MINERVA: Metadata for data dIscoverability aNd study rEplicability in obseRVAtional studies

Strengthening Use of Real-World Data in Medicines Development: Metadata for Data Discoverability and Study Replicability (EMA/2017/09/PE/16)

MINERVA Deliverable 9: Final Good Practice Guide for Metadata Collection for Real-World Data Sources

A Partnership of the EU PE&PV Research Network, the SIGMA Consortium, and Collaborators

EUPAS39322

10 January 2022

Prepared for:

European Medicines Agency

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

Task 9 Leads and Co-Leads

ARS, Rosa Gini
UU, Romin Pajouheshnia
UMCG, Morris Swertz, Eleanor Hyde
UMCU, Miriam Sturkenboom
RTI-HS, Lia Gutierrez, Susana Perez-Gutthann
On behalf of the MINERVA project consortium
Contact: sperez@rti.org

Disclaimer

This document is a deliverable of the MINERVA project funded by the European Medicines Agency through the framework contract No EMA/2017/09/PE/16. The views expressed in this document are the personal views of the authors and may not be understood or quoted as being made on behalf of or reflecting the position of the European Medicines Agency or one of its committees or working parties.

Document history

Version	Date	Authors	Changes
Draft 0.1	27 Sep 2021	Rosa Gini	First draft for Consortium review
Draft 0.2	6 Oct 2021	Consortium Review	Comments including changes in sequence and sustainability recommendations
Draft 0.3	18 Oct 2021	Rosa Gini	Comments incorporated
Draft 0.4	20 Oct 2021	Consortium Review	
Draft 1.0	22 Oct 2021	Rosa Gini, Romin Pajouheshnia, Morris Swertz, Eleanor Hyde, Lia Gutierrez, Susana Perez-Gutthann, and Miriam Sturkenboom, on behalf of the MINERVA Consortium	Document finalisation
Draft for Version 2.0	3 Dec 2021	Consortium Review	Draft addressing EMA comments
Draft 2.0	6 Dec 2021	Rosa Gini, Romin Pajouheshnia, Morris Swertz, Eleanor Hyde, Lia Gutierrez, Susana Perez-Gutthann, and Miriam Sturkenboom, on behalf of the MINERVA Consortium	Draft addressing EMA comments and Consortium
Draft 2.1	14 Dec 2021	Rosa Gini, Romin Pajouheshnia, Morris Swertz, Eleanor Hyde, Lia Gutierrez, Susana Perez-Gutthann, and Miriam Sturkenboom, on behalf of the MINERVA Consortium	Draft addressing EMA comments of 13 Dec 2021
Draft for Final document	28 Dec 2021	Consortium Review	Draft addressing EMA comments
Final Document	10 Jan 2022	Rosa Gini	Incorporated comments from Consortium

MINERVA Consortium	Key Roles Organisation	Organisation, Country
Susana Perez-Gutthann, Lia Gutierrez, Carla Franzoni, Andrea Margulis, 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, Carlos Duran, Vjola Hoxhaj	**Lead Task 7** *President VAC4EU* *PI ConcePTION*	UMCU, Netherlands
Romin Pajouheshnia, Olaf Klungel, Satu Johanna Siiskonen, 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, Eric Zwart, Marije van der Geest, Brenda Hijmans, Sido Haakma, Connor Stroomberg	**Lead Task 8, Co-lead Task 4**	UMCG, Netherlands
Cécile Droz-Perroteau, Nicolas Thurin, Régis Lassalle, Séverine Lignot	**DAP, SNDS, France**	BPE, University of Bordeaux, France
Vera Ehrenstein	**DAP, Denmark**	AU-DCE, Aarhus, Denmark
Ron Herings, Josine Kuiper, Ella Jansen	**DAP, PHARMO; Netherlands** *Co-lead SIGMA office*	PHARMO, Netherlands
Ulrike Haug, Wiebke Schäfer, 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 García 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, Janja Jazbar, Špela Žerovnik	**DAP, Slovenia**	UL FFA, Ljubljana, Slovenia
Manuel Barreiro, Anna Casas, Eugeni Domènech, Yamile Zabana	**DAP, ENEIDA patient registry, Spain**	GETECCU, Spain
Bas Middelkoop, Eoin McGrath, Silvia Zaccagnino, Maria Paula Busto, Andreu Gusi	**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 partner; 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 **7**

1 Background **10**

2 Guidance **10**

2.1 FAIR principles 10

2.2 Conceptual framework for describing data sources 11

2.3 Guidance on the structure of the Catalogue 12

2.4 Guidance on sustainable metadata collection into the Catalogue 13

2.5 Recommendations based on use cases for testing the Catalogue 15

3 Considerations detailing and supporting the recommendations **18**

3.1 Institution, Data Source, and Data Bank sections: qualitative metadata, initial population 18

3.2 Common Data Model and Network sections: initial population 18

3.3 Study-independent and study-specific updates of the qualitative metadata of all sections 18

3.4 Quantitative metadata in the Data Source and Data Bank sections 19

3.5 List of metadata 20

3.6 Functionalities to support data entry and discoverability 21

3.7 Interoperability sof catalogues 21

3.8 Quality checks 22

3.9 Required efforts for implementing and maintaining a sustainable catalogue 22

3.10 Applied examples of use 24

4 Conclusions **29**

5 References **30**

Annex 1: MINERVA project work substantiating the recommendations **32**

Annex 2: Final metadata list **38**

Abbreviations

AAI	authentication and authorisation infrastructure
ARS	Agenzia Regionale di Sanità della Toscana (Italy)
AU-DCE	Aarhus University, Department of Clinical Epidemiology (Denmark)
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)
CDM	common data model
COVID-19	coronavirus disease 2019
CPE KI	Centre for Pharmacoepidemiology, Karolinska Institutet
CPRD	Clinical Practice Research Datalink
DAP	data access partner
EBMT	The European Society for Blood & Marrow Transplantation
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 (inflammatory bowel disease registry in Spain)
EpiChron	a data source for studying chronic diseases (Spain)
ETL	extract, transform, load
EU	European Union
EU PAS Register	European Union electronic register of post-authorisation studies
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)
GDPR	General Data Protection Regulation 2016/679 (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
HMA	Heads of Medicines Agencies
IACS	Instituto Aragonés de Ciencias de la Salud (Spain)
ID	identification
IMI	Innovative Medicines Initiative
LSHTM	London School of Hygiene and Tropical Medicine
MINERVA	Metadata for data dIscoverability aNd study rEplicability in obseRVAtional studies
MOLGENIS	the software platform that will be used for the metadata Catalogue
OMOP	Observational Medical Outcomes Partnership
PHARMO	PHARMO Institute for Drug Outcomes Research (Netherlands)
PID	persistent identifier
POC	proof of concept
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 subtables: *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 partner (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) and that may contribute expertise on the data bank(s).
- Data bank: Data collections sustained by a specified organisation, which is the data originator. The data bank is characterised 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.
- 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 (or, joint controllers): The organisation(s) that determine(s) the purposes for which and the means by which personal data are processed. The qualification as joint controllers may arise when more than one actor is involved in the data controlling function (GDPR).
- 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 (GDPR).
- Data source: All the data banks referring to the same or partially overlapping underlying (sometimes called source) 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 Clinical Practice Research Datalink (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.

- Extract, transform, load (ETL): A repeatable process for converting data from one format to another, such as from a source format to a common data model format. 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 [PIDs], 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 is a population defined by an administrative characteristic, a geographical location, a disease, a medical condition or any other relevant characteristic.
- Unit of observation of a table: the item that is observed in a table. For example, in a table listing name, address, expertise of institutions, the unit of observation is the institution. In a table listing the role, data sources, publications, that institutions contribute to studies, the unit of observation is the study *and* the institution.
- 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

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" (MINERVA; EU PAS Register number EUPAS39322) was initiated. The main focus of the study was the definition of the set of metadata including engagement with stakeholders to reach broad agreement, and the development of a good practice guide describing the metadata and recommendations based on this pilot. The vision of the project was also to pilot a metadata list and a FAIR data proof-of-concept (POC) Catalogue designed to assist investigators, data access partners, and evidence consumers in the process that goes from discovery and feasibility to study design and conduct, and to document the strength and limitations of the study results.

This document represents the good practice guide on metadata for data discoverability and study replicability and contributes to EMA's vision on enabling the use of real-world evidence (RWE) across the spectrum of regulatory use cases (Arlett et al., 2021). Section 2 of this guide describes the conceptual framework, the structure of the catalogue, the sustainable collection of metadata in the catalogue, and recommendations for use cases. Section 3 covers the considerations supporting these recommendations. Annex 1 provides more details on the MINERVA pilot, on which the recommendations are based, and Annex 2 contains the final list of metadata.

2 Guidance

The goal of the good practice guide is to provide a description of the defined metadata and recommendations on use of metadata for the purpose of identifying real-world data sources for a specific study purpose, assessing the suitability of data sources used in previous studies, and contributing to assessment of the evidentiary value of study results. The good practice guide provides regulatory use cases and examples of the use of the metadata upon consultation with the EMA. This guide incorporates the experience of the MINERVA pilot in populating the list of metadata and the POC catalogue.

Our guidance recommendations are based on developmental and testing work that is described in separate MINERVA deliverables, which are summarised in Annex 1, and on the FAIR principles and a conceptual framework, which are summarised below.

2.1 FAIR principles

The FAIR principles should be applicable to the development of the metadata catalogue. The MINERVA POC catalogue was developed based on these principles. In 2016, the "FAIR Guiding Principles for scientific data management and stewardship" were published (Wilkinson, et al., 2016). The authors intended to provide guidelines to improve the Findability, Accessibility, Interoperability, and Reuse (FAIR) of digital assets. The principles emphasise machine-actionability (i.e., the capacity of computational systems to find, access, interoperate, and re-use data with no or minimal human intervention) because humans increasingly rely on computational support to deal with data as a result of the increase in volume, complexity, and creation speed of data.

2.2 Conceptual framework for describing data sources

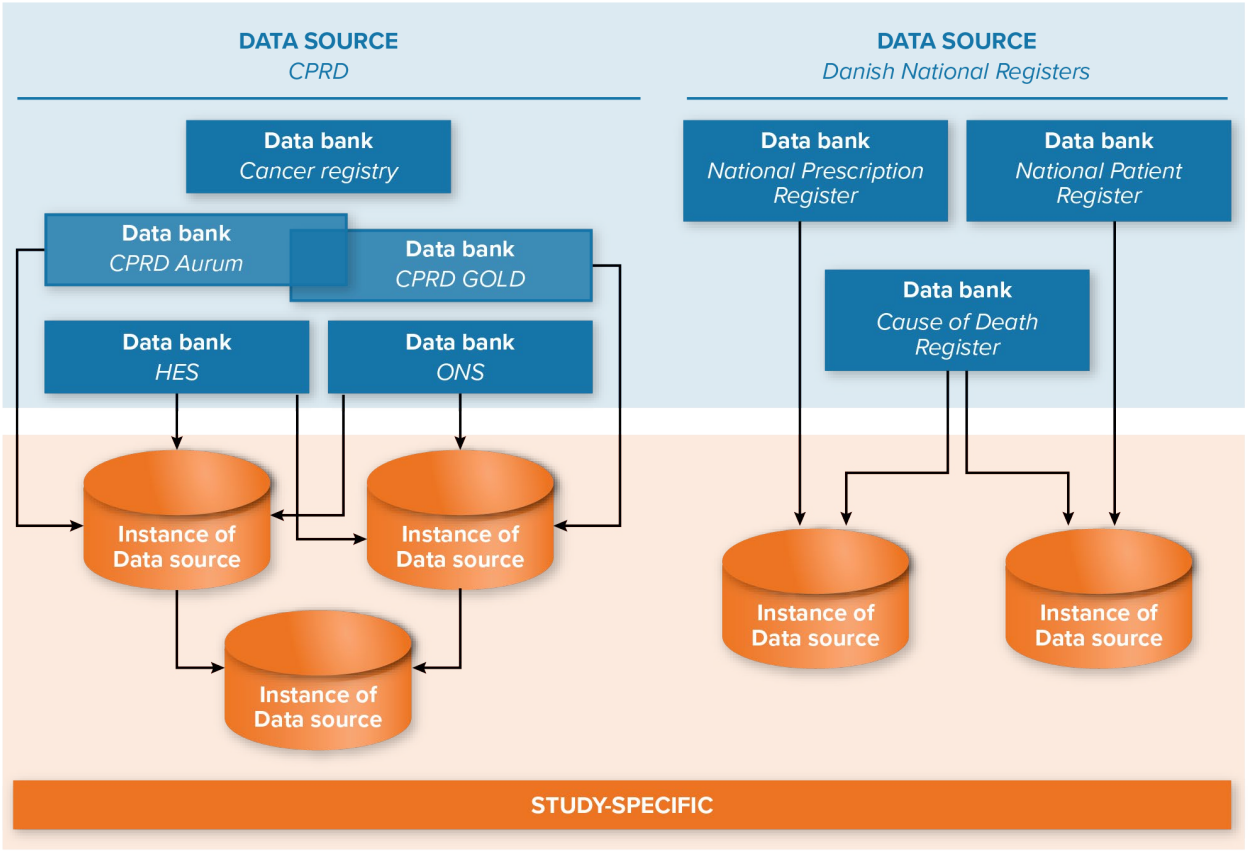
The conceptual framework of the MINERVA POC catalogue pilot project stems from acknowledging that the mechanisms that put data into existence are heterogeneous across data sources and even within the specific data banks of a data source. This variability needs to be clearly described to permit data discoverability. The conceptual framework arose from a qualitative study conducted prior to the MINERVA project (Thurin et al., 2021) and is summarised in the next paragraphs.

Data sources are composed of ***data banks***, which are collections of data sustained by a specified organisation (***originator of the data bank***), 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, for a purpose that may be unrelated to research (e.g., reimbursement). Each data bank comprises tables, variables, values, and labels that can be further annotated with the following information for better interpretation:

- ***Prompts****:* the situations/context that prompted the creation of a record in the data bank, e.g., access to an emergency room; a visit to a primary care centre; the dispensing of a reimbursable medication. In the specific case that data bank A automatically collects information originally recorded in data bank B, then data bank A will have among its own prompts the indication "A record in data bank A is also created whenever it is created in data bank B"
- ***Underlying population****:* the population whose events would prompt the creation of a record in the data bank if the prompt happened, e.g., the persons entitled to receive healthcare assistance funded by a specific payer; persons in a geographical area assigned to a specific network of primary care centres; persons with a disease, a medical condition; legal inhabitants of a region or country
- ***Data model:*** overview of the tables and variables in the data format for the data access provider, including the data dictionary describing the native variables, labels, and format of the variables
- ***Rules*** to convert the format of the tables of the data access partner (DAP) of the data bank to a common data model (CDM) that is used by multiple DAPs

A data source is composed of data banks having overlapping underlying populations and linkable to one another at an individual level (either probabilistic or deterministic). A data source could be a single data bank or any combination of data banks that relate to the same or partially overlapping underlying populations. The relationship between data sources and data banks is illustrated in Figure 1.

Figure 1. Illustration of the conceptual framework underlying the catalogue design



CPRD = Clinical Practice Research Datalink; GOLD = general practitioner online data source of the CPRD; HES = Hospital Episode Statistics; ONS = Office for National Statistics (cause-of death data).
Note: instances of data sources might include any combination of available data banks in the data source

A data access partner (***DAP***) is an organisation that is capable of obtaining access to relevant data. DAPs have research experience and capabilities in RWE research supporting regulatory decisions. Access may be complete or partial and may be subject to authorisation for use; for example, based on a study protocol, an ethical approval, or other local regulations. From the point of view of the General Data Protection Regulation 2016/679 (GDPR), the DAP may have any of the following roles: data controller, joint controller, or data processor.

2.3 Guidance on the structure of the Catalogue

We recommend that the future Catalogue of metadata comprises at least the following 6 sections:

- **Institutions.** This section should provide information on institutions of relevance for data discoverability, including those responsible for entries to the Catalogue, and others that are related to the data banks (e.g., data originators, data controllers or processors). Each institution is listed with its roles and expertise.
- **Data Source.** This section should describe the data source: (a) qualitative information about the data source, including information on possibilities and routes for access to the data source or instances thereof by potential DAPs for research purposes, and (b) quantitative metadata, e.g., counts and demographic distributions of the underlying population in the data source. If the data banks of a data source have only partially overlapping underlying populations, the quantitative metadata will be adapted to the configuration, e.g., all the relevant subpopulations will be described.

- **Data Bank.** This section should describe individual data banks in the data source with respect to (a) qualitative information—descriptors of the underlying population (including metadata that allow for linking with other data banks that have a partially or fully overlapping underlying population); possibilities and routes for access to the data bank thereof by potential DAPs for research purposes; what prompts records to be created; the data source(s) in the Data Source section that includes the data bank; the content, including the data model of the tables in the data bank; and the design of their extract, transform, and load (ETL) procedures to common data models (if any)—and (b) quantitative metadata, e.g., counts and demographic distributions of the underlying population of the data bank and/or completeness of the cells in the data bank tables.
- **Study.** This section should give access to descriptions of studies using data sources or/and data banks described in the Catalogue. This section should cross-link and be interoperable or merged with the EU PAS Register/Study metadata catalogue to link information on studies with information on the use, content, and quality of data sources and their respective data banks. This section should record strengths and limitations of each data source, specifically referring to the study questions addressed in the study.
- **Common Data Models (CDMs).** This section should describe information on CDMs to which data sources described in the Catalogue or their data banks have been converted.
- **Network.** The Network section should provide descriptions of research networks that include institutions described within the Institutions section.

A preliminary list of the metadata elements included in the six sections was listed and described in Deliverable 5—which is available on the EU PAS Register (http://www.encepp.eu/encepp/openAttachment/documentsLatest.otherDocument-0/43021)—and has been further updated in this document.

Many questions and information items are not standardised in other metadata schema or ontologies. Pharmacoepidemiologic expert knowledge was used to define such items, based on metadata items from existing catalogues (ENCePP, GRiP, ADVANCE, ConcePTION, EMIF, MOLGENIS) and interviews with representatives from large networks deploying real-world data for regulatory decision making.

2.4 Guidance on sustainable metadata collection into the Catalogue

Available metadata catalogues often have poor sustainability. They were created by several organisations on a project basis and consistently suffer from lack of updating, inaccessibility, and lack of interoperability; hardly any persistent identifiers (PIDs) are available (EMA, 2021a). The future Catalogue should be sustainable, meaning that the Catalogue is accessible and interoperable with other FAIR metadata catalogues, that the metadata content is findable and reusable, and that there are financial resources for this work. An organisation will need to be tasked with the role of controller of the Catalogue, hereinafter referred to as "Catalogue controller." For it to be reusable, it should be kept updated and be compliant with the GDPR. It also requires that there is a common agreement that it is justifiable and acceptable to collect the information in the Catalogue, either directly from organisations that access the data banks or through linking to publicly available information from the data banks, including other FAIR metadata catalogues.

We recommend the following practices:

1. FAIR and open-science principles are followed
    a. The Catalogue should be accessible to the public.
    b. When available, open-source software should be used for the Catalogue.
    c. Existing PIDs for digital objects should be used (e.g., ORCID, for persons).

d. For the key concepts of data sources and data banks, the Catalogue controller should take the lead in creating PIDs and engaging with the wider community of European metadata catalogue holders, encouraging them to adopt PIDs to facilitate interoperability. This would then facilitate exchange of metadata for the same data sources and data banks across metadata catalogues.

e. Global ontologies should be used.

2. Metadata completion is supported

   a. The Catalogue controller should install a service office/desk for the Catalogue.

   b. First-time collection of study-independent core metadata directly from a DAP for the purpose of populating the metadata catalogue should be conducted online together with a catalogue/metadata expert. Refer to Annex 1, Table 1-1, Challenge C.

   c. Procedures should be put in place to build a trusted relationship between the DAP and the Catalogue controller, namely a relationship with clear a priori set-up and mutually beneficial goals with transparent communication in which mutual feedback is acknowledged and acted upon and both parties' interests are respected. The result and proof of a trusted relationship is that the DAP continues to be engaged.

   d. Update of qualitative metadata in the Data Source and Data Bank sections should be based on updates inputted in the Study section during studies. Thanks to interoperability, it should be possible to update the Study section from the EU PAS Register, and in regulatory-requested studies, study documentation should include updating this section of the Catalogue. Through interoperability, it should be possible to also update the study-independent metadata, if the DAP requests so.

   e. To minimise double-entry work, support should be technically enacted to interoperate metadata within the Catalogue—from the Study section to the Data Source and Data Bank sections, and vice versa—and to obtain publicly available information from the data banks provided by an organisation with experience with the data banks.

3. Change management and reproducibility are enacted

   a. Constantly search and align with global alliances and initiatives for standards in labelling/ annotating metadata and update the metadata to keep it up to date and machine readable; in particular, align with PIDs and global ontologies. In the absence of ontologies, free-text entries should be used.

   b. Those entering and editing metadata information should use a robust authentication system—e.g., Elixir AAI: Authentication and Authorisation Infrastructure (https://elixir-europe.org/services/compute/aai)—to allow for an audit trail. Refer to Annex 1, Table 1-1, Challenge A.

   c. All DAPs of the same data source/data bank, including data originators that become DAPs should be enabled to edit the corresponding metadata, but ensuring that the attribution of each data entry is traceable via appropriate version control. An amendment/versioning overview for the Data Source and Data Bank sections should be available. Refer to Annex 1, Table 1-1, Challenge B.

   d. For reproducibility, for each entry in the Study section, a copy of the metadata for data sources and data banks should be created at the beginning of the study and edited by the DAPs participating in the study, under their own responsibility. Such copy may be edited by

the DAPs throughout the study, as needed to reflect study-tailored metadata, but left unchanged after the end of the study.

e. Procedures to compute quantitative metadata should support multiple common data models.

4. Epidemiologic principles are complied with

   a. Have a study protocol that allows for repeated recalculation of quantitative metadata, signed off by at least one DAP per data source/data bank.

   b. Compute quantitative metadata for data sources and for data banks at the level of the underlying population.

5. GDPR is complied with

   a. Define and communicate the purpose and use of the Catalogue, as well as the duration of the availability of metadata.

   b. Include in the metadata entry process the active approval for metadata to be re-used, re-edited, and published under an appropriate licence.

   c. Ensure a technical option to delete metadata.

   d. Establish appropriate measures to protect the privacy of institutions and individuals.

6. Quality management system is in place

   a. A quality management system is in place by the Catalogue controller that has procedures for metadata collection and updating and for evaluating the services.

   b. An incident management system is in place.

   c. A disaster recovery plan is available.

   d. A quality-assurance officer is available.

2.5 Recommendations based on use cases for testing the Catalogue

No.	Use case	Recommendations
1	**A DAP provides expertise throughout the cycle of a study**	The data access partner (DAP) accesses the Catalogue when the organisation is involved in a study. The DAP first considers editing the Institution, Data Source, and Data Bank sections that are under the DAP's responsibility, if appropriate. The updated sections are then replicated in the Study section of the Catalogue, under a subsection of the new entry of the current study. In this study-specific subsection, the DAP records strengths and weaknesses of the data source that are particularly relevant for the study at hand and the algorithms to define specific study variables, including strengths and limitations of the algorithms themselves. The subsection may be updated again during the study if new expertise is obtained by the DAP on its data source, e.g., information gained via validation studies. At the interpretation stage, some lessons learnt on strengths and limitations of the data may also be incorporated in the Catalogue.

No.	Use case	Recommendations
		If some of the study-specific information is deemed relevant for the study-independent sections of the Catalogue, the DAP migrates such information as appropriate. After the end of the study, the study-specific subsection will remain frozen to ensure transparency and replicability.
2	**An investigator uses the Catalogue throughout the life cycle of a study**	The investigator accesses the Catalogue prior to initiation of a study. Two sections are of interest for discoverability: the Data Source and the Study sections. In the former, study-independent metadata on data sources are available. In the latter, the information on data sources is tailored to specific research questions that have been utilised in the past. The Study section of the Catalogue allows finding studies addressing research questions similar to the one at hand, details on how the data sources involved were used, and on their strengths and limitations. A link to the EU PAS Register with the protocols and reports of these studies is available. Quantitative quality measures may also be included in the Study section, as well as results, which may allow a preliminary power calculation. The Data Source section of the Catalogue allows follow-up on the latest updated information on relevant data sources, in particular, based on geographic coverage. For some data sources, quantitative metadata resulting from "repeated analyses" may be available in the future. If in a data source of interest such metadata include contents of interest (e.g., a condition that is an outcome of the study), then a preliminary power calculation is possible. The investigator can use the contact information to reach out to the DAP(s) who worked with the data source. Once the DAPs are involved, they can use the metadata Catalogue based on Use Case 1. At the design stage and when study results are interpreted, the investigator is supported by the information on strengths and limitations of the data sources included in the Catalogue by the DAPs.
3	**A programmer programs a study**	The programmer accesses the Catalogue. As in Use Case 1, the DAPs have suggested in the Study section how to operationalise the variables of the study in their respective data sources. The decision is stored in the statistical analysis plan of the study, but the metadata Catalogue contains technical details, such as the name of the tables and columns of the data banks in the data source. If the study is implemented in a common data model (CDM), the Data Source section may contain the specifications of the extract, transform, and load (ETL) procedure from the data source to the CDM. Irrespective of whether the DAP has converted to the CDM the entire data source, or only an extraction thereof, this information supports the programmer in developing the study script. Using the details on all the data sources participating in the study, the script can be tailored to each data source. If the script is available in a public repository, e.g., a GitHub repository, at the end of the study the programmer can record in the Catalogue the link to the repository, thus enabling transparency and quality control and facilitating reproducibility.

No.	Use case	Recommendations
4	**A consumer of evidence assesses reproducibility and quality of evidence generated by a study**	The consumer of evidence has become aware of a study that they want to understand more in detail. The publications associated with a study mention that the study is documented in the Catalogue. The user accesses the Catalogue in the Study section and finds the entry of the study. The user can explore detailed information on the strengths and limitations of the data sources involved in the study at the time when the study was conducted and obtain complete documentation of the implementation choices, possibly including the study script, if available. The user can find the study report and potentially access to study quality measures and results, if these are open.
5	**A FAIR process launched by an external organisation, including a data originator, populates the Catalogue**	Another organisation has implemented the Catalogue standards in its own findable, accessible, interoperable, and reusable (FAIR) catalogue, or a data originator has implemented interoperability standards on the documentation of a data model of its data bank(s). When an update is available, the information is automatically offered to the Catalogue for the Quality Office to consider.
6	**An institution with research capabilities becomes a DAP for a data bank and/or a data source**	The user can explore in the Data Bank section of the Catalogue the detailed description of each data bank, including its prompts, the data model and dictionary of its tables, and a study-independent description of its quality. The user can understand which constellations of data banks from the same underlying population have been used in the past as part of the same data source, if any. If such a data source exists, the user can explore in the Data Source section of the Catalogue its study-independent description. The user can then access the Study section of the Catalogue and read through the entries on the data source referring to past studies to gain an understanding of its strengths and limitations, possibly via past validations. If quantitative metadata have been designed and calculated on the data source, then they can be browsed in the Data Source section. The user interested in a data bank has the opportunity to find additional data banks on overlapping underlying population that have never been used before in studies as part of a same data source. The user can understand if the data source of interest has been converted already to a CDM. If so, they can access details of the ETL design and information about what data can be re-used.

Note: See Section 3.10, Applied examples of use.

3 Considerations detailing and supporting the recommendations

3.1 Institution, Data Source, and Data Bank sections: qualitative metadata, initial population

One of the challenges reported from the MINERVA pilot was that a deep understanding of the conceptual framework underpinning the metadata Catalogue is needed to correctly capture the DAPs' knowledge about the data source (refer to Annex 1, Table 1-1, Challenge C). To overcome this challenge, an iterative approach was successful: the DAP first shared the documents they used internally and filled out a simple template, the Catalogue expert reviewed them, a live interview allowed completion of the main part of the metadata while clarifying important details, and the DAP completed the metadata. Finally, a round of quality checks was done by the Catalogue experts of the MINERVA Consortium. The experts inspected the inputted information for the following: (1) incomplete or missing items, (2) possible misunderstanding of metadata items, and (3) typographical errors. The DAPs were notified of the errors observed and asked to revise their input, possibly with suggestions. This was done manually and was a pilot for the quality-assurance activity.

Although time-consuming, this process was key to ensuring good quality of the metadata.

3.2 Common Data Model and Network sections: initial population

The Common Data Model and Network sections are expected to have fewer entries. The number of metadata is small. As a consequence, the completion of such sections is expected to be easier to initiate and manage.

3.3 Study-independent and study-specific updates of the qualitative metadata of all sections

DAPs who had already populated compatible metadata in a previously compiled catalogue were not interviewed but were able to follow instructions to complete or update their metadata. Quality checks were needed to align some of the misunderstanding on metadata items, but overall, this pilot activity supported the recommendation to enable interoperability between existing FAIR metadata catalogues.

The EMA envisions that regular, study-independent updates of such qualitative metadata would be needed at specific times (e.g., annually); for example, when new data banks are included in the data source and when there are changes in governance and access to data sources and data banks. This would need to be supported through an appropriate funding mechanism.

Refinements of the qualitative metadata are foreseen on a study-by-study base, as part of the study activity, e.g., in a study-specific copy of the metadata of the data source, the strengths and limitations of the data source and specific variables to address a specific research question will be described and discussed based on the expertise of the DAP (irrespective of whether the DAP is a data controller, a data joint controller, or a data processor). A DAP participating in a study is responsible and scientifically accountable for interpretation of the study results based on the strength and limitations of the data source to which it has access. This pathway for update is therefore a powerful way of conveying into the Catalogue valuable knowledge and expertise. This type of discussion is feasible only in a study-specific entry since limitations are linked to a specific research question; however, the information can then be used in future occasions when a similar research question needs to be addressed by the same DAP, by another DAP, or by any investigator or consumer of evidence.

In order to pilot a study-specific update of the metadata, the example of two studies was retrospectively reproduced. In both studies, participating DAPs had already been interviewed based on

the conceptual framework underpinning the Catalogue and updated their interview to include new information about their data source (namely, inclusion of a COVID-19 registry as a novel data bank). Moreover, one of the studies requested an extraordinary timeliness in data access, and not all the data banks were available in time to be included. The requested information was shared during regular project activities and recorded in the Catalogue retrospectively. Had the Catalogue been available, it would have been recorded live, therefore supporting the study development. Entry of metadata into the Catalogue is available now to document the characteristics of the data sources at the time of the study.

3.4 Quantitative metadata in the Data Source and Data Bank sections

In most data sources included in the POC catalogue, for each individual, the date of entrance to and date of exit from the underlying population are recorded in one of the data banks (see definition of Underlying population in the Glossary). This characteristic is recorded in Section C1 of a data bank, and in Section B1 of the corresponding data source. In such cases, the underlying population can be seen as an epidemiological cohort, and the data banks allow investigators to derive study variables, which can be analysed and provide quantitative metadata. Easy examples include distributions of age and sex, all of which can be captured at the data source level in Section B6 of the catalogue and at the data bank level in Section C7 of the catalogue. Moreover, in many such cases, the underlying population can be considered a representative sample of a general population: for example, the underlying population may include all the inhabitants of a geographic area or a subpopulation selected based on characteristics that can be argued to act as a random sampling, such as registration with a selected sample of primary care physicians, which is a form of cluster sampling. Therefore, quantitative metadata should be computed whenever possible at the data source level. If the underlying populations of the data banks in a data source overlap only partially, the quantitative metadata can be computed in all the subpopulations deemed significant. Whenever the underlying population is a representative sample of a population, this creates the base for quantitative metadata to provide information on the population itself. Moreover, in this case, the quantitative metadata (for example, the prevalence of diabetes) are comparable across data sources. However, to support comparability, interpretation of the variables observed on the underlying population must take into account the limitations of the data source. As a simple example, consider data source A, where records of diagnoses are prompted solely by primary care visits; data source B, where records of diagnoses are prompted solely by inpatient visits; and data source C, where records of diagnoses are prompted by both inpatient and primary care visits. The comparison between the occurrence of the same diagnosis in data sources A, B, and C must consider this difference to ensure a correct interpretation of the metadata: in data source A it can be interpreted as occurrence of *access to primary care* for that condition; in data source B, as occurrence of *hospitalisation* for that condition; in data source C, as occurrence of *either hospitalisation or access to primary care* for that condition.

In our pilot, the technical challenge of creating a unique script able to run on multiple CDMs to produce quantitative metadata was successfully piloted for the variables describing the age and sex distributions in the underlying population.

Conversely, governance proved more challenging than expected (refer to Annex 1, Quantitative metadata of Data Source section, study-independent update, and Annex 1, Table 1-1, Challenge E). Even if the use of the data to compute the quantitative metadata had already been approved in the context of a different study, re-using the same data to produce the same numbers in a different way for a different project was perceived as "repurposing" the data and considered not authorised by 3 of 4 DAPs. This experience highlights that data use outside of a specific study is often at odds with local regulations with which DAPs must comply.

In the future, to update quantitative metadata, the Catalogue controller may seek to obtain data access through organisations that have more direct and timely access in terms of processing, including data originators. For an organisation to be recommended for this role, however, some conditions need to be fulfilled. First, to allow computation of many meaningful quantitative metadata, multiple data banks of the same data source must be accessed at the same time; for example, to be able to define the cohort of the underlying population. Second, the organisation must have research capabilities including formally trained and experienced epidemiologists, statisticians, and data scientists including clinically trained experts. Third, it must have expertise and capacity to identify the limitations of study variables that must be derived from the data banks to calculate quantitative metadata. One of the use cases of the Catalogue is to support organisations that aim to acquire this expertise to become DAPs.

A recommended solution to populate quantitative metadata in a regular fashion is framing this activity as "repeated analysis," as formalised by the DARWIN-EU classification of observational studies and analyses (EMA, 2021b): once the protocol to capture metadata from a data source has been defined and its limitations have been assessed, then the DAP should have authorisation to ensure regular repetition if this is sustainable.

3.5 List of metadata

Based on the pilot, the following updates were included in the core metadata list:

- It was clarified that the study-specific metadata should incorporate a copy of metadata for each data source and corresponding data banks: Section B is transported to the new Table F8, and Section C is transported to the new Section F9 (see Annex 2). Those metadata are expected to be populated at the beginning of the study, and then further edited, along with the rest of Section F. If the DAP believes that some updates to the metadata in Sections F8 and F9 can be transported back to sections B and C, this should be allowed by the rules on internal interoperability between sections (or between the Catalogue and the EU PAS Register). This new set of metadata enables Recommendation 2d.

  Discussions with the EMA also highlighted that the building of a new Study metadata catalogue (replacing the EU PAS Register) will allow linking the Data Source catalogue to the Study metadata catalogue. This technical solution would allow researchers to introduce study data, together with the databank identification number, only in the Study metadata catalogue. An automated process would populate the Data Source metadata catalogue with key study information to the record of the corresponding data bank.

- In Table F3, additional fields F3.12 were included, aimed to capture expertise and lessons learnt specific to the study question at hand, specifically for study variables (outcomes, exposure, covariates).

- The unit of observation of some metadata was revised: in the final version, the unit of observation is recorded in the tab "Full metadata list," column "Unit of observation."

- It is recommended that the data bank model (metadata C6.1.1) be considered as a priority variable, as it contains a complete representation of the contents of a data bank. In the current MINERVA POC Catalogue, this information is implemented as a machine-readable table, which is recommended according to the FAIR principles. The ETL specification to a common data model is also implemented as a machine-readable table in the current MINERVA POC Catalogue. During the pilot, in the absence of a consolidated ontology, it was apparent that some metadata that had been envisioned to be categorical could not be captured in the form provided by the DAPs. Such metadata were recast and entered as free text. In the case of metadata with multiple choice, a category "other" was added. For a future Catalogue, the recommendation is to provide the option for multiple selection.

The updated, final metadata list is available in Annex 2.

3.6 Functionalities to support data entry and discoverability

To facilitate data entry, the following functionalities would be useful:

- Some metadata are replicated across all the data banks of a data source. A useful functionality would allow the contributor, once a metadata item has been entered about a data bank, to replicate it across all the data banks of the same data source and to manually edit only metadata about those data banks that differ from the others (which may be the exception rather than the rule).
- Some metadata of a data source are just the summary of the corresponding metadata of its composing data banks. For example, the 'Areas of Information' of a data source are the unique list of all the 'Areas of Information' included in the data banks. A functionality could be developed that offers to compute the data source entry, based on the corresponding data bank entries.
- Some metadata are equal across different sections; they are repeated in order to be displayed in multiple places but should be collected only once. For example, the "linkageVariable" of a data bank is collected both in the Data Bank section and in the Data Source sections including that data bank.
- In the future, the Catalogue should be programmed to distinguish between true non-applicable metadata and metadata fields with missing values, which need to be completed.

On the other hand, discoverability through the data banks of a data source should be facilitated, by enabling searches, for example,

- "Is there any data bank in this data source that has *these metadata* equal to/matching *this value*?"
- "What is the list of values of *these metadata* in the data banks in this data source?"
- "Do all the data banks in this data source have *these metadata* equal to/matching *this value*?"

3.7 Interoperability of catalogues

Several initiatives coexist in Europe that involve a structural collection of metadata on data sources including the EU PAS Register (ENCePP, 2020) and the ENCePP Resources Database (ENCePP, 2021). The MINERVA metadata POC catalogue tested the interoperability between such catalogues in a test with the IMI-ConcePTION catalogue.

The MINERVA pilot started the interaction with other catalogue controllers/processors, which showed that an a priori semantic mapping between compatible structures is needed. For the future, systematic interoperability with other catalogues might reduce metadata entry frequency, increase quality, foster perceived value, and increase sustainability. Given the lack of ontologies and standards at this moment, exchange is technically possible but cannot be automated; therefore, it will need to be manually curated. This fact originated the recommendation to constantly search initiatives for standards and to update the Catalogue to keep it up to date and machine readable. In particular, align with PIDs and global ontologies. However, for the key concepts of data sources and data banks, the future Catalogue would be in an ideal position to take the lead.

To this aim, EMA could identify an organisation to be tasked with the role of providing PIDs, and a PID is created whenever a new entry of the Data Source and Data Bank section is created. The PID is then shared with other institutions that maintain resources willing to be interoperable with the Catalogue and linked to the corresponding data sources and/or data banks. This will then ensure that complementary descriptions of the same data source and data bank can be identified and potentially merged. To enable catalogues a pathway to provide a PID and metadata list, we recommend that

catalogues do not need to implement a complete metadata list but also can comply with a selected subset of metadata items.

The responsibility of identifying a data source/data bank with the entry in the Catalogue would stay with the organisation that maintains the external catalogue. In the case of the ENCePP Resources Database and of the EU PAS Register, the organisation is EMA itself. In the case of the VAC4EU Catalogue, the organisation is VAC4EU (Vaccine monitoring Collaboration for Europe).

Moreover, the metadata catalogues willing to be interoperable with the Catalogue need to agree with the Catalogue controller on a mapping between their resource and the metadata list (or its operationalisation as data model of the Catalogue). Once the two elements (PIDs and mapping between catalogues) are in place, interoperability can be implemented.

Still, until standards grow to maturity, whenever content is offered to the Catalogue from one of the interoperable resources, the Data Quality Officer of the Catalogue should inspect it before it is accepted.

Refer to Annex 1, Table 1-1, Challenge F.

3.8 Quality checks

A consistent finding from the interviews conducted during the first part of the project with experts in the field was that central quality checks are necessary: first, to ensure that the Catalogue is still functioning; second, to make sure that the updates are consistent with the definitions; and third, to increase the perceived value of the Catalogue across users.

This finding and the experience during the MINERVA pilot indicate that the Catalogue Quality Office should have roles at multiple levels. Automatic checks (such as field validation rules, including format of values, checks for completeness of mandatory fields, and checks for data value distributions and abnormalities) can be implemented in the software managing the Catalogue. Checking the consistency of the content was piloted during MINERVA and proved to be an intensive task that required a deep understanding of both the metadata definitions and the content that the DAP was conveying. Such a task requires expertise in pharmacoepidemiology/RWE and regulatory-grade multiple-database studies.

The intense activity of quality checking the metadata content will increase the value of the Catalogue as perceived by contributors, which in turn will foster their engagement.

3.9 Required efforts for implementing and maintaining a sustainable catalogue

For each activity, Table 1 and Table 2 indicate the source of the required effort and some notes on quantification of such effort: orange cells indicate that resourcing of the Catalogue controller is required, although specifics may be outsourced; brown cells indicate need of external resources to be funded; green cells indicate external resources funded by studies. The effort is quantified whenever this is possible, based on the pilot or on expertise of the MINERVA Consortium. Refer to Annex 1, Table 1-1, Challenge D.

Table 1. Activities during Catalogue set-up

Asset of the metadata Catalogue	Activity during set-up	Source of effort	Quantification of effort
Catalogue software deployment	Installation	Catalogue controller	Use of existing open-source Catalogue resources or tailor-made software
	Set-up of authentication system	Catalogue controller	Possible use of existing open-source resources
	Service desk implementation	Catalogue controller	Personnel and ticket service. Service-level agreement of 24/7 ticket service and 8-hour availability during working days in the Western, Central and Eastern European time zones
	Updating catalogue software	Catalogue controller/ developer	On demand, expect 1 day per week
Population of Institution, qualitative metadata in Data Source and Data Bank sections	First completion	DAP	3 person-days per data source, or interoperability with existing FAIR catalogues
	Expert support for initial manual population	Catalogue controller/expert	1 person-day per data source, and set-up of interoperability with other FAIR catalogues
Population of quantitative metadata in Data Source and Data Bank sections	Quality check feedback	DAP/DARWIN-EU Catalogue controller	Quality checks on data from, e.g., DARWIN-EU or dedicated catalogue study
Population of Common Data Model section	Manual entry of ETL (extract, transform, load) design	DAP	3 person-days per CDM
Population of Network section	Manual	DAP	1 person-day per network
Population of Study section	Manual	DAP	Study based

CDM = common data model; DAP = data access partner.

Table 2. Activities during Catalogue maintenance

Asset of the metadata Catalogue	Activity during maintenance	Source of effort	Quantification of effort
Tool	Maintenance & updating of software	Catalogue controller	To be specified
Updating of metadata content	Applied study specific	DAP	Studies
	Annual, independent of applied studies but update of overall metadata content	DAP	1 day
Population of quantitative metadata in Data Source and Data Bank sections	Repeated update	DARWIN-EU	DARWIN-EU
Population of Common Data Model section	Update, manual	DAP	Studies
	Annual, study independent	DAP	0.5 day
Population of Network section	Update, manual	DAP	Studies
	Annual, study independent	DAP	0.5 day
Population of Study section	Regular population	DAPs	Studies
Quality Office	Quality control of updates, using both automatic and manual checks; update of educational material	Catalogue controller	1 full-time equivalent over entire lifetime, periodically, recommended annually

DAP = data access partner.

3.10 Applied examples of use

In this section, some practical examples of use of the Catalogue are illustrated.

Example 1. A researcher wants to identify which data sources/data banks contain sufficient information about a rare disease to be able to (a) calculate prevalence of the disease in the underlying population of the data source, (b) follow the natural history of patients with the disease over time, or (c) study treatment of the disease over time

This is an instance of Use Case 2. The researcher (who also may be a regulator) accesses the Catalogue and enters the Studies section first. Using the searching functionalities on the metadata of Section **F2**, the researcher can identify studies conducted in the past on the same rare disease. Alternatively, once the Data Source metadata catalogue will be linked to the Study metadata catalogue, the researcher may search for corresponding studies in the Study catalogue and, if such study has been performed using a data source registered in the Data Source catalogue, access linked information of the Data Source catalogue. This process allows the researcher to inspect the data sources that contributed to such studies and learn which strengths and limitations were encountered during their conduct, as a preliminary step to build their discovery of data sources of interest. First, they can check which algorithms were used to identify persons with the rare disease in the data

sources of the existing studies (Sections **F3**, **F9C5**, and **F9C6**). Moreover, they can learn whether validation studies were conducted on such algorithms, in which data sources, and what the results were (Section **F3**). Validity indices may allow to include in the study design a quantitative adjustment of the observed prevalence. Second, the investigators can assess whether the severity of the disease could be measured in the data sources participating in the study, which allows studying the natural history of the disease, and identify which prompts and contents were used in such algorithm(s) (Sections **F3, F9C5**, and **F9C6**). Additionally, the investigators can assess whether the data sources with the best quality of identification of the disease were also able to identify the treatment of interest, and how (Section **F3**). Finally, the researcher can access quantitative metadata (Section **F5.8**) to assess data characterisation of the data sources. In case in the available studies a limitation is acknowledged that no data source is optimal to identify all the variables of interest (diagnosis of the disease, levels of severity, and treatments), then a strategy has to be devised to complement the information obtained from one data source with what can be obtained from another. The experience recorded in the previous studies can inform the search for new data sources whose data banks have prompts (Section **C5**) and content (Section **C6**) similar to the best ones used in the available studies and inspect whether there are general comments on the data banks quality section (Section **C9**) that would guide in assessing whether the data source is fit-for-purpose. The investigator can additionally perform a general search, see the next paragraph.

If past studies on the same disease cannot be found, the investigator should first search whether disease registries for the rare disease of interest are recorded in the Data Banks section of the Catalogue, using the metadata on family of data banks (Section **C5**). If such data banks exist, the Catalogue can help in understanding whether they belong to data sources where the underlying population can be identified as well (Sections **B1**, **B4,** and **C1**), to be able to estimate prevalence. If metadata on the content of the disease registries (Section **C6**) report that they contain information on history of the disease and/or treatment, the data source can be used to address all the research questions. Otherwise, if the disease registries belong to data sources where other data banks can be linked (Section **B4**), the investigator can understand whether the resulting data sources can be used to also address the other research questions. The investigator can also understand the geographic area of reference of the underlying population (Section **C1**). Moreover, if quantitative metadata of the data source are available (Section **B6**), the investigator may have a chance to perform a preliminary power calculation.

If the data source(s) containing a disease registry for the rare disease are not sufficient (because the power is insufficient, or the geographic coverage not appropriate, or for other reasons), an a priori knowledge of the disease clinical history is to be gathered outside of the Catalogue before the search can be performed. Namely, the investigator must gather information on the disease typical pathway of treatment in Europe and map the steps in the diagnosis and treatment of the disease to the set of prompts documented in the Catalogue. The investigator can then use the metadata in Section **C6** to search for data sources that include all the prompts, or at least some of them, and use metadata in Section **C5** to check whether the content that is prompted is sufficient to gather the study variables of interest, that is, the onset of the disease, the progression to severity levels, and the treatment evolution. Since prevalence estimates are requested, data sources should be chosen that also include the entry and exit from the underlying population, which can be checked in Sections **B1** and **B4** of the data sources.

In all cases, once a shortlist of data sources is identified, the researcher can use the Catalogue to explore quantitative metadata (Sections **B6**), Finally, the investigator can use the Catalogue to find DAPs that would be available to further contribute to the fitness-for-purpose assessment. The expertise, type of data access, and contact point of the institutions can be found (Sections **B2** and **A1**).

Example 2. A regulator wants to identify which data sources/data banks can be used to study the relationship between a drug treatment prescribed in first-trimester pregnant women and a health outcome in the offspring up to the age of 5 years

This is an instance of Use Case 2.

The researcher (who also may be a regulator) accesses the Catalogue and enters the Studies section first. Using the search functionalities on the metadata of Section **F2**, the researcher can identify studies conducted (a) on the drug treatment of interest, (b) on association between an exposure at the early stages of pregnancy and a mid-term outcome in the offspring, and (c) on the specific health outcome of interest. Alternatively, once the Data Source metadata catalogue will be linked to the Study metadata catalogue, the researcher may search for corresponding studies in the Study catalogue and, if such study has been performed using a data source registered in the Data Source catalogue, access linked information of the Data Source catalogue.

This process allows the researcher to inspect the data sources that contributed to such studies and learn which strengths and limitations were encountered during their conduct, as a preliminary step to build their discovery of data sources of interest. First, they can check which algorithms were used to identify utilisation of the drug of interest in the data sources of the existing studies, and the health outcome of interest (Sections **F3**, **F9C5**, and **F9C6**). Moreover, they can learn whether validation studies were conducted on such algorithms, in which data sources, and what the results were (Section **F3**). Validity indices may allow to include in the study design a quantitative adjustment of the observed health outcome. Second, the investigators can assess how pregnancies resulting in a live birth were identified in the selected studies, in particular which prompts and contents were used in such algorithm(s) (Sections **F3, F9C5**, and **F9C6**). Additionally, even if no study is found addressing all the research questions (a), (b), and (c) listed above, the investigators can assess whether there are data sources that have prompt and content similar to those used to address all the research questions separately (Sections **F9C5** and **F9C6**). Finally, the researcher can access quantitative metadata (Section **F5**) to assess data characterisation of the data sources. In case no data source is optimal to identify all the variables of interest (pregnancy and exposure in women, health outcome in linked offspring), then a strategy has to be devised to complement the information obtained from one data source with what can be obtained from another. The experience recorded in the previous studies can inform the search for new data sources whose data banks have prompts (Section **C5**) and content (Section **C6**) similar to the best ones used in the available studies and inspect whether there are general comments on the data banks quality section (Section **C9**) that would guide in assessing whether the data source is fit-for-purpose. The investigator can additionally perform a general search, see the next paragraph.

If past studies on some of the research questions (a), (b), and (c) listed above cannot be found, an a priori knowledge of the clinical history of the health outcome of interest is to be gathered outside of the Catalogue before the search can be performed. Namely, the investigator must gather information on the disease typical pathway of treatment in Europe and map the steps in the diagnosis and treatment of the disease to the set of prompts documented in the Catalogue. The investigator can then use the metadata in Section **C6** to search for data sources that include all the prompts, or at least some of them, and use metadata in Section **C5** to check whether the content that is prompted is sufficient to gather the study variable of interest. At the same time, the investigator can search birth registries and/or congenital anomaly registries recorded in the Data Banks section of the Catalogue, using the metadata on family of data banks (Section **C5**). Linking back to the Data Source section (Section **B4**), the investigator can understand whether data sources exist that have access to all the data banks needed to identify both treatment exposure and the health outcome in a population of, respectively, pregnant women and their offspring: in a data source, data banks can be linked to one another at the individual level, and Section **B5** can be explored to learn whether linkage is complete.

The investigator can also understand the geographic area of reference of the underlying population (Section **B1**).

In all cases, once a shortlist of data sources is identified, the researcher can use the Catalogue to explore quantitative metadata (Sections **B6**), Finally, the investigator can use the Catalogue to find DAPs that would be available to further contribute to the fitness-for-purpose assessment. The expertise, type of data access, and contact point of the institutions can be found (Sections **B2** and **A1**).

Example 3. A regulator wants to study a drug treatment prescribed (a) in primary care or (b) during hospital care and follow the treated patients for outcome events that lead to hospitalisation or are fatal

This is an instance of Use Case 2.

The strategy to utilise the Catalogue is similar to the case of Example 1 and 2: explore the Studies section first, or use the Study metadata catalogue, to build on the knowledge recorded from previous studies, and then dive into the Data Banks and Data Source section.

The specific challenge of this example is that hospital administrative records normally do not record administration of medications during inpatient stays. Data sources detailing treatments administered in an inpatient setting are typically hospital *electronic medical records*, which rarely in Europe can identify the underlying population and/or link to other hospitals, because linkage to the corresponding data banks is rarely available, because originators and purposes of the data banks are different. This makes it challenging to observe on a same study population at the same time inpatient treatment, hospitalisation (which make take place in a different hospital) and death (which may take place in a different hospital or at home). Conducting the study on such data sources would therefore lie on assumptions, that can be tested on the few data sources that happen to be able to link all the necessary data banks.

However, this challenge may be less of a concern for specific classes of drugs (e.g., some cancer treatments), that in recent years have been recorded in some hospital administrative records.

The functionalities of the Catalogue illustrated in Examples 1 and 2 allow shortlisting the candidate data sources and investigate their fitness-for-purpose. In particular, the description of all the data banks included in the data source allows to understand whether information on other diagnostic tests (laboratory, or imaging or other) is also prompted, and what assumptions can be made according to the experience of the DAP that has entered the metadata.

Example 4. A regulator wants to assess the suitability of data sources used in a study

This is an instance of Use Case 4. The consumer of evidence (who also may be a regulator) accesses the Catalogue, and enters the Study section. Using the search functionalities, the regulator finds the entry of the study of interest. Links to EU PAS Register entry (Section **F2**), the protocol (Section **F2**), and the publications (Section **F6**) of the study are available. A full description of the data sources involved in the study is included, as described by expert DAPs at the time of the study, including all details of the corresponding data banks (Sections **F8** and **F9**). This allows to assess in a succinct manner strengths and limitations of the data sources that are independent of the study question. Additionally, qualitative (Section **F3**) and quantitative (Section **F5**) assessments of data quality are available, that specifically detail strengths and limitations tailored to the research question. Such assessments were recorded by DAPs who are scientifically accountable for the accuracy of their entries at the time when the study was conducted (Sections **F1** and **F3**), and that can be contacted in case additional detail is needed.

Example 5. A data originator of a data bank updates the data model of the data bank tables

This is an instance of Use Case 5. The Use Case applies when the data originator of a data bank maintains a repository of metadata describing the data model of the data bank tables, and this repository complies with the FAIR principles and is interoperable with the Catalogue. Whenever the repository of the data originator is updated, the Catalogue is automatically offered the update for incorporation into the Catalogue.

Example 6. A FAIR metadata catalogue is updated with new information

This is an instance of Use Case 5. The Use Case applies when an entire metadata catalogue is FAIR and is interoperable with the Catalogue. Whenever the external catalogue is updated, the Catalogue is automatically offered the update.

Example 7. A DAP wants to convey both study-independent and study-specific strengths and limitations of a complex data source

This is an instance of Use Case 1.

Imagine that hospital A makes its electronic medical records (or its administrative records) available for research. In Section **C5,** the prompt of the data bank is described as an admission or visit in hospital A. In Section **C1,** the DAP must describe what is the best definition of underlying population. Four scenarios are plausible (the last one is documented in the POC Catalogue).

The first scenario is when no data bank exists that describes the population accessing hospital A. If this is the case, the description of the underlying population is only qualitative. Based on the characteristics of the healthcare system where hospital A is located and on the characteristics of hospital A, the most appropriate qualitative definition may be "persons living in the geographical area where hospital A is located" or "persons with health insurance that has hospital A as a recommended healthcare facility" or some other description. The data source comprising only the single hospital A data bank must be used with caution for research questions that require follow-up since it is possible to follow the persons contributing to the data source only when they access hospital A again. Moreover, between two such records for the same person, it is impossible to ascertain whether the person has visited hospital B, which serves the same population. However, there may be studies where such limitations of hospital A data can be addressed. For example, when studying a health condition for which hospital A is a reference centre in the geographical area, it may be assumed that all the contacts relative to care of that condition will happen within hospital A and therefore prompt records in the data source. When the DAP initiates a study using this data source, this assumption is recorded in the study-specific area of the data source, Sections **F8B1** and **F3**.

In the second scenario, a data bank P records when persons enter and exit the population served by hospital A (for instance, the geographical area), then the data source including data bank P and hospital A has the potential of following study subjects, for example, for their vital status or to assess whether they leave the population. This possibility is recorded in Section **C1** of the hospital A data source and in Section **B1** of the data source including hospital A and data bank P. Whenever a study starts, if the DAP accessing the data source has access both to hospital A and data bank P, this information is copied in Sections **F8B1** and **F9C1**. If hospital B serves the same population, but no data bank is available where access to hospital B prompts records, then metadata in Section **F3.12** will document how this limitation is addressed in the study for each study variable of interest.

The third scenario is a special case of the second scenario. If there is reason to assume that a subpopulation that can be identified in data bank P is accessing mostly hospital A instead of hospital B (for example, inhabitants of a subregion within the geographical area or persons with a specific

insurance scheme), the DAP may record this information in Section **B1**. When a study starts, if deemed useful, the DAP may record in Section **F3.12** that restricting the analysis to such a subpopulation is recommended, possibly as a sensitivity analysis. Indeed, the analysis restricted to such a subpopulation will have less power but may provide more accurate results. The usefulness of this strategy is dependent on the study question.

The fourth scenario is a special case of the third scenario. In this case, there is a subpopulation in data bank P that is accessing mostly hospital A instead of hospital B, but the data source includes other data banks whose underlying population is the whole data bank P. An example of a data source with this configuration available in the POC Catalogue is BIFAP, where primary care medical records (data bank BIFAP_EMR_PRIMARY_CARE) are collected for an underlying population of 9 Spanish regions, but only 5 of the regions have administrative hospital records available (data bank BIFAP_DIAGNOSIS_HOSPITAL_INPATIENTS). In this case, the information recorded in Section **C1** of the data bank BIFAP_DIAGNOSIS_HOSPITAL_INPATIENTS is "the entire population belonging to the Spanish National Health system and located in a subset of patients living in 5 of the 9 regions participating in the BIFAP programme." When BIFAP participated in the ACCESS study (Willame et al., 2021), the data source was queried twice: the first time, the whole population was analysed, using only BIFAP_EMR_PRIMARY_CARE (which was indicated in the report as BIFAP_PC); the second time, the analysis was restricted to the subpopulation of 5 regions, using both BIFAP_DIAGNOSIS_HOSPITAL_INPATIENTS and BIFAP_EMR_PRIMARY_CARE (which was indicated in the report as BIFAP_PC_HOSP). In this manner, the results of BIFAP_PC_HOSP had less power but were more accurate than those in BIFAP_PC, but both results could be interpreted transparently. This information could be recorded in the ACCESS entry in the Studies section of the Catalogue, in Sections **F8B**, **F9C**, and **F3**.

In all the scenarios, during a study, the DAP can describe in Section **F3.10** the strengths and limitations of its data source in addressing the study question; in Section **F3.12,** the DAP can enter details of strengths and limitations of each specific study variable; and finally, in Section **F3.11**, text can be recorded that links such strengths and limitations to the actual interpretation of the results.

4 Conclusions

The research community has been involved for decades in RWE research for regulatory decision making. For all, increased data discoverability has a clear positive impact; it allows more efficient and higher quality studies. It will also increase the transparency, discoverability, and reproducibility of the studies we conduct. Developments in this area will result in the rebuilding of tools that the research community and other stakeholders currently use, namely the ENCePP Resource Catalogue and the ENCePP Study Registry, which also serves as the EU PAS Register. The challenge has always been to create sustainable and FAIR data source and study metadata catalogues.

The MINERVA Consortium, 18 ENCePP centres and collaborators, in collaboration with the Agency, experts from many European and international RWE research networks, along with the input of stakeholders during a public workshop in April 2021, developed and discussed a set of recommendations for the sustainable set-up and maintenance of a future metadata Catalogue for data sources for RWE studies.

These recommendations were based on a stepwise approach of defining an initial list of metadata, a stakeholder consultation to consolidate the list, and a pilot with 15 heterogeneous European data sources, including electronic health records and patient registers, to collect a defined subset of metadata about the selected data sources. The metadata were collected, and a proof-of-concept catalogue to display and search through the metadata was developed for several use cases and applied

examples of use. The use cases included a DAP provides expertise throughout the cycle of a study; an investigator uses the Catalogue throughout the life cycle of a study; a programmer programs a study; a consumer of evidence assesses reproducibility and quality of evidence generated by a study; a FAIR process launched by an external organisation, including a data originator, populates the Catalogue; and an institution with research capabilities becomes a DAP for a data bank and/or a data source.

Based on the experience of the MINERVA pilot, setting up and maintaining an operating metadata Catalogue on real-world data sources will require a substantial effort to implement the FAIR principles, adhere to data protection rules, and effectively support discoverability of data sources and reproducibility of studies in Europe. We have assessed the effort required to implement and maintain a sustainable catalogue—the specific assets of the Catalogue, its activities, and the source of the effort—and created a qualitative categorisation for the foreseen effort.

The key deliverables of the MINERVA project are the list of metadata and this guidance document. This collaboration across research networks and stakeholders within Europe and across the world needs to continue in the future to take this work to the next level of fully sustainable and FAIR data source and study metadata catalogues.

5 References

Aetion. 2021. Real-world evidence solution | RWE Analytics | Aetion. [online]. https://aetion.com. Accessed 1 March 2021.

Arlett P, Kjaer J, Broich K, Cooke E. Real-world evidence in EU medicines regulation: enabling use and establishing value. Clin Pharmacol Ther. 2021 Nov 19. doi:10.1002/cpt.2479.

Aspennet.asia. 2021. Asian pharmacoepidemiology network [online]. https://aspennet.asia/. Accessed 1 March 2021.

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

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

EMA. European Medicines Agency. Annex VIII to Technical Specifications no. EMA/2021/08/TDA. Personal data processing risk mitigations and controls. 3 June 2021b. https://etendering.ted.europa.eu/cft/cft-documents.html?cftId=8503. Accessed 21 October 2021.

EMA. European Medicines Agency. Data Analytics and Methods Task Force. Technical workshop on real-world metadata for regulatory purposes. EMA/139729/2021. 2 September 2021a. https://www.ema.europa.eu/en/documents/other/summary-report-technical-workshop-real-world-metadata-regulatory-purposes_en.pdf. Accessed 22 October 2021.

EMA. European Medicines Agency. Data Analytics and Methods Task Force. Technical workshop on real-world metadata for regulatory purposes. 12 April 2021c. https://www.ema.europa.eu/en/events/technical-workshop-real-world-metadata-regulatory-purposes. Accessed 22 October 2021.

ENCePP. European Network of Centres for Pharmacoepidemiology and Pharmacovigilance. The European Union Electronic Register of Post-Authorisation Studies (EU PAS Register). 2020. http://www.encepp.eu/encepp_studies/indexRegister.shtml. Accessed 2 August 2021.

ENCePP. European Network of Centres for Pharmacoepidemiology and Pharmacovigilance. ENCePP Resources Database. 2021. http://www.encepp.eu/encepp/resourcesDatabase.jsp. Accessed October 2021.

European Data Protection Board. Guidelines 07/2020 on the concepts of controller and processor in the GDPR. Version 1.0. Adopted on 02 September 2020. https://edpb.europa.eu/sites/default/files/consultation/edpb_guidelines_202007_controllerprocessor_en.pdf. Accessed 19 October 2021.

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

Gini R, Schuemie M, Brown J, et al. Data extraction and management in networks of observational health care databases for scientific research: a comparison of EU-ADR, OMOP, Mini-Sentinel and MATRICE Strategies. EGEMS (Wash DC). 2016 Feb 8;4(1):1189. doi:10.13063/2327-9214.1189.

HMA-EMA Big Data Steering Group. Workplan. 27 Jul 2020. 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 December 2020. https://www.ema.europa.eu/en/documents/other/priority-recommendations-hma-ema-joint-big-data-task-force_en.pdf. Accessed 3 March 2021.

IMI. Innovative Medicines Initiative. ConcePTION D7.5 Report on existing common data models and proposals for ConcePTION. 2020. 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 21 October 2021.

Imi-conception.eu. 2021. ConcePTION [online]. https://www.imi-conception.eu/. Accessed 1 March 2021.

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

Sturkenboom M. Cohort monitoring of adverse events of special interest and COVID-19 diagnoses prior to and after COVID-19 vaccination (ECVM study). 2021. http://www.encepp.eu/encepp/viewResource.htm?id=40974. Accessed 22 September 2021.

Thurin NH, Pajouheshnia R, Roberto G, Dodd C, Hyeraci G, Bartolini C, et al. From inception to ConcePTION: genesis of a network to support better monitoring and communication of medication safety during pregnancy and breastfeeding. Clin Pharmacol Ther. 2021 Nov 26. doi: 10.1002/cpt.2476. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/cpt.2476.

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

Wilkinson M, Dumontier M, Aalbersberg I. 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.

Willame C, Dodd C, Gini R, et al. Background rates of adverse events of special interest for monitoring COVID-19 vaccines (2.0). Zenodo. 2021. https://doi.org/10.5281/zenodo.5255870. Accessed 22 September 2021.

Annex 1: MINERVA project work substantiating the recommendations

Literature review, interviews with stakeholders and stakeholders' consultations to define the metadata

This section summarises Deliverable 3 (Preliminary set of metadata and definitions, process and catalogue tool), Deliverable 4 (Technical workshop on real-world metadata for regulatory purposes), and Deliverable 5 (Final set of metadata and definitions, process, and catalogue tool).

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.

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 (US Food and Drug Administration [FDA], 2021)
- CNODES (Canadian Network for Observational Drug Effects Studies) (Cnodes, 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)

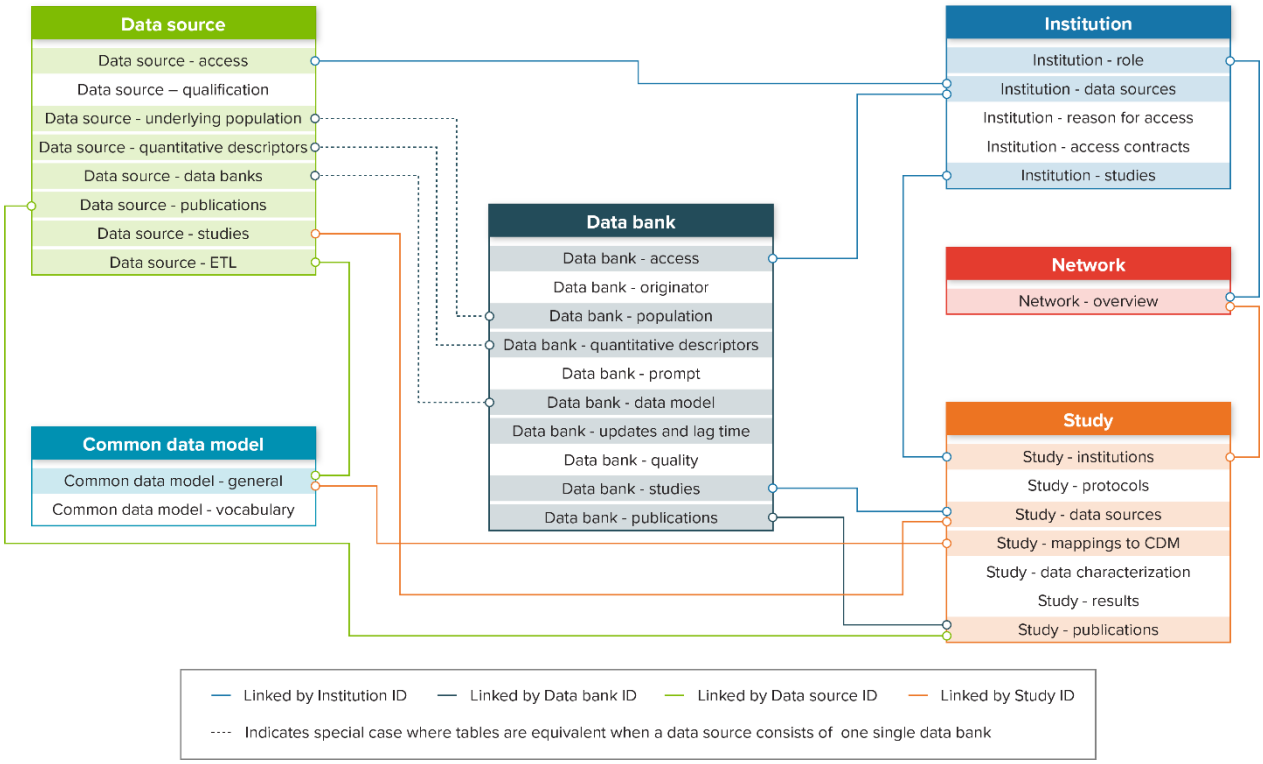
The interviews followed a standard set of questions, which were piloted in the first interview with representatives from IMI-ConcePTION. The information from the interviews was used to (1) refine the scope of metadata envisioned for the catalogue, (2) identify additional key metadata variables to collect, (3) gather additional resources that could be used to inform the metadata list, (4) identify existing examples of tools that can be used to access or visualise metadata, and (5) identify potential challenges or barriers to the implementation of the envisioned catalogue.

Finally, we incorporated the recommendations developed by ENCePP Working Group 3 based on its review of the EU PAS Register (ENCePP, 2020). Conclusions of the Working Group are that entries in the current EU PAS Register are often incomplete, show various degrees of accuracy, and seem to suggest that some fields may be poorly understood. The interpretation by the Working Group is that the Register is only perceived as a formal requirement and has issues around completeness, maintenance and sustainability.

The preliminary list of metadata was the object of a stakeholder's consultation that culminated in an online recorded technical workshop on April 2021 (EMA, 2021c).

Figure 1-1 shows a high-level relational model of the metadata tables.

Figure 1-1. 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.

MINERVA pilot

The MINERVA pilot is described in Deliverable 6 (Collection of metadata & tool) and Deliverable 7 (Report on data collection and tool). We summarise here the findings most relevant for the development of the list of recommendations in Section 2.5 of this document.

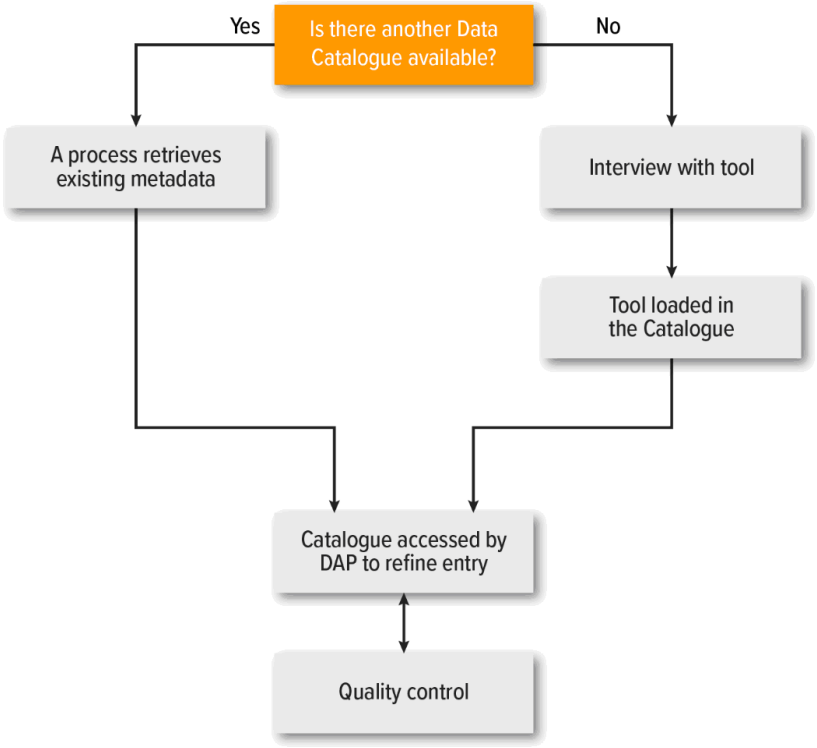
Institution section and qualitative metadata of Data Source and Data Bank sections: first study-independent completion of the POC Catalogue

Two different methods to populate the list of metadata were used in this pilot. In total, the POC catalogue was piloted by 15 DAPs for 15 data sources.

- A total of 11 DAPs had collected a subset of metadata in a compatible way in previously used tools and projects; the data models of those tools were mapped to the MINERVA metadata list. These metadata were then transferred into the MINERVA POC catalogue, and the missing values could be added manually. This process, depicted in the left column in Figure 1-2, was a pilot of the interoperability between FAIR catalogues.
- The remaining four DAPs had no pre-existing metadata available. Metadata were collected using a process developed initially for the IMI-ConcePTION project (project deliverable 7.9, https://www.imi-conception.eu/wp-content/uploads/2021/01/D7.9-test-report-v1.0.pdf) and tailored to the MINERVA POC catalogue. A data entry tool was created specifically for the MINERVA POC catalogue and was used to capture information from the interview, which was preceded by a request for preliminary information. The interview tool consisted of a specifically curated

spreadsheet of metadata, which was filled in during interviews with DAPs. After a round of offline completion and revision, the metadata were loaded manually from the interview tool into the MINERVA POC catalogue. This process, depicted in the right column in Figure 1-2, was a pilot of direct data entry into the catalogue.

Figure 1-2. Retrieval of pre-existing metadata and entry of new metadata



DAP = data access partner.

In both cases, the catalogue entry was accessed again by the DAPs and further refined.

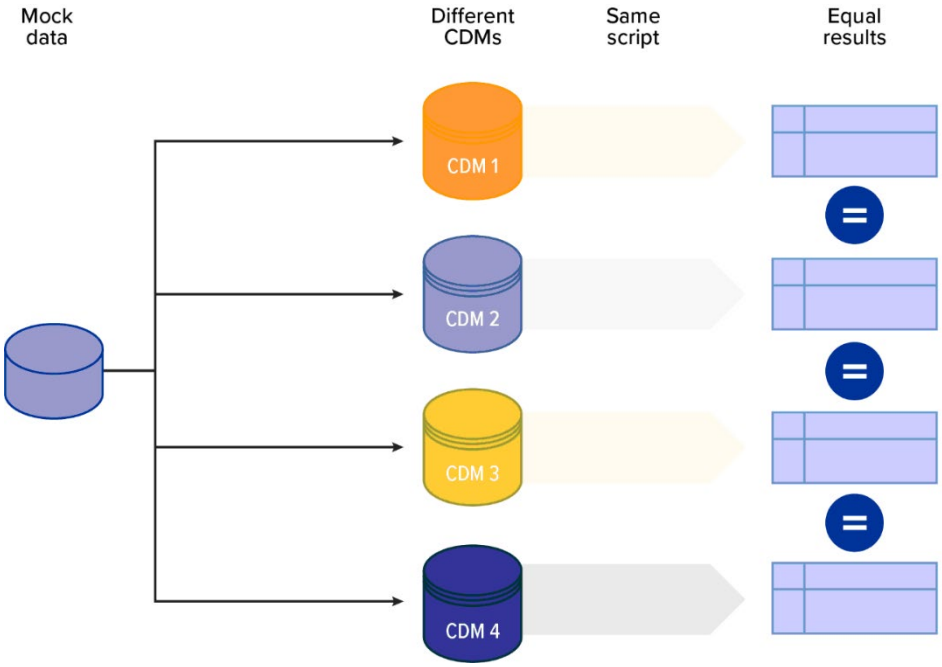
Finally, a round of quality checks was performed by the MINERVA experts. The experts inspected the inputted information for the following: (1) incomplete or missing items, (2) possible misunderstanding of metadata items (where the entry suggested that the meaning of the field was misinterpreted), and (3) typographical errors. The DAPs were notified of the errors observed and were asked to revise their input, possibly with suggestions. This was done manually and was a pilot for the quality-assurance activity.

Quantitative metadata of Data Source section: study-independent update

We selected the example of age and sex distribution of the underlying population of a data source to develop and test a script that could retrieve quantitative metadata from any of four supported common data models (CDMs). The following steps were enacted:

- A data set of a mock population data was created
- The data set was mapped to the four CDMs: Observational Medical Outcomes Partnership [OMOP], ConcePTION, Nordic, and TheShinISS
- A script was built to run on multiple CDMs, which is a parameter of the script
- The script was run against each of the four CDMs
- The four resulting output data sets were proven to be the same (see Figure 1-3)

Figure 1-3. Flow diagram of the data set output test

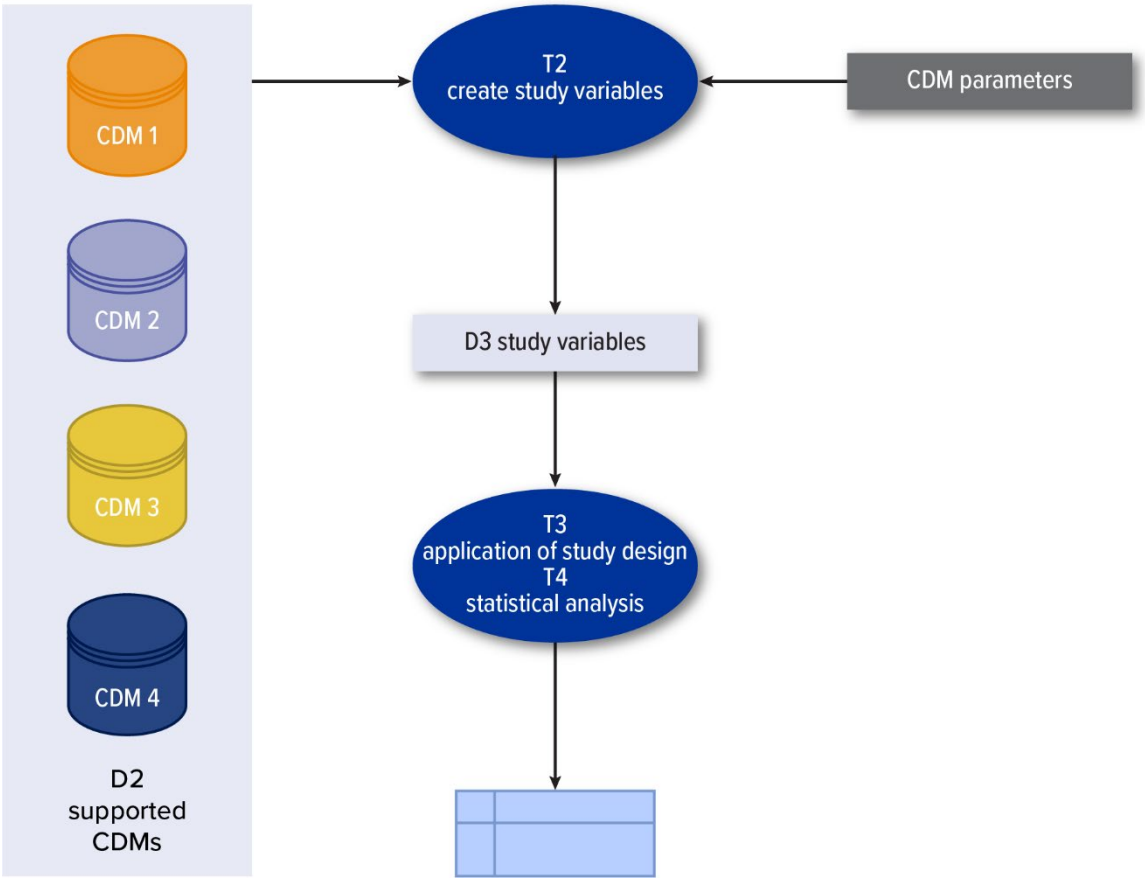


CDM = common data model.

This exercise is available on GitHub (https://github.com/ARS-toscana/MINERVA_samplescript).

Support for multiple CDMs is obtained by splitting the script in steps, as illustrated in Gini et al. (2016). The step "T2 – create study variables" needs to be adapted to the different CDMs and could also be applied outside the scope of a study (provided that data banks undergo an ETL [extract, transform, load] procedure to an applicable CDM prior to running the script). The output of this step is the same, irrespective of the CDM, so the next steps of the script (applying the study design to create the analytical data set and running the statistical analysis) is the same irrespective of the CDM, see Figure 1-4.

Figure 1-4. Steps of a script to generate quantitative metadata from multiple common data models



CDM = common data model; D*n* = dataset; T*n* = transformation.
Source: Adapted from Gini et al. (2016).

Four DAPs were asked to run the script on an instance of their data source already extracted and converted to one of the four CDMs for different studies. Two DAPs were allowed to repurpose the extraction for the MINERVA pilot. The other DAPs would have needed to obtain additional approvals. Therefore, they provided the requested results using publicly available data.

Study section

We piloted the population and maintenance of the metadata catalogue for study-specific data for two recent studies to which MINERVA data sources had contributed, ACCESS (Willame et al., 2021) and ECVM (Sturkenboom, 2021). Data entry for these two studies was manually opened and edited with general metadata such as name and description. Subsequently, a copy of the metadata of the corresponding data sources was copied into the Study section. The metadata of the data sources were then edited manually to reflect the actual use of the data sources in the two studies.

Common Data Model and Network sections

The sections Common Data Model and Network were updated manually as part of the population of the Study section. Information on the ConcePTION CDM adopted in the studies (IMI, 2020) was transferred from another compatible FAIR catalogue and edited to complete the entry. The Network section was edited manually.

Lessons learnt from the pilot

During the pilot, a number of challenges were identified, as listed in Table 1-1.

Table 1-1. Challenges during the population of the MINERVA pilot POC catalogue

	Challenge	Description of the challenge
A	Authentication of contributors	Contributors need to be identified within and across catalogues, and the role of each contributor needs to be defined Resulted in practice Recommendation 3b in Section 2.4
B	Responsibility of contributors, across updates of the same entry/version control	Multiple contributors can provide input on the same entry at different points in time. Each contributor may edit the content provided by others in previous instances. Resulted in practice Recommendation 3c in Section 2.4
C	Populating the qualitative metadata when expertise on the conceptual framework is lacking	The metadata catalogue requires alignment of heterogeneous data sources to the same conceptual framework. DAPs with a deep understanding of the data sources to which they have access, but lacking the expertise in comparing them with others, may find it challenging to position their knowledge in this conceptual framework. This is particularly true for data sources including a single data bank. Resulted in practice Recommendation 2b in Section 2.4 and Section 3.1
D	Effort to populate qualitative metadata	Even though the pilot was focused on 202 priority metadata variables, less than half of the desired metadata variables, their collection and review required intensive effort both by contributors and expert reviewers Resulted in Section 3.9, Required efforts for implementing and maintaining a sustainable catalogue.
E	Authorisation to provide quantitative metadata	Repurposing existing instances of a data sources to obtain quantitative metadata is incompatible with the governance of many data sources Resulted in last 4 paragraphs of Section 3.4, Quantitative metadata in the Data Source and Data Bank sections.
F	Ontologies	For many metadata variables, specific ontologies need to be defined or refined Reflected in Section 3.7, Interoperability of catalogues.
G	Quality control	Quality control is necessary to ensure that the system remains functional after population and/or update and is useful to provide feedback and increase the perceived value Resulted in Section 3.8, Quality checks.

DAP = data access partner.

Annex 2: Final metadata list

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