogue. When a responsible person from an institution enters information in the catalogue, the equivalent information across these tables will need to be entered only once. The connections between the tables in the metadata catalogue can be seen in Figure 4 in Section 4.1 and is described in the metadata catalogue tables and metadata full list in the Excel document.

4.1. Overview of proof-of-concept catalogue tables

4.1.1. Catalogue

The Catalogue domain contains metadata that will be recorded on creation of any entry into the catalogue. This "metadata on metadata" captures the provenance of metadata in the catalogue and may be implemented in a catalogue alongside individual metadata entries (as a time stamp) or presented as an overall log of when entries are made in the catalogue. This will be integrated in the proof-of-concept catalogue by the automated creation of a time stamp for each metadata entry, as well as a label to describe who entered the metadata. In a future implementation, this metadata could be created when contributors log into the catalogue with a verified account, and a historical record of entries could be supported to facilitate version control.

M1 Catalogue – entry: This table contains metadata describing who entered information in the catalogue and when. This will not in itself be displayed as a domain on the homepage but will be used to support addition of metadata on metadata entered in each of the other tables.

4.1.2. Institution

The Institution domain contains metadata describing contributors to the catalogue and other organisations listed as an institution in the Study domain of the catalogue. This domain provides information on the institutions responsible for entries to the catalogue and on the expertise of institutions and their relations to data sources and data banks.

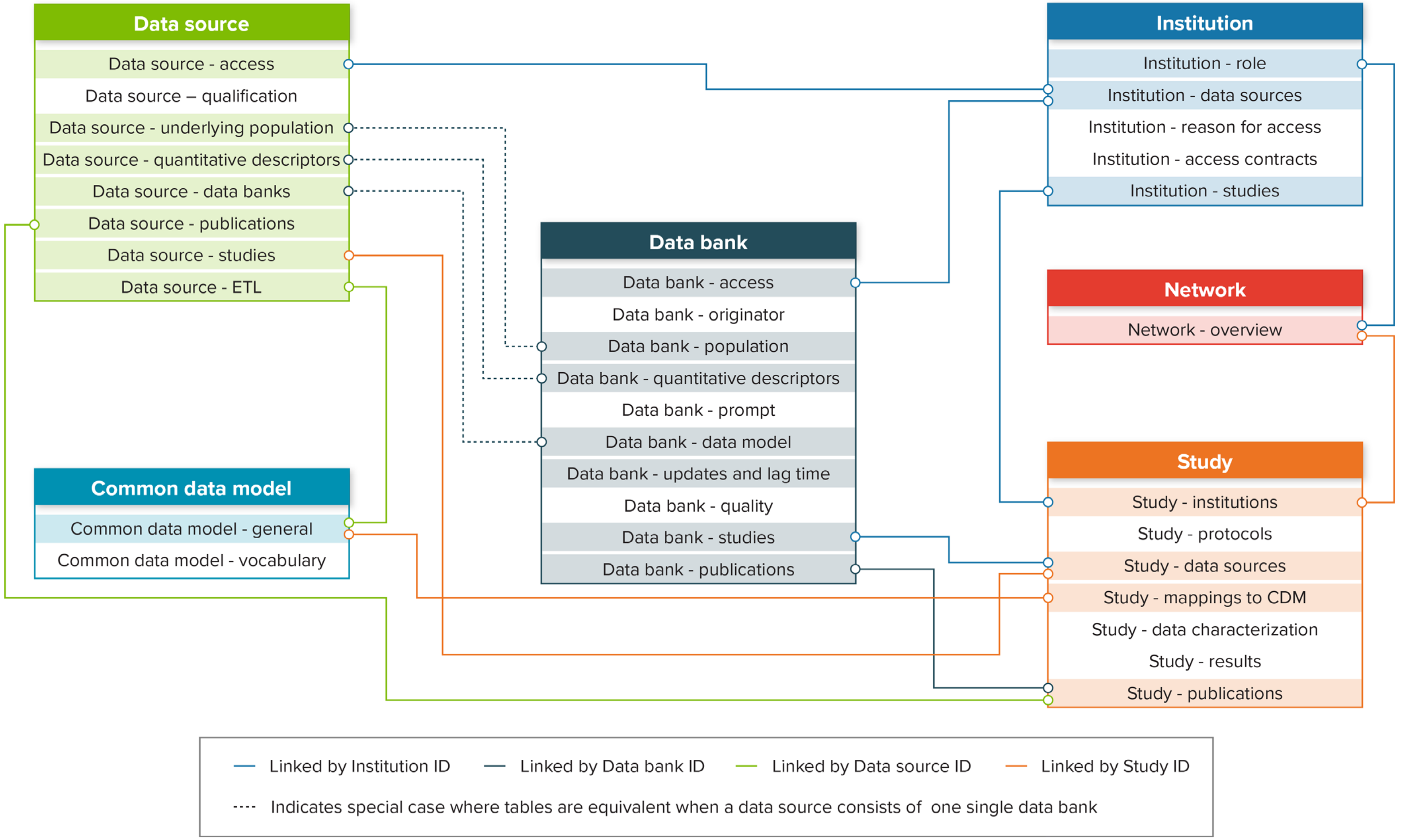
A1 Institution – role: This table contains metadata describing an institution (any institution listed as an institution in the Study table, or any other contributor to the catalogue, such as a data originator organisation), its role in studies, and expertise.

A2 Institution – data sources and data banks: This table contains metadata describing the data sources or data banks to which an institution can obtain access, as well as details on whether only a subset of the underlying population can be accessed.

A3 Institution – access contracts: This table contains metadata describing how access to data sources is permitted for the institution and the time it takes between applying for access to a data source or an extract of the data source and obtaining the access for that specific institution.

A4 Institution – studies: This table links directly to the studies entered in the *Study* domain, where the institution is listed either as the lead study institution or an additional institution that contributed to the study.

Figure 4. A high-level relational model of the metadata tables



CDM = common data model; ETL = extract, transform, load; ID = identification.

Note: Shaded rows indicate tables that are connected to other tables.

4.1.3. Data source

The data source domain describes information on collections of data banks, which contain data on a specific underlying population, as well as information on access to the data source for research purposes. The metadata tables will collect information (e.g., on population counts, demographic distributions) that can be presented in the catalogue to (1) display the information stratified on a yearly basis (if available) and (2) changes in information as entries are updated over time.

B1 Data source – underlying population: This table contains metadata describing the population that can potentially be captured in the data source.

B2 Data source – access: This table contains metadata describing the institutions that are able to access the data source, as denoted in the Institution domain of the catalogue. This table links directly to metadata provided in the *Institution – data sources* table.

B3 Data source – qualification: This table contains metadata describing any qualifications that the data source has successfully undergone, such as EMA qualification.

B4 Data source – data banks: This table contains metadata describing the data banks that make up the data source and summary information of their contents.

B5 Data source – linkage: This table contains a description of existing linkages between data banks in the data source.

B6 Data source – quantitative descriptors: This table contains numerical summaries of the data source population.

B7 Data source – ETL: This table contains metadata describing existing extract-transform-load specifications for mapping of the data source to a CDM.

B8 Data source – studies: This table links the data source to any studies that listed it as a data source in the *Study – data sources* table.

B9 Data source – publications: This table contains metadata describing publications relevant to the data source, such as those that describe the contents of the data source.

4.1.4. Data bank

The data bank domain describes information on individual data banks, their content and what prompts records to be created, the underlying population, and the data model of the data bank. The metadata tables will collect information (e.g., on population counts, demographic distributions) that can be presented in the catalogue to (1) display the information stratified on a yearly basis (if available) and (2) changes in information as entries are updated over time. In the catalogue it will be possible to identify data banks covering the same or an overlapping underlying population in two ways. First, the Data source domain will show visible links where data banks have already been linked together in a study and will contain information on existing linkages. Second, the Data bank domain, will show where data banks cover an underlying population from the same geographic population, as well as information on unique identifiers in the data banks (which could provide a potential linkage key). In a future implementation of the metadata catalogue, this could potentially be visualised to support data discoverability.

C1 Data bank – underlying population: This table contains metadata describing the population that can potentially be captured in the data bank.

C2 Data bank – access: This table contains metadata describing the list of institutions entered in the catalogue that are able to obtain access to the data bank, as denoted in the Institution domain of the catalogue. This table links directly to the *Institution – data sources* table.

C3 Data bank – qualification: This table contains metadata describing any qualifications that the data bank has successfully undergone, such as EMA qualification.

C4 Data bank – originator: This table contains metadata describing the institution or body that sustains or maintains the collection of records in the data bank.

C5 Data bank – prompt: This table contains metadata describing the event(s) that trigger(s) the creation of a record in the data bank.

C6 Data bank – data model and contents: This table contains metadata describing the data model (i.e., standardised dictionary) of the data bank, including vocabulary.

C7 Data bank – quantitative descriptors: This table contains numerical summaries of the data bank population.

C8 Data bank – updates and lag time: This table contains metadata describing the regularity of updates and time lags of the data bank.

C9 Data bank – quality: This table contains metadata describing qualitative descriptions of quality and qualitative and quantitative descriptors of completeness of the data bank.

C10 Data bank – studies: This table links the data bank to any studies that listed it as a data bank in the *Study – data sources and data banks* table.

C11 Data bank – publications: This table contains metadata describing publications relevant to the data bank, such as those that describe the contents of the data bank.

4.1.5. Common data model

The CDM domain describes information on CDMs to which data sources described in the catalogue have been converted. Details on the CDM mappings can be found in DS – ETL and Stu – mappings

D1 Common data model – general: This table contains metadata describing CDMs utilised within studies captured in the catalogue.

D2 Common data model – vocabulary: This table contains metadata describing standard vocabularies used in the CDM.

4.1.6. Network

The network domain provides descriptions of research networks that include member institutions described within the Institutions domain. This provides information on how data sources, as well as expertise, may be brought together for the purpose of a study and could potentially interface with the ENCePP Resources Database.

E1 Network – overview: This table contains metadata describing networks/consortia linking to institutions and studies in the catalogue.

4.1.7. Study

The Study domain describes studies conducted using data sources described in the catalogue. This should interface with the EU PAS Register to link information on studies with information on the use, content, and quality of data sources, and their respective data banks, described in the catalogue. It is envisioned that, providing both the metadata catalogue and EU PAS Register operate on FAIR principles, it would be possible for metadata on studies to be imported from the EU PAS Register to the metadata catalogue, whenever a study is entered in the EU PAS Register, and where the information (variables) in the two catalogues overlaps.

F1 Study – institutions: This table contains metadata describing institutions involved in the study.

F2 Study – protocols: This table contains metadata describing protocols related to the study.

F3 Study – data sources and data banks: This table contains metadata describing data sources and data banks included in the study and software used for extraction and processing of data.

F4 Study – mappings to CDM: This table contains metadata describing mappings that have been done to CDMs in the study.

F5 Study – data characterisation: This table contains quantitative descriptors of the completeness of data and numeric quality indicators, if generated within the study (e.g., level 1-3 checks on the CDM instance). It is envisioned that the catalogue will support standard tables to which this information can be transformed. Standard data characterisation scripts or programmes can also be supported, for DAPs to run on their data (if converted to a CDM in the catalogue) and return standard measures of completeness of the CDM tables. Metadata entered in this table can be transformed into a dashboard with visualisations.

F6 Study – results: It is envisioned that the catalogue would support standard tables in which study results could be entered. Metadata entered in these tables can be transformed into a dashboard with visualisations.

F7 Study – publications: This table contains metadata describing publications generated by the study.

5. Plans for further development of the metadata list and catalogue

The final metadata list presented herein represents a final list of the metadata content that is proposed to be collected and piloted in the catalogue. The following additional steps will be taken during the development of the proof-of-concept catalogue, to ensure that the metadata list follows the FAIR principles and is adherent to recognised international standards and to clarify potential strategies for implementing the catalogue in practice.

First, the content variables in the metadata list will be mapped to international standards using existing ontologies. This will include mapping of terminology used in the catalogue to applicable standards from those proposed in Section 2.4.1. Where there is no existing vocabulary for a necessary concept, new terminology will be submitted to be added to these standards. Additional suggested standard vocabularies for metadata variables will be included based on international standards (such as the International Organization for Standardization standards for identification of medicinal products).

Second, feedback from the stakeholder workshop highlighted the need for the automation of metadata extraction and entry to the catalogue. Section 2.4 of this report proposes one possible solution, whereby CDMs can facilitate the use of tools for automated extraction of metadata. Additional solutions will be theoretically explored, and solutions for extracting quantitative information on completeness and quality will be more thoroughly investigated.

Third, as a part of the development of the proof-of-concept catalogue, preliminary dashboards will be designed for the visualisation of metadata, to support use of the catalogue.

Finally, further investigation of processes for collecting, maintaining, and sustaining the catalogue is planned alongside the development of the proof-of-concept catalogue. This will include a thorough exploration of the possible options for collecting and maintaining metadata information in the catalogue, as well as considerations for long-term sustainability, which is only briefly described in this report.

These steps will be completed during the second half of the MINERVA project. The resulting development of the metadata list underlying the proof-of-concept catalogue, with integration of international standards, will be presented alongside MINERVA project Deliverable 6 (Collection of metadata and the tool) and Deliverables 8 and 9 (draft and final Guidance).

6. References

Aetion. 2021. *Real-World Evidence Solution | RWE Analytics | Aetion*. [online] Available at: <https://aetion.com> [Accessed 1 March 2021].

Andersen M, Thinsz Z, Citarella A, et al. implementing a Nordic common data model for register-based pharmacoepidemiological research. *Nor Epidemiol*. 2015b;Sep;25(Suppl 1):P27.

Andersen M, Thinsz Z, Ekström N, et al. Use of a concept dictionary to integrate different medical classification systems in a multicountry study. *Nor Epidemiol*. 2015a;25(Suppl 1):P27.

Aspennet.asia. 2021. *Asian Pharmacoepidemiology Network*. [online] Available at: <https://aspennet.asia/> [Accessed 1 March 2021].

Athlete. 2021. *Home – Athlete*. [online] Available at: <https://athleteproject.eu/> [Accessed 1 March 2021].

Bioschemas.org. 2021. Bioschemas. [online] Available at: <https://bioschemas.org/> [Accessed 12 March 2021].

But A, De Bruin ML, Bazelier MT, et al. Cancer risk among insulin users: comparing analogues with human insulin in the CARING five-country cohort study. *Diabetologia*. 2017;60(9):1691-1703. doi:10.1007/s00125-017-4312-5.

CINECA. 2021. *CINECA | Common Infrastructure for National Cohorts in Europe, Canada, and Africa*. [online] Available at: <https://www.cineca-project.eu/> [Accessed 1 March 2021].

Cnodes.ca. 2021. *CNODES | Canadian Network for Observational Drug Effects Studies*. [online] Available at: <https://www.cnodes.ca> [Accessed 1 March 2021].

Dodd C, Rosa G, Sturkenboom M, et al., 2020. D7.5 Report on existing common data models and proposals for ConcePTION. Available at: <https://www.imi-conception.eu/wp-content/uploads/2020/10/ConcePTION-D7.5-Report-on-existing-common-data-models-and-proposals-for-ConcePTION.pdf> [Accessed 16 March 2021].

Ehden.eu. 2021. *European Health Data Evidence Network – ehden.eu*. [online] Available at: <https://www.ehden.eu/> [Accessed 1 March 2021].

EMA. 2021, *EMA | About us | How we work | Big data | Research projects*. [online] Available at <https://www.ema.europa.eu/en/about-us/how-we-work/big-data#research-projects-(new)-section> [Accessed 3 March 2021].

Eucanconnect.com. 2021. *EUCAN Connect*. [online] Available at: <https://eucanconnect.com/> [Accessed 1 March 2021].

Fairplus-project.eu. 2021. *FAIRplus | Home page*. [online] Available at: <https://fairplus-project.eu> [Accessed 1 March 2021].

Gini R, Sturkenboom MCJ, Sultana J, et al. Different strategies to execute multi-database studies for medicines surveillance in real-world setting: a reflection on the European model. *Clin Pharmacol Ther*. 2020;108(2):228-235. doi:10.1002/cpt.1833.

GO FAIR. 2021. Virus Outbreak Data Network (VODAN) – GO FAIR. [online] Available at: <https://www.go-fair.org/implementation-networks/overview/vodan/> [Accessed 12 March 2021].

HMA-EMA Big Data Steering Group. Workplan. 27 Jul 2020. Available at: <https://www.ema.europa.eu/en/documents/work-programme/workplan-hma/ema-joint-big-data-steering-group_en.pdf> [Accessed 3 March 2021].

HMA-EMA. Priority recommendations of the HMA-EMA joint Big Data Task Force. HMA-EMA Big Data Steering Committee Group; 15 Dec 2020. Available at: <https://www.ema.europa.eu/en/documents/other/priority-recommendations-hma-ema-joint-big-data-task-force_en.pdf> [Accessed 3 March 2021].

Imi-conception.eu. 2021. *ConcePTION*. [online] Available at: <https://www.imi-conception.eu/> [Accessed 1 March 2021].

LifeCycle. 2021. *Home – LifeCycle*. [online] Available at: <https://lifecycle-project.eu/> [Accessed 1 March 2021].

LongITools. 2021. *LongITools | Exposome application*. [online] Available at: <https://longitools.org/> [Accessed 1 March 2021].

Maelstrom-research.org. 2021. *Home page | Maelstrom research*. [online] Available at: <https://www.maelstrom-research.org> [Accessed 1 March 2021].

National Center for Biomedical Ontology. *NCBO BioPortal*. [online] Available at: <https://bioportal.bioontology.org> [Accessed 10 June 2021].

Offord C. 2020. The Surgisphere scandal: what went wrong? 1 October 2020. Available at: <https://www.the-scientist.com/features/the-surgisphere-scandal-what-went-wrong--67955>. [Accessed 16 March 2021].

Schema.org. 2021. [online] Available at: <https://schema.org/> [Accessed 12 March 2021].

Schriml LM, Chuvochina M, Davies N, et al. COVID-19 pandemic reveals the peril of ignoring metadata standards. *Sci Data*. 2020;7(1):188. doi:10.1038/s41597-020-0524-5.

Swertz MA, Dijkstra M, Adamusiak T, et al. The MOLGENIS toolkit: rapid prototyping of biosoftware at the push of a button. *BMC Bioinformatics*. 2010;11(Suppl 12):S12. doi:10.1186/1471-2105-11-S12-S12.

Trifirò G, Gini R, Barone-Adesi F, et al. The role of European healthcare databases for post-marketing drug effectiveness, safety and value evaluation: where does Italy stand? *Drug Saf*. 2019;42(3):347-363. doi:10.1007/s40264-018-0732-5.

United States Food and Drug Administration. 2021. FDA's Sentinel Initiative. [online] Available at: <https://www.fda.gov/safety/fdas-sentinel-initiative> [Accessed 1 March 2021].

7. Appendices

7.1. Appendix 1: List of articles and grey literature reviewed for extraction of metadata concepts and variables

Afonso, A., S. Schmiedl, Claudia Becker, S. Tcherny-Lessenot, P. Primatesta, E. Plana, P. Souverein et al., 2016. 'A methodological comparison of two European primary care databases and replication in a US claims database: inhaled long-acting beta-2-agonists and the risk of acute myocardial infarction.' European journal of clinical pharmacology 72, no. 9 1105-1116.

Ball, Christopher, and Nicholas Hudson. 2020. 'Selecting Real World Data – a Guide to Selecting the Right Real World Data to Support Your Research.' https://www.iqvia.com/-/media/iqvia/pdfs/library/white-papers/iqvia-insights-guide_selecting-real-world-data.pdf?_=1613642511752.

Berger, Marc L., Harold Sox, Richard J. Willke, Diana L. Brixner, Hans-Georg Eichler, Wim Goettsch, David Madigan, et al., 2017. 'Good Practices for Real-World Data Studies of Treatment and/or Comparative Effectiveness: Recommendations from the Joint ISPOR-ISPE Special Task Force on Real-World Evidence in Health Care Decision Making.' Value in Health 20 (8): 1003-8.

Bergeron, Julie, Dany Doiron, Yannick Marcon, Vincent Ferretti, and Isabel Fortier. 2018. 'Fostering population-based cohort data discovery: The Maelstrom Research cataloguing toolkit.' PloS one 13, no. 7: e0200926.

Brauer, Ruth, Ana Ruigómez, Gerry Downey, Andrew Bate, Luis Alberto Garcia Rodriguez, Consuelo Huerta, Miguel Gil, et al., 2016. 'Prevalence of Antibiotic Use: A Comparison across Various European Health Care Data Sources.' Pharmacoepidemiology and Drug Safety 25 (S1): 11-20.

Brown, Jeffrey S., Michael Kahn, and Sengwee Toh. 2013. 'Data Quality Assessment for Comparative Effectiveness Research in Distributed Data Networks.' Medical Care 51(8 0 3): S22-29.

Butler, Amanda Leanne, Mark Smith, Wayne Jones, Carol E Adair, Simone Vigod, Paul Kurdyak, and Alain Lesage. 2018. 'Multi-Province Epidemiological Research Using Linked Administrative Data: A Case Study from Canada.' International Journal of Population Data Science 3 (3).

Callahan, Tiffany J, Alan E Bauck, David Bertoch, Jeff Brown, Ritu Khare, Patrick B Ryan, Jenny Staab, Meredith N Zozus, and Michael G Kahn. 2017. 'A Comparison of Data Quality Assessment Checks in Six Data Sharing Networks.' 5 (1): 15.

Dodd, Caitlin, Rosa Gini, Miriam Sturkenboom, Vjola Hoxhaj, Marieke Hollestelle, Nicolas Thurin, Claudia Bartolini et al., 2020. 'IMI ConcePTION Deliverable 7.5 – Report on existing common data models and proposals for ConcePTION.' https://www.imi-conception.eu/wp-content/uploads/2020/10/ConcePTION-D7.5-Report-on-existing-common-data-models-and-proposals-for-ConcePTION.pdf [Accessed on 01 February 2021].

Doyle, Carla M., Lisa M. Lix, Brenda R. Hemmelgarn, J. Michael Paterson, and Christel Renoux. 2020. 'Data Variability across Canadian Administrative Health Databases: Differences in Content, Coding, and Completeness.' Pharmacoepidemiology and Drug Safety 29 (S1): 68-77.

European Medicines Agency. 2020. 'Guideline on registry-based studies – Draft.' < https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-registry-based-studies_en.pdf> [Accessed 01 February 2021].

European Network of Centres for Pharmacoepidemiology and Pharmacovigilance. 'ENCePP Resources Database.' Available at: http://www.encepp.eu/encepp/resourcesDatabase.jsp [Accessed 01 February 2021].

Ferrer, Pili, Elena Ballarín, Mònica Sabaté, Joan-Ramon Laporte, Marieke Schoonen, Marietta Rottenkolber, Joan Fortuny, Joerg Hasford, Iain Tatt, and Luisa Ibáñez. 2014. 'Sources of European Drug Consumption Data at a Country Level.' International Journal of Public Health 59 (5): 877-87.

Fortier, I., P. R. Burton, P. J. Robson, V. Ferretti, J. Little, F. L'Heureux, M. Deschenes, et al., 2010. 'Quality, Quantity and Harmony: The DataSHaPER Approach to Integrating Data across Bioclinical Studies.' International Journal of Epidemiology 39 (5): 1383-93.

Fortier, Isabel, Nataliya Dragieva, Matilda Saliba, Camille Craig, and Paula J. Robson. 2019. 'Harmonization of the Health and Risk Factor Questionnaire Data of the Canadian Partnership for Tomorrow Project: A Descriptive Analysis.' CMAJ Open 7 (2): E272-82.

Fortier, Isabel, Parminder Raina, Edwin R Van den Heuvel, Lauren E Griffith, Camille Craig, Matilda Saliba, Dany Doiron, et al., 2016. 'Maelstrom Research Guidelines for Rigorous Retrospective Data Harmonization.' International Journal of Epidemiology, June.

Ilomäki, Jenni, J. Simon Bell, Adrienne YL Chan, Anna-Maija Tolppanen, Hao Luo, Li Wei, Edward Chia-Cheng Lai et al. "Application of Healthcare 'Big Data' in CNS Drug Research: The Example of the Neurological and mental health Global Epidemiology Network (NeuroGEN)." CNS drugs 34, no. 9 (2020): 897-913.

Kahn, Michael G., Tiffany J. Callahan, Juliana Barnard, Alan E. Bauck, Jeff Brown, Bruce N. Davidson, Hossein Estiri, et al., 2016. 'A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data.' EGEMs (Generating Evidence & Methods to Improve Patient Outcomes) 4 (1): 18.

Lai, Edward Chia-Cheng, Kenneth K. C. Man, Nathorn Chaiyakunapruk, Ching-Lan Cheng, Hsu-Chih Chien, Celine S. L. Chui, Piyameth Dilokthornsakul, et al., 2015. 'Brief Report: Databases in the Asia-Pacific Region.' Epidemiology 26 (6): 815-20.

Lai, Edward Chia-Cheng, Patrick Ryan, Yinghong Zhang, Martijn Schuemie, N Chantelle Hardy, Yukari Kamijima, Shinya Kimura, et al., 2018. 'Applying a Common Data Model to Asian Databases for Multinational Pharmacoepidemiologic Studies: Opportunities and Challenges.' Clinical Epidemiology Volume 10 (July): 875-85.

Mack, Christina, and K. Lang. 2014. 'Using real-world data for outcomes research and comparative effectiveness studies.' Drug Discovery & Development 4 < https://chip.unc.edu/wp-content/uploads/2016/duke-unc-presentations/pdf/Mack_RW%20data_CHIP%20presentation%2011Nov%20v1.pdf> [Accessed 01 February 2021].

Mikita, J. Stephen, Jules Mitchel, Nicolle M. Gatto, John Laschinger, James E. Tcheng, Emily P. Zeitler, Arlene S. Swern, et al., 2021. 'Determining the Suitability of Registries for Embedding Clinical Trials in the United States: A Project of the Clinical Trials Transformation Initiative.' Therapeutic Innovation & Regulatory Science 55 (1): 6-18.

Observational Health Data Sciences and Informatics. 'Chapter 15 Data Quality' in The Book of OHDSI. Independently published.August 29, 2019.

Oliveira, José Luís, Alina Trifan, and Luís A. Bastião Silva. "EMIF Catalogue: a collaborative platform for sharing and reusing biomedical data." International journal of medical informatics 126 (2019): 35-45.

Sørensen, Henrik Toft, Svend Sabroe, and Jørn Olsen. 1996. 'A Framework for Evaluation of Secondary Data Sources for Epidemiological Research.' International Journal of Epidemiology 25 (2): 435-42.

Suissa, Samy, David Henry, Patricia Caetano, Colin R. Dormuth, Pierre Ernst, Brenda Hemmelgarn, Jacques LeLorier et al., 2012. 'CNODES: the Canadian network for observational drug effect studies.' Open Medicine 6, no. 4: e134.

Swertz, Morris, Maria Loane, Joanne Given, Florence Coste, Helen Dolk, David Lewis, Eugene van Puijenbroek, et al., 2020 'IMI ConcePTION Deliverable 7.1: User requirements and metadata model for the FAIR data catalogue from QP1, 2 &7 – task 7.4.' <https://www.imi-conception.eu/wp-content/uploads/2020/04/ConcePTION-D7.1-User-requirements-and-metadata-model-for-the-FAIR-data-catalogue-from-WP1-2-7.pdf> [Accessed 01 February 2021].

The National Patient-Centered Clinical Research Network. 2021. Data | The National Patient-Centered Clinical Research Network. [online] Available at: <https://pcornet.org/data/> [Accessed 16 March 2021].

US Food and Drug Administration, Sentinel Operations Center and Cross-Network Directory Service Collaborating Partners. 2018. 'Cross-network Directory Service Project – Design and Technical Documentation.' <https://www.sentinelinitiative.org/sites/default/files/Methods/CNDS_Design_Technical_Documentation.pdf> [Accessed 01 February 2021].

US Food and Drug Administration, Sentinel Operations Center and Cross-Network Directory Service Collaborating Partners. 2018. 'Cross-network Directory Service Project – Final Report.' <https://www.sentinelinitiative.org/sites/default/files/Methods/CNDS_Final_Report_v10.pdf> [Accessed 01 February 2021].

US Food and Drug Administration, Sentinel Operations Center and Cross-Network Directory Service Collaborating Partners. 2018. 'Data Quality Review and Characterization Programs – Quality Assurance (QA) Package.' <https://dev.sentinelsystem.org/projects/QA/repos/qa_package/browse> [Accessed 01 February 2021].

US Food and Drug Administration, Sentinel Operations Center and Cross-Network Directory Service Collaborating Partners. 'Sentinel | Key Database Statistics.' Available at: <https://www.sentinelinitiative.org/about/key-database-statistics#section-1593025578856> [Accessed 01 February 2021].

US Food and Drug Administration, Sentinel Operations Center and Cross-Network Directory Service Collaborating Partners. 2020. 'Standardization and Querying of Data Quality Metrics and Characteristics for Electronic Health Data – Data Quality Metrics System Final Report.' <https://www.sentinelinitiative.org/sites/default/files/Methods/Standardization_and_Querying_of_Data_Quality_Metrics.pdf> [Accessed 01 February 2021].

Van Bochove, Kees, Emma Vos, Anne van Winzum, Julia Kurps, and Maxim Moinat. 'Implementing FAIR in OHDSI: Challenges and Opportunities for EHDEN,' <https://www.ohdsi.org/wp-content/uploads/2020/05/Implementing-FAIR-in-OHDSI.pdf> [Accessed 16 March 2021].

Wang, Shirley V., Olga V. Patterson, Joshua J. Gagne, Jeffrey S. Brown, Robert Ball, Pall Jonsson, Adam Wright, Li Zhou, Wim Goettsch, and Andrew Bate. 2019. 'Transparent Reporting on Research Using Unstructured Electronic Health Record Data to Generate "Real World" Evidence of Comparative Effectiveness and Safety.' Drug Safety 42 (11): 1297-1309.

Wang, Shirley V., Simone Pinheiro, Wei Hua, Peter Arlett, Yoshiaki Uyama, Jesse A. Berlin, Dorothee B. Bartels, Kristijan H. Kahler, Lily G. Bessette, and Sebastian Schneeweiss. 2021. 'STaRT-RWE: structured template for planning and reporting on the implementation of real world evidence studies.' BMJ 372.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al., 2016. 'The FAIR Guiding Principles for Scientific Data Management and Stewardship.' Scientific Data 3 (1): 160018.

7.2. Appendix 2: Technical description of MOLGENIS

MOLGENIS is a FAIR data platform for researchers to accelerate scientific collaborations that has been under continuous development since 2001. While originally developed for molecular genetics research (Swertz et al., 2004, Swertz & Jansen 2007), it is now a broadly applicable data tool thanks to many sponsors and contributors. The MOLGENIS community boasts being a front-runner in the FAIR movement, i.e., leading methodologies to make data findable, accessible, reusable, and interoperable (Wilkinson et al., 2016), and is recognised by ELIXIR as 'recommended interoperability resource.' MOLGENIS is used in many scientific areas such as biobanking, rare disease research, multicentre cohort studies, genomics, patient registries, and other scientific organisations all around the world. Currently, more than 100 servers and more than 40 projects use MOLGENIS as their platform, and more than 20 active projects sponsor the MOLGENIS core team, enabling 30+ engineers, data stewards, system administrators, and project leads to sustain the platform and its users. MOLGENIS software is free to download and use, and to modify as open source, under LGPLv3 licence. In addition, there are pay-for-service options if sustainable hosting is required.

Below provides a summary of MOLGENIS use cases, features, governance, and literature references.

7.2.1. Portfolio

MOLGENIS has been applied in many ways, and great data solutions have been created by projects of varying size and backgrounds. Typically, different stakeholders from multiple projects or organisations join hands around one application area, resulting in a family of MOLGENIS instances. Please see examples of currently active MOLGENIS user communities and their applications below.

7.2.1.1. FAIR catalogue

The MOLGENIS community has created a 'catalogue' application family that is now used as the basis by many health (research) data infrastructures in Europe. Motivation for use of this application is that many relevant data for health research are privacy sensitive and hence data cannot be put on the Internet. The MOLGENIS catalogue has enabled data providers to make large collections of data and samples findable and promotes request and access. MOLGENIS catalogue is now used by BBMRI-ERIC (Holub et al., 2016, Holub et al., 2020), BBMRI-NL, BBMRI-DE biobank catalogues, as well as LifeLines Biobank, and RD-connect rare disease samples catalogue, to name a few (e.g., https://directory.bbmri-eric.eu/).

7.2.1.2. Multicentre data federations

A natural extension of the aforementioned catalogue has been to catalogue not only individual data collections, but also multicentre collaborations such as LifeCycle (Jaddoe et al., 2020) that integrate data from multiple cohorts, 'harmonise' their contents onto a common standard, and subsequently make the harmonised data available for analysis. In particular, MOLGENIS has joined hands with the DataSHIELD community, an open-source software that enables data analysis across multiple institutes without needing to collect personal data centrally but instead by sharing only summary statistics. Such privacy-preserving federated analysis depends on high-quality data harmonisations and documentation thereof, which is now possible in MOLGENIS (e.g., http://catalogue.lifecycle-project.eu/).

7.2.1.3. Patient registries

Research into rare diseases and personalised medicine greatly depends on access to patient data. However, in many cases these data are hard to find if you want to look for rare phenotypes or genotypes. Therefore, the MOLGENIS community has developed a ‘registry’ application family. Applications in this family share patient, mutation, and disease knowledge to understand relations between genetics, environment, and disease (e.g., van den Akker et al., 2011). Most recently, four European Expert Reference Networks—Ithaca, SKIN, Genturis, and CRANIO—have chosen to use MOLGENIS for their central ‘meta’ registry. Example: https://deb-central.org/.

7.2.1.4. Central data warehouse for research groups/consortia

While the above projects are mainly about sharing metadata, many groups use MOLGENIS as their workhorse for data management within research groups or consortia. Typically, they create a bespoke data management plan and implement this using MOLGENIS as their platform. For example, the 1000IBD project (Imhann et al., 2019) aimed to prospectively follow more than 1,000 patients with inflammatory bowel disease from the northern provinces of the Netherlands. For these patients, they have collected a uniquely large number of phenotypes and generated multi-omics profiles (dietary and environmental factors, drug responses and adverse drug events, genotyping and sequencing, transcriptome information, and microbiome information). They make all this available using MOLGENIS, both internally for controlled access, as well as externally to aid data requests.

7.2.1.5. Genotype-phenotype analysis

Finally, the MOLGENIS community is strongly rooted in innovative research and healthcare using novel genomics methods. As a consequence, a large set of MOLGENIS applications centres around data management and analysis of (multi-)omics data. This has resulted in specialised data models and applications for genomics. Examples include large multicentre genomics studies, such as BBMRI BIOS and X-omics, as well as vast multicentre rare disease studies such as Solve-RD and European Joint Programme on Rare Diseases. MOLGENIS typically provides two features: (a) integrated metadata/ warehouse for the project (e.g., ’RD3’ in Solve-RD catalogues all samples and omics files in the project) and (b) a ‘digital research environment’ for large-scale genomics analysis, combining the MOLGENIS expertise on genomics data analysis pipelines with the capability to deliver high-performance computer ‘clouds’ (often called a ‘Sandbox’ for research).

7.2.2. Features

7.2.2.1. Rich model to define, capture, and manage your data

The core differentiation feature of MOLGENIS is the richness of its data structure options. MOLGENIS allows bioinformaticians/health information technology staff to rapidly customise MOLGENIS to the needs of its users (e.g., van der Velde et al., 2014). Because it is typically richer than its competitors, it is often used as the big ‘integrator,’ the system into which data from multiple sources is integrated. It enables uploading data in large batches or entering data via user-friendly forms. Dynamic refinement of the user’s data model is possible using the MOLGENIS advanced ‘object-relational’ data definition format and the online metadata editor.

7.2.2.2. FAIR interoperability interfaces

MOLGENIS has a long track record in making data findable, interoperable, accessible, and reusable (FAIR). For example, MOLGENIS is used as the FAIR catalogue software for the BBMRI, the European Biobanking and BioMolecular Research Infrastructure (http://directory.bbmri-eric.eu). In addition, MOLGENIS automatically generates technical and semantic interfaces from the data, such as industry standard REST web service interfaces using OpenAPI standard documentation, which enables programmers to connect systems where appropriate. MOLGENIS was the first to implement the new FAIR data point specification.

7.2.2.3. Secure access

MOLGENIS enables control of group, role, and individual access. MOLGENIS data are organised following scientific best practices. Data can be divided into research groups; within the groups, the user can assign roles such as 'data manager,' 'data editor,' and 'data user.' Authentication can be ensured by connecting the user's institute account via SURFconext (NL) or BBMRI/ELIXIR AAI (Europe) or by using Google two-factor authentication.

7.2.2.4. Scripting & visualisation

Bioinformaticians can take full control in MOLGENIS. It provides a rich script framework that allows users to program analyses and processing tools in their favourite programming languages (e.g., R, javascript, python) and connect to the data using APIs to add great analysis tools and views. Users can even create complete html + javascript apps. Example: http://molgenis.org/ase.

7.2.2.5. Harmonisation and integration

MOLGENIS provides specific tools to make data interoperable. Combined analysis is much more powerful than running smaller analyses on each data set separately, but data integration is hard. MOLGENIS offers multiple 'FAIRification' tools to find related data, codify data contents and transform different tables into one standardised table. These tools include SORTA (for recoding data), Mapping service (for ETL) and applications on top such as BiobankConnect (for cohort harmonisation).

7.2.2.6. Task automation

MOLGENIS has tools to automate recurring tasks such as data upload, transformation, and statistics. Frequently, data from multiple sources must be combined for success. Therefore, data exchanges, transformations, and analyses must be repeated often.

7.2.2.7. Extreme customisation

The portfolio shows MOLGENIS can support a wide range of use cases. That is only possible because the user can change MOLGENIS' complete data structure, menus, logic, and layout. MOLGENIS is the only software in its class that allows complete data model freedom. In addition, the user can completely change menu structure, user interface, and permission system. Finally, theme styles can be used to fit the user's website seamlessly.

7.2.2.8. App development platform

Still, when MOLGENIS configuration is still too limiting, users can add their own user interfaces to the MOLGENIS app store. MOLGENIS gives programmers the complete freedom to create HTML+JavaScript applications using MOLGENIS REST-style programmer interfaces. The user can upload these apps like plugins to become part of MOLGENIS itself and use them seamlessly. For example, Johansson et al. (2018) implemented a complete Non-invasive prenatal testing pipeline software this way.

7.2.2.9. High-performance computing and genomics 'cloud'

MOLGENIS enables use of large-scale analysis jobs on a computer cluster. MOLGENIS developers have used that capability for genomics research as well as in routine DNA diagnostics for > 10 years (e.g., Li et al., 2020). MOLGENIS also provides a high-performance computing framework called the 'Sandbox,' which uses simple spreadsheets to define workflows and templates to define workflow steps. It works on PBS and SLURM. EasyBuild is used for deployment and OpenStack is used to make it a true cloud facility.

7.2.2.10. Industry standard deployment

MOLGENIS is built using industry standards such as Java, PostgreSQL, Elasticsearch, Minio, VueJS, R project, Python. In addition, we deliver prepackaged deployments into Docker, Kubernetes, Redhat and using Ansible. Finally, we provide hosting as a service (i.e., 'MOLGENIS cloud') under the contractual framework of UMCG (University Medical Centre Groningen) research information technology support facility, including security and GDPR (General Data Protection Regulation) compliance.

7.2.3. Literature references

Selected publications:

Arends D, van der Velde KJ, Prins P, et al. xQTL workbench: a scalable web environment for multi-level QTL analysis. Bioinformatics. 2012 Apr 1;28(7):1042-4. doi: 10.1093/bioinformatics/bts049. PMID: 22308096.

Fokkema IFAC, van der Velde KJ, Slofstra MK, et al. Dutch genome diagnostic laboratories accelerated and improved variant interpretation and increased accuracy by sharing data. Hum Mutat. 2019 Dec;40(12):2230-2238. doi: 10.1002/humu.23896. PMID: 31433103.

Holub P, Kozera L, Florindi F, et al.; BBMRI-ERIC community. BBMRI-ERIC's contributions to research and knowledge exchange on COVID-19. Eur J Hum Genet. 2020 Jun;28(6):728-731. doi: 10.1038/s41431-020-0634-8. PMID: 32444797.

Holub P, Swertz M, Reihs R, et al. BBMRI-ERIC Directory: 515 Biobanks with over 60 million biological samples. Biopreserv Biobank. 2016 Dec;14(6):559-562. doi: 10.1089/bio.2016.0088. PMID: 27936342.

Imhann F, Van der Velde KJ, Barbieri R, et al. The 1000IBD project: multi-omics data of 1000 inflammatory bowel disease patients; data release 1. BMC Gastroenterol. 2019 Jan 8;19(1):5. doi: 10.1186/s12876-018-0917-5. PMID: 30621600.

Jaddoe VWV, Felix JF, Andersen AN, et al.; LifeCycle Project Group. The LifeCycle Project-EU Child Cohort Network: a federated analysis infrastructure and harmonized data of more than 250,000 children and parents. Eur J Epidemiol. 2020 Jul;35(7):709-724. doi: 10.1007/s10654-020-00662-z.

Janssen N, Bergman JE, Swertz MA, et al. Mutation update on the CHD7 gene involved in CHARGE syndrome. Hum Mutat. 2012 Aug;33(8):1149-60. doi: 10.1002/humu.22086. PMID: 22461308.

Johansson LF, de Weerd HA, de Boer EN, et al. NIPTeR: an R package for fast and accurate trisomy prediction in non-invasive prenatal testing. BMC Bioinformatics. 2018 Dec 17;19(1):531. doi: 10.1186/s12859-018-2557-8. PMID: 30558531.

Krops LA, Bouma AJ, Van Nassau F, et al. Implementing individually tailored prescription of physical activity in routine clinical care: protocol of the Physicians Implement Exercise = Medicine (PIE=M) Development and Implementation Project. JMIR Res Protoc. 2020 Nov 2;9(11):e19397. doi: 10.2196/19397.

Lazzarini E, Jongbloed JD, Pilichou K, et al. The ARVD/C genetic variants database: 2014 update. Hum Mutat. 2015 Apr;36(4):403-10. doi: 10.1002/humu.22765. PMID: 25676813.

Li S, van der Velde KJ, de Ridder D, et al. CAPICE: a computational method for Consequence-Agnostic Pathogenicity Interpretation of Clinical Exome variations. Genome Med. 2020 Aug 24;12(1):75. doi: 10.1186/s13073-020-00775-w. PMID: 32831124.

Merino-Martinez R, Norlin L, van Enckevort D, et al. Toward global biobank integration by implementation of the Minimum Information About BIobank Data Sharing (MIABIS 2.0 Core). Biopreserv Biobank. 2016 Aug;14(4):298-306. doi: 10.1089/bio.2015.0070. PMID: 26977825.

Netea MG, Joosten LA, Li Y, et al. Understanding human immune function using the resources from the Human Functional Genomics Project. Nat Med. 2016 Aug 4;22(8):831-3. doi: 10.1038/nm.4140. PMID: 27490433.

Pang C, Hendriksen D, Dijkstra M, et al. BiobankConnect: software to rapidly connect data elements for pooled analysis across biobanks using ontological and lexical indexing. J Am Med Inform Assoc. 2015 Jan;22(1):65-75. doi: 10.1136/amiajnl-2013-002577. PMID: 25361575.

Pang C, Kelpin F, van Enckevort D, et al. BiobankUniverse: automatic matchmaking between datasets for biobank data discovery and integration. Bioinformatics. 2017 Nov 15;33(22):3627-3634. doi: 10.1093/bioinformatics/btx478. PMID: 29036577.

Pang C, Sollie A, Sijtsma A, et al. SORTA: a system for ontology-based re-coding and technical annotation of biomedical phenotype data. Database (Oxford). 2015 Sep 18;2015:bav089. doi: 10.1093/database/bav089. PMID: 26385205.

Pang C, van Enckevort D, de Haan M, et al. MOLGENIS/connect: a system for semi-automatic integration of heterogeneous phenotype data with applications in biobanks. Bioinformatics. 2016 Jul 15;32(14):2176-83. doi: 10.1093/bioinformatics/btw155. PMID: 27153686.

Scholtens S, Smidt N, Swertz MA, et al. Cohort profile: LifeLines, a three-generation cohort study and biobank. Int J Epidemiol. 2015 Aug;44(4):1172-80. doi: 10.1093/ije/dyu229. PMID: 25502107.

Snoek LB, Joeri van der Velde K, Li Y, et al. Worm variation made accessible: take your shopping cart to store, link, and investigate! Worm. 2014 Jan 1;3(1):e28357. doi: 10.4161/worm.28357. PMID: 24843834.

Snoek LB, Van der Velde KJ, Arends D, et al. WormQTL--public archive and analysis web portal for natural variation data in Caenorhabditis spp. Nucleic Acids Res. 2013 Jan;41(Database issue):D738-43. doi: 10.1093/nar/gks1124. PMID: 23180786.

Swertz MA, De Brock EO, Van Hijum SA, et al. Molecular Genetics Information System (MOLGENIS): alternatives in developing local experimental genomics databases. Bioinformatics. 2004 Sep 1;20(13):2075-83. doi: 10.1093/bioinformatics/bth206. PMID: 15059831.

Swertz MA, Dijkstra M, Adamusiak T, et al. The MOLGENIS toolkit: rapid prototyping of biosoftware at the push of a button. BMC Bioinformatics. 2010 Dec 21;11(Suppl 12):S12. doi: 10.1186/1471-2105-11-S12-S12. PMID: 21210979.

Swertz MA, Jansen RC. Beyond standardization: dynamic software infrastructures for systems biology. Nat Rev Genet. 2007 Mar;8(3):235-43. doi: 10.1038/nrg2048. PMID: 17297480.

van den Akker PC, Jonkman MF, Rengaw T, et al. The international dystrophic epidermolysis bullosa patient registry: an online database of dystrophic epidermolysis bullosa patients and their COL7A1 mutations. Hum Mutat. 2011 Oct;32(10):1100-7. doi: 10.1002/humu.21551. PMID: 21681854.

van der Velde KJ, de Haan M, Zych K, et al. WormQTLHD--a web database for linking human disease to natural variation data in C. elegans. Nucleic Acids Res. 2014 Jan;42(Database issue):D794-801. doi: 10.1093/nar/gkt1044. PMID: 24217915.

van der Velde KJ, Dhekne HS, Swertz MA, et al. An overview and online registry of microvillus inclusion disease patients and their MYO5B mutations. Hum Mutat. 2013 Dec;34(12):1597-605. doi: 10.1002/humu.22440. PMID: 24014347.

van der Velde KJ, Imhann F, Charbon B, et al. MOLGENIS research: advanced bioinformatics data software for non-bioinformaticians. Bioinformatics. 2019 Mar 15;35(6):1076-1078. doi: 10.1093/bioinformatics/bty742. PMID: 30165396.

Van Gijn ME, Ceccherini I, Shinar Y, et al. New workflow for classification of genetic variants’ pathogenicity applied to hereditary recurrent fevers by the International Study Group for Systemic Autoinflammatory Diseases (INSAID). J Med Genet. 2018 Aug;55(8):530-537. doi: 10.1136/jmedgenet-2017-105216. PMID: 29599418.

Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016 Mar 15;3:160018. doi: 10.1038/sdata.2016.18. PMID: 26978244.