Article published online: 2022-12-04

Towards an Interoperable Ecosystem of Research Cohort and Real-world Data Catalogues Enabling Multi-center Studies

Morris Swertz[1]*, Esther van Enckevort[1], José Luis Oliveira[2], Isabel Fortier[3], Julie Bergeron[3], Nicolas H. Thurin[4], Eleanor Hyde[1], Alexander Kellmann[1], Romin Pahoueshnja[5], Miriam Sturkenboom[6], Marianne Cunnington[7], Anne-Marie Nybo Andersen[8], Yannick Marcon[9], Gonçalo Gonçalves[10], Rosa Gini[11]*

[1] Department of Genetics, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands
[2] DETI/IEETA, University of Aveiro, Portugal
[3] Research Institute of the McGill University Health Center, Montreal, Canada
[4] Univ. Bordeaux, INSERM CIC-P 1401, Bordeaux PharmacoEpi, Bordeaux, France
[5] University of Utrecht, The Netherlands
[6] Department of Datascience & Biostatistics, Julius Center, University Medical Center Utrecht, Utrecht, The Netherlands
[7] GlaxoSmithkline, Stevenage, Herts, SG1 2NY, UK
[8] University of Copenhagen, Copenhagen, Denmark
[9] Epigeny, France
[10] Human-Centered Computing and Information Science, INESC TEC, Portugal
[11] ARS Toscana, Florence, Italy

* Corresponding authors

Summary

Objectives: Existing individual-level human data cover large populations on many dimensions such as lifestyle, demography, laboratory measures, clinical parameters, etc. Recent years have seen large investments in data catalogues to FAIRify data descriptions to capitalise on this great promise, i.e. make catalogue contents more Findable, Accessible, Interoperable and Reusable. However, their valuable diversity also created heterogeneity, which poses challenges to optimally exploit their richness.

Methods: In this opinion review, we analyse catalogues for human subject research ranging from cohort studies to surveillance, administrative and healthcare records.

Results: We observe that while these catalogues are heterogeneous, have various scopes, and use different terminologies, still the underlying concepts seem potentially harmonizable. We propose a unified framework to enable catalogue data sharing, with catalogues of multi-center cohorts nested as a special case in catalogues of real-world data sources. Moreover, we list recommendations to create an integrated community of metadata catalogues and an open catalogue ecosystem to sustain these efforts and maximise impact.

Discussion: We propose to embrace the autonomy of motivated catalogue teams and invest in their collaboration via minimal standardisation efforts such as clear data licensing, persistent identifiers for linking same records between catalogues, minimal metadata 'common data elements' using shared ontologies, symmetric architectures for data sharing (push/pull) with clear provenance tracks to process updates and acknowledge original contributors. And most importantly, we encourage the creation of environments for collaboration and resource sharing between catalogue developers, building on international networks such as OpenAIRE and research data alliance, as well as domain specific ESFRIs such as BBMRI and ELIXIR.

Keywords

Catalogs as topic; data collection; metadata

Yearb Med Inform 2022:262-72
http://dx.doi.org/10.1055/s-0042-1742522

1 Introduction

Existing individual-level human data cover large groups of human individuals, populations and environments through high-value and longitudinal data capturing multiple dimensions, such as lifestyle, demography, laboratory measures, omics, clinical parameters, as well as data accounting for use of healthcare. Many data have been collected primarily for the purpose of research, such as population-based and clinical trial cohorts. In addition, even greater volumes of data have been collected for surveillance purposes, such as congenital anomalies, pathology, birth registries or infectious disease registries. Another example is data collected for the primary reason of healthcare conduct and/or administration that can also be used for secondary purposes such as research.

This broad range of pre-existing medical/administrative data that can be reused for research is often referred to as 'secondary use' or *real-world* (RW) data [1].

However, certain challenges must be overcome to fully benefit from the content, temporal and geographic diversity of information [2-4] and to support co-analysis of data across studies to reach the statistical power needed to elucidate the complex relationships between genetic traits, environment and disease and enhance capacity to undertake comparison, cross validation or replication analysis. In each situation and for each type of data collection, effective and efficient use and reuse of data critically depends on availability of detailed and specific metadata, as again recently acknowledged in the context of the COVID-19 pandemic [5].

Recent years have therefore seen large investments in data catalogues within the health research domain. This development has been in part due to the focus on data reuse and the FAIR principles movement, i.e., that data should be Findable, Accessible, Interoperable and Reusable [6], and in part because larger datasets are needed for much health-related research that cannot be provided by one single data source.

In this context, an increasing number of data catalogues are being developed to support:

- *discovering* data, to facilitate access to information and optimise usage of available data;
- *understanding* data, to assess their fitness-for-purpose to address specific research questions;
- *leveraging* integration of data, to address a broad range of research questions, increase statistical power, and enable comparison, cross validation or replication analysis.

While these catalogues are of great value, their heterogeneity poses challenges to optimally exploit their richness. This opinion review focuses on catalogues that describe data sources for human subject research such as cohorts, biobanks, registries, administrative and healthcare records. We propose a pathway towards a unified framework to enable interoperability and reuse of existing catalogues.

The review is organised as follows:

- In Section 2, we review examples of catalogues to understand their concepts and use cases;
- In Section 3, we analyse the conceptual overlap between existing catalogues as the basis for a unified conceptual framework for metadata catalogues for health research;
- In Section 4, we analyse requirements and pre-conditions to promote catalogue collaborations and reuse of catalogue contents;
- In Section 5, we make recommendations towards interoperable representation of data collections to enable sharing and cross-querying.

2 Existing Catalogues for Human Data

The subset of harmonisation catalogues is of particular interest in this review. Data collection may be prospectively harmonised, compatible with certified standards requiring straightforward transformation to the standard, although these standards might only apply to a given area, province, or country. However, implementation of a prospective approach is not always possible or suitable for research data, for example due to novel research questions or technical limitations [7], especially where data are generated for purposes other than research (e.g., as a part of routine healthcare). Retrospective harmonisation (i.e., harmonisation after data collection) is thus often the only option to permit data integration [8]. Below, we describe examples of catalogues based on experiences from the co-authors, first reviewing cohort data catalogues, then RW data catalogues, possibly also including cohorts.

2.1 Cohort Data Catalogues

One of the largest catalogues of human studies is the Directory of Biobanks of BBMRI-ERIC [9] with 1,829 collections from 619 organisations (https://directory.bbmri-eric.eu/#/), which is based on the MIABIS (Minimum Information About BIobank data Sharing) minimum information about a biobank standard [10, 11]. This catalogue does not provide information on the individual variables collected, nor does it emphasise data harmonisation. Another example is the International HundredK+ Cohorts Consortium catalogue (https://atlas.ihccglobal.org/), but this catalogue also does not provide information on variables. Examples of domain specific catalogues are Birthcohorts (http://birthcohorts.net), describing cohorts specifically from the perspective of data before or during pregnancy or latest at birth, including data on mother-child pairs, and ImmPort (https://www.immport.org/home) for open access immunological assay data for translational and clinical research [12].

An example of a catalogue that provides sufficient information to inform multi-center data analysis is the catalogue of the Maelstrom initiative. Maelstrom is an international collaboration of epidemiologists, statisticians, and computer scientists with a focus on enabling multi-cohort data analysis using data harmonisation for which they provide rigorous guidelines on what metadata to collect [7]. Their catalogue, https://www.maelstrom-research.org/, documents cohorts related to a particular research domain, e.g. paediatrics/parent-child, and networks of cohorts and researchers, such as international research consortia/projects, that mostly retrospectively harmonise data following the Maelstrom data harmonisation principles [13]. This catalogue provides detailed listings of 23 Networks, Individual studies (177 with variables) and Harmonisation projects (7, numbers checked Jan 2022). The Networks' list contains detailed information on collaborations between multiple cohort studies. The Individual Studies' list contains detailed descriptions of cohort studies, including the definition of their data collection timeline (e.g., baseline, follow-up), (sub)populations sampled (e.g., children from northern England), and most notably, in many cases, detailed listing of the variables collected (i.e., variable name, type, description, code list) (see Figure 1). Finally, the Harmonisation projects' list documents objectives of the harmonisation project, selection criteria for cohort studies and participants to be included, and detailed

descriptions to map collected variables from cohort studies onto harmonised variables using Maelstrom harmonisation guidelines [7]. This includes a detailed listing of the harmonised variables defined and then, for each cohort study, a mapping detailing whether variables from this cohort study could be harmonised, and if so, what algorithm was used. Of particular use is the 'Areas of information' classification that enables users to quickly find cohorts, networks, or variables on a particular topic (topics available in 18 domains and 135 sub-domains), for example: 'tobacco use' or 'cognitive functioning'. The Maelstrom catalogue software is available as open source via the Mica software in OBiBa (http://www.obiba.org) [14].

In Europe, several catalogues have emerged following the convention of Maelstrom, typically as consortia funded by the European Commission, starting with BioSHaRE [15, 16]. In the terminology of Maelstrom, each such consortium is a Network, and its catalogue a Network catalogue. A recent example is the LifeCycle catalogue (https://catalogue.lifecycle-project.eu/) [17]. This catalogue provides two entry points. By default, users are led to browse all harmonised variables, based on a keyword classification similar to (but not the same as) Maelstrom. For each variable, users can see if and how variables from partner cohorts in LifeCycle have been mapped. Secondly, users can browse the cohorts to receive general descriptions of the cohort. Currently, the LifeCycle catalogue is being expanded to include metadata from other harmonisation consortia, notably LongITools [18] and ATHLETE [19] and to be expanded to other consortia in the European Human Exposome Network (https://www.humanexposome.eu/). These catalogues are available as open source via the MOLGENIS software suite [20]. Another recent example is the RECAP Preterm catalogue (https://platform.recap-preterm.eu), containing information about pregnancy cohorts, which also makes use of OBiBa's Mica software to extensively catalogue the studies and data available on the project's platform. A variable classification system (in the vein of Maelstrom's 'Areas of information') was developed [21] and the project's variables were mapped to it in order to improve variable searching and harmonisation on the platform.

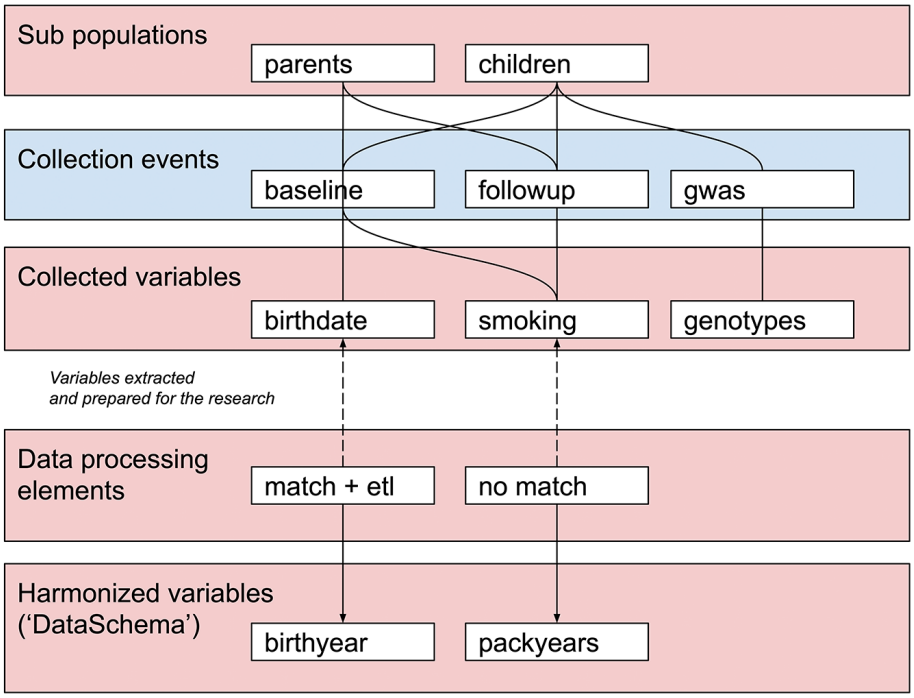

Fig. 1 Concept of collection events, populations and variables. Adapted from [13].

The large boxes define areas of metadata for multi-center cohort study catalogues: the population(s) being samples, e.g., 'Mothers in Northern Netherlands'; rounds/events of data collection, e.g., 'baseline' and 'followup'; and data dictionaries of variables collected, e.g., on 'cigarettes smoked per week'. In addition, these catalogues describe what harmonised research variables were defined for pooled analysis across multiple cohorts, e.g., 'pack years smoked'; and, for each cohort, if and how these harmonised variables can be constructed from the collected variables, e.g., match including algorithm or 'no match'. The small boxes are examples.

2.2 Real World Data Catalogues

Also, consortia using RW data have been establishing catalogues to document 'secondary use' data sources. The European Medicines Agency (EMA) in 2021 funded the MINERVA project to define a set of metadata to describe RW data sources. As an introductory step, the MINERVA project executed a search for relevant catalogues to describe RW data sources and we report here a selection of that list [22, 23].

The catalogue of the IMI-EHDEN project includes a dashboard allowing to display characteristics of the data sources participating in the project, all converted to a same common data model, and graphical and aggregated information about their content. It also allows the comparison of the harmonisation potential between different data sources [24]. The EMIF catalogue is an example of a catalogue representing both RW and cohort-derived data sources [25]. It is organised as a set of data research communities, each covering distinct data types, diseases, and users. It does not require pre-specified metadata to be catalogued: each community defines its own data model, to describe the catalogue entries. Despite this customisation facility, it lacks a common schema to integrate easily with external catalogue systems. The ENCePP Resource Database is an initiative of the EMA-sponsored European Network of Centres for Pharmacoepidemiology and Pharmacovigilance, which is freely accessible (https://www.encepp.eu/encepp/resourcesDatabase.jsp).

However, it does not provide a description of the data model of the data sources, and EMA funded the above-mentioned MINERVA project as part of the effort to improve the ENCePP catalogue.

With this mission, the MINERVA consortium has developed common data elements for RW data catalogues [26]. This catalogue was based on a conceptual framework from the IMI-ConcePTION Project [27], which is also embedded in the IMI-ConcePTION catalogue [28], and that we briefly summarise here: a data source is a collection of *data banks*, e.g. the National Health Registries of Denmark is a data source including the Danish National Patient Registry, the Danish Medical Birth Registry, and others. Each data bank is a collection of data that is mandated and sustained by an organization (in the example, both data banks are mandated by the Danish Health Data Authority) and where data generation is prompted by a class of events called *prompts*: in the example, data in the Danish National Patient Registry are prompted by discharges from Danish hospitals, and data in the Danish Medical Birth Registry are prompted by births happening in Denmark. Each data bank has a specified *underlying population*, i.e. the population whose events prompt records in the data bank. In a data source, all data banks must have the same underlying population (in the example, the inhabitants of Denmark), or populations that partially overlap, and must have the potential to be linked to one another at an individual level.

3 Comparison of Metadata Collected and How They Map onto Each Other, as The Basis for a Unified Conceptual Framework

In this section, we analyse the potential to harmonise and/or standardise catalogue contents between catalogues. We therefore analysed existing catalogues from Maelstrom, MINERVA, IMI-ConcePTION, LifeCycle, BBMRI/MIABIS, EMIF, ENCePP and IMI-EHDEN (described above). Subsequently, we analysed how similarities could be harnessed towards a harmonised framework for both research cohort and RW data catalogues.

3.1 Concepts Commonly Collected

We analysed the user interfaces of existing catalogues and, where available, we also read their publications and associated literature, and downloaded their data models, i.e. definitions of tables/column/properties. We created a large spreadsheet where metadata items collected in the various catalogues were compared. For this review, we did not aim for a complete mapping of all metadata items collected, but instead aimed to assess commonalities and differences in terms of main topics, entities and features collected. This spreadsheet is included as Supplementary Table 1 (available at: https://docs.google.com/spreadsheets/d/1PvLpZgJ0jTwc9d8jxAt2Ex3nepUH3-Z4ImgH4-UUwnI/edit#gid=0). Table 1 below summarises the results.

Table 1 Overview of concepts collected. Detailed mapping on which catalogue is associated to each term can be found in Supplementary Table 1.

Terms	Concept
Concepts for collected data	
Data source, Data bank, Database, Study, Cohort, Registry, Collection	Unit of observation of the catalogue. Note that the term 'study' is used in other catalogues to describe a research activity using the data
Data source type, Data source category, Data bank family, Study design	Most catalogues contain a categorical classification of the type of data source, e.g., based on the design (cross-sectional, longitudinal, etc.) or on the type of data collection (population cohort, clinical cohort, biobank, registry)
Collected variable, Table column, Event	Observations collected, measured, or constructed within a data source
Collection event, Prompt	Event that triggers data collection. In cohort studies, these are typically planned, such as 'baseline' and 'first follow-up' while in RW, these are ad-hoc events such as 'admission in hospital' or 'birth'
Population, Study	Population whose data is collected in the data source
Access conditions, Data use conditions, Informed consent, Privacy, Data licence	Conditions in which data may be used, e.g., privacy, ethical/legal, or commercial considerations
Concepts for harmonisation/standardisation projects	
Common data model (CDM), DataSchema, Standard data dictionary	Data dictionary that is defined as part of an agreed-upon standard
DataSchema variable, Target variable, Harmonised Variable, Core variable, CDM variable, Common data element	Elements within a common data model. 'Standard' could be standard metadata (i.e., information to be used to define the variable; label, data source, categories, values, etc.) or can be the format of the variable (e.g. smoking status (0=non smoker, 1=smoker)).
Study variable, Derived variable, Computed variable, Measure, Phenotype	Data derived based on the collected data, on top of the standardisation/harmonization.
Research project, Harmonization study, Study	The term study is used for different purposes in different catalogues. In some cases, this describes the use of data in one or multiple data sources for a particular research project
Concepts used to support above concepts	
Institution, Organisation, Centre, Biobank, Data access partner, Data access partner	Organisation responsible for one or more data sources, or with other roles in studies (e.g., investigators)
Network, Consortium	Collaborations, denoted as relationships between data sources and/or institutions, for example between biobanks (e.g., BBMRI is a network) or in the context of a specific long-term project (e.g., ConcePTION is a consortium)
Areas of Information, Data types, Data categories, Keywords, Variable taxonomy	Categorical classifications used to classify variables collected in the data source and/or common data element/harmonized variables

3.2 Analysis Towards Common Conceptual Framework

In this section, we discuss three areas where we find correspondences between cataloguing cohort and RW data sources: the data collection itself (Table 2); the steps to address a research question in one single database (Table 3); the steps to enact interoperability across data sources to address a research question (Table 4).

From the point of view of a catalogue, the most important difference between research cohorts and RW data is the *unit of the data collection*: in the cohort domain, the unit is the cohort itself, which is also the data collection needed to conduct a study; in the RW domain, the unit is the data bank (e.g., registry, primary care records, pharmaceutical database), but a second-level unit, the data source, possibly including multiple data banks, is what is used to conduct a study [27]. For this reason, both the MINERVA and the ConcePTION catalogues have both levels of data collections separately described. Another important difference is the *population.* In the case of a cohort, the population is the set of persons whose data is collected, while the population of a data bank may not be included in the data bank itself: for example, in the case of a death registry, the population is not the deceased, but the persons whose death, were it to happen, would be recorded in the data bank. In general, the population is not the set of persons who have experienced the prompt, but the set of persons whose prompts, were it to happen, would prompt a record in the data bank.

In Table 3, five dimensions are listed. In a catalogue of cohorts, most of the information relative to a study can be described independently of the research question: with the exception of large general cohorts, the institution using the data is very close to the process of data collection, the research question belongs to a set of pre-specified questions incorporated in the design of the cohort, and both selection of the study population and study variables are embedded in the cohort data model. On the other hand, in the RW domain, the selection and study variables observed on the data source population to conduct a study are not observed *directly*, but rather are *derived* based on the collected data, a process referred to in the literature also as *measuring* [29] or *phenotyping* [30]: if a study must select persons with diabetes, an algorithm to identify diabetes from the data banks of each data source needs to be specified [31]. Since every study has different selection criteria and study variables (associated with the specific research question), to enumerate such information in a catalogue it is convenient to include study-specific sections. Indeed, the choice of phenotype may depend on the research question: for a surveillance study, a more sensitive phenotype is preferrable; to select a population with a disease, a more specific phenotype may be advisable. The MINERVA catalogue includes a section describing studies, also listing the study variables derived in each participating data source [26].

Table 4 describes the steps that need to be taken by a network of institutions accessing data collections to address together a same research question.

In the cohort domain, this typically involves a process where first a research consortium is formed, then variables for the research are defined, and subsequently all participating cohorts are asked to retrospectively harmonise their data onto these research variables. In the RW domain, the first step towards interoperability is the conversion to a common data model (CDM), such as the Sentinel CDM (https://www.sentinelinitiative.org/methods-data-tools/sentinel-common-data-model), the OMOP CDM [32], the ConcePTION CDM [27]. On top of this, the derivation of study variables must be enacted based on the CDM [33, 34], possibly using prompt-tailored algorithms [35, 27].

In summary, cohort catalogues can be nested in RW catalogues where:

- *Cohorts* are a special case of a *data source* with a single *data bank*, having a set of research questions embedded in the data bank design, and having as population the set of persons whose data are included in the cohort;
- *Collection events* can be considered a special case of *prompts*;
- *Harmonisation* of items is a special case of conversion to a *common data model*, since in RW additional information about the data sources must be stored, i.e., how to *derive* study variables to address research questions.

Table 2 Representation of data collections.

Concept	Domain	
	Research	RW data
Unit of data collection	Cohort, Cross sectional sample	Data bank
Data collection to conduct a study	Cohort	Data source
Purpose of the data collection	Research	Other purposes (e.g., routine healthcare)
Organisation sustaining the data collection	Organisation having research as a mission	Often an organisation without research in their mission (originator of the data bank)
Population	Conceptualised as the persons whose data is collected. It is a selected sample of the general population, and selection criteria are associated with a broad set of research questions	Conceptualised as the persons whose data would be collected, were the prompts to happen. Often a sample of a general population unrelated with research questions
Unit of observation within the data collection	Person (study subject)	An event prompting a record in the data bank, e.g. a hospitalisation, birth, medicine dispensing
Timing of data collection	Once, or longitudinal (collection events)	Whenever the prompt happens
Items collected	Associated with a broad set of research questions	Associated with the prompt
Codes used in categorical values	Often research-specific code systems from commonly used surveys or tests	Often international coding systems

Table 3 Concepts involved in the process of addressing a research question in one single data collection.

Concept	Domain	
	Cohorts	RW data
Actor conducting research	Research institution often close to the data collection	Research institution often lacking any control over data collection.
Actor populating the catalogue	Often one institution per cohort	Could be one or more institutions with expertise to conduct studies on the data
Research questions that can be addressed	Pre-specified	Independent of original reason for data collection
Selection of the study population to address a research question	Pre-specified (sub)population	Population selected using derived data
Variables entering the analysis to address a research question	Mostly: items observed	Items observed or derived algorithms

Table 4 Concepts involved to enact interoperability across multiple data collections to address a research question.

Concept	Domain	
	Cohorts	RW data
Actors conducting the research	Networks of institutions	Networks of institutions
Tool for interoperability across data collections	Harmonisation of items	Conversion to a common data model
Harmonized data collection	Harmonized cohort	Common data model
Actions to be enacted on top of data collection to obtain study variables	No further action is needed before application of study design	Data processing (measure/phenotype) to derive study variables on top of the common data model

4 Towards a Sustainable Catalogue Ecosystem

Creating high quality catalogues takes considerable time, energy, expertise, and motivation [36]. This section analyses requirements for catalogue collaboration in an open catalogue ecosystem.

Main actors in this endeavour are: the *data partners* - the individuals who collect and/or provide access to the human/health related data sources such as owners/hosts of cohorts, registries, biobanks; the *data consumers* - researchers, data scientists, funders of research, policy makers, etc. who aim to find, access and reuse these data sources; and *catalogue developers* - the individuals responsible for the development of the technical infrastructure of catalogues, specifically its metadata structure (i.e. metadata elements documented), and for the supply and curation of its contents, such as catalogue data managers/curators, scientific staff involved and software developers/administrators.

Based on the experience of the co-authors, we observe opposite forces to desire both central and distributed cataloguing efforts. Some would desire to have one single catalogue. The motivation from the data partners' perspective is that they ideally would catalogue their resources only once, instead of now receiving repeated cataloguing requests, and thus having to provide descriptions of their data sources many times. The motivation from data consumers is that they only need to search one catalogue, instead of finding and searching multiple catalogues. However, there are also many good reasons to have multiple catalogues developed. Depending on the research domain or the purpose of a catalogue, there can be quite different requirements on what and how data should be catalogued, i.e. the metadata items which should be described – and to what level of detail – depend on the use cases of the catalogue and the perspective of its curators. Moreover, these descriptions can be highly dependent on the context of use, i.e. often data sources are catalogued to serve particular types of studies or research questions, e.g. a catalogue that describes cohorts and harmonised variables for 'environmental factors influencing child development'.

Also, from a practical perspective, we observe benefits for distributing the cataloguing efforts. Cataloguing scientific resources currently requires a great effort from the data partners and typically a data management/curator team that assists data partners when cataloguing. We observe that such teams are highly dependent on domain knowledge and regional proximity to enable development of relationships with data partners and user communities. More importantly, we observe that cataloguing efforts for both data partners and the user community are more easily focused on a particular research area, as compared to a general all-purpose cataloguing effort. We speculate that it is much easier to catalogue information of interest that is required by a specific research group than for a diverse and more distant user community that is hard to engage. Finally, we have observed in other scientific databasing efforts that having one central effort creates potential single points of failure when funders and institutions priorities shift, while networks of such database efforts are easier to sustain [http://www.insdc.org/]. Meanwhile,

small catalogues developed for a particular research project are also hard to sustain over the long term, potentially wasting cataloguing efforts.

Moving forward, we believe we should thankfully embrace existing data curation in distributed cataloguing efforts. However, we also believe it timely to promote the sharing and reuse of catalogue contents to enable data consumers to navigate the catalogue contents more easily, enabling integrated or cross catalogue searches, reduce duplicated efforts, improve quality and completeness and ease sustainability of cataloguing contents in case the original catalogue hosting is discontinued. We identify the following requirements for successful sharing and reuse of catalogue contents:

- Ensure autonomy of metadata collection;
- Have clear licensing conditions in place for catalogue contents;
- Enable symmetric data sharing, syndication and cross indexing;
- Add provenance and attribution records (metadata on the metadata).

4.1 Autonomy of Metadata Collectors

In the next section, we will recommend standardisation and harmonisation at many levels. However, while centralisation and standardisation have many advantages, research is developing quickly, creating new research applications for existing and newly collected data. Consequently, the metadata that are desirable are different for different research use cases. Therefore, we recommend as a principle to maintain the full autonomy of catalogue developers to expand and refine the catalogue structure (i.e., metadata elements collected) and contents. Standardisation for interoperability is not a goal in itself but serves a purpose. We recommend the use of micro standards, i.e., instead of enforcing catalogue developers to immediately implement 100+ metadata items, allow for a modular approach such that catalogues can keep their autonomy to adhere to relevant standard elements where useful but without large standardisation requirements creating a barrier to connecting a catalogue.

4.2 Clear Data Licence

Without a licence that explicitly allows metadata sharing across catalogues, copyright laws and database laws will prohibit it and it might be even unclear who would be the legal entity to approach for permission. General copyright law will reserve all rights to the creator, which depending on the local legislation, might be the creator of the catalogue or the original content provider whose data is being catalogued or a combination thereof. Similarly, database laws will reserve rights to a structured database to the creator of that database. A licence should clarify how to handle data sharing and contributions. Aspects such as whether records are allowed to be copied and modified and to what extent changes should be given back, should be covered, as well as whether catalogue contents can be included in other catalogues or can be used in commercial software applications, and, importantly, whether references or citations must be made to the data partner and/or catalogue provider. Even though it is definitely possible to draft your own licence, the OpenAIRE Guide "How do I licence my research data" [https://www.openaire.eu/how-do-i-license-my-research-data] advises to use the CC BY 4.0 licence [https://creativecommons.org/licenses/by/4.0/] for works that fall under the definition of a creative work under copyright law and to use CC0 licence [https://creativecommons.org/publicdomain/zero/1.0/] for databases or datasets. Using these licences will ensure compatibility with the definition of Open Access data.

4.3 Symmetric Data Sharing and/or Aggregation

Catalogues aim to improve findability and reuse, and therefore a key performance indicator is to attract as many (relevant) users as possible, and ideally to convert these users into successful research projects. For this, indexing into general search engines such as Google is the primary aim. The aggregation of their catalogue into more general catalogues, i.e. to have a partial copy of the catalogue record with a link to the original catalogue so that it can also be found in other catalogue instances can also aid this goal. However, catalogues also have the existential challenge that in order to keep securing their funding they must be recognizable/visible. There can also be a concern that copies of the metadata are added, changed or removed in ways not approved by the original catalogue developer, raising concerns about quality and/or completeness of the records. Moreover, there is the concern that a catalogue record is fully copied and/or no link back to their catalogue is provided, thus taking away users from the original catalogue. All these trust issues can make catalogues hesitate to allow cross-catalogue indexing.

To overcome these issues, we recommend that the sharing of catalogue contents should always be reciprocal. Potential ways could be by synchronising contents between two similar catalogues, or by aggregating metadata into an 'umbrella' catalogue (for example, a catalogue for 'all child cohorts'). This kind of 'cross indexing' would then enable the source catalogue to show aggregated catalogue contents from other sources and thus also provide a limited form of the integrated search experience instead of fearing loss of their users to the aggregated catalogue. Obviously, links back to the aggregator must then be shown so that the aggregator is also rewarded for its efforts, with the potential for more users and greater impact. An example of this is being developed in EUCAN-connect, where cross-search between the Maelstrom, LifeCycle, Birth cohorts and RECAP Preterm catalogues is being enabled [https://catalogue.eucanconnect.eu/]. Another example of this is being piloted in the European Joint Programme for Rare Disease [https://github.com/ejp-rd-vp], where a 'search widget' has been developed to enable database users to do 'live' searches that also include the other databases in their network, via a single user interface.

4.4 Clear Procedures for Changes and Updates

Finally, when catalogue records are indeed exchanged and there is clarity on the licence, there should also be clarity on how changes to records should be processed. Ideally, it should be clear which metadata items are expected to be universally the same and which items can be different when cataloguing records

in multiple catalogues. We would expect that administrative information, e.g., name, host institute and certain contact details, would be independent from the purpose of the catalogue. Meanwhile, study-specific or application domain-specific details, such as inclusion criteria, data collected and design of the study, could vary. Ideally, these context-specific metadata items should also be marked as such.

When additional details and improved descriptions have a universal added value, one would want to merge these improvements with the original record. Therefore, each catalogue should also provide a clear procedure for submitting and processing requests for updates of catalogue records, and for downstream users of the catalogue to then retrieve such updates. This procedure should minimally include technical submission procedures (webform and/or programmatic) and an indication of the last changed date. Ideally, it should also include review/acceptance workflow (e.g., involving the original author of the record, often associated with the data source being catalogued), versioning, tracking of changes, and clarity of ways to attribute/name the individuals involved in the cataloguing effort who should be acknowledged.

5 Interoperability Recommendations

To implement interoperability across catalogues, technical standardisation is also desirable. Practically, while catalogues are usually proposed as a method to implement FAIR principles for datasets, the contents of the catalogue itself must also adhere to the FAIR principles [6]. In the past decade, converging recommendations have emerged to improve metadata catalogues [FAIRsFAIR on metadata catalogue interoperability, https://zenodo.org/record/5744913#.YeEB3VjMLzc]. Also, recent projects in the context of the European Open Science Cloud (EOSC), the European Health Data Space (EHDS), and Research Data Alliance (RDA) produce relevant recommendations, such as via the TEHDAS (towards European health data space) project [https://tehdas.eu] and the SHAring Rewards and Credit Interest Group [36] with practical recommendations, FAIR assessment templates, and lessons learned such that FAIR criteria should be implemented gradually, and that training, support and rewards should be considered. Based on these elements and the practical experience from the co-authors, the following main requirements emerge:

- Persistent identifiers for cohorts, data banks, data sources, institutions, contributors;
- Use of common data elements and ontologies to identify content overlap between catalogues;
- Easy to use syntaxes and architecture to pull/push catalogue contents;
- Addition of provenance records to track how catalogue contents have been copied and changed.

5.1 Persistent Identifiers

When data records are copied across catalogues, it is essential that data curators can check where the original records reside. In addition, the same resource (cohort, data bank) might be catalogued multiple times with different names/acronyms. Persistent identifiers such as the DOIs or Handle Identifiers could be used to unambiguously identify all these different types of records. These systems provide a separation between the identifier and the location of the (digital) object that they refer to and have agreements to maintain the service in the long term.

The identifiers should be opaque and unique, where the structure and the content of the identifier is devoid of any significance about the object that they describe. This allows the identifier to stay the same, while the object that it describes can change. For instance, if we use the name of a data bank as an identifier, this causes problems when the data bank changes its name and website or if another organization creates a new data bank with the same name. Any catalogue that uses the identifier can use it to retrieve the current name and link to the website and maintain all references to the data bank. At the same time, a catalogue system can easily tell the two data banks apart based on their different identifiers.

There are many persistent identifier services available starting for a few hundred dollars a year, however the costs are in identity maintenance. We envision that large organisations, such as the European Medicine Agency via their ENCePP system [https://www.encepp.eu/] and BBMRI-ERIC via their biobank directory [http://directory.bbmri-eric.eu], would be willing to assign and maintain persistent identifiers for a large community of catalogue maintainers to use.

5.2 Common Data Elements and Ontologies

As described above, we can easily observe overlap between the catalogues. Also, we believe that a combination of Maelstrom, LifeCycle, BBMRI-ERIC and MINERVA provides the coverage that can provide a basis to standardise these concepts. However, it would be very labour-intensive to search, compare and combine the contents from multiple catalogues manually today. Standardisation in using the same metadata elements and code systems/ontologies where possible can greatly alleviate this challenge. And ideally, metadata items should be coded inputs, using values from an ontology, instead of free-text that allows non-standard metadata values (although in our experience, there is added value for free-text in descriptive notes, and to chart candidate codes before a code system has been established).

In recent years, general agnostic catalogue standards have emerged, but they lack details about the domain (in our case, cohorts, RW data). Recommended standards include DCAT [Data Catalogue - https://www.w3.org/TR/vocab-dcat-2/], and a specific implementation thereof called FDP [FAIR data point, https://www.fairdata-point.org/]. DCAT defines: (1) the 'catalog' which is a dataset in which each individual item is a metadata record describing some resource; (2) the 'resource' which represents a dataset, a data service or any other resource; (3) the 'dataset', a collection of data published or curated by a single agent; (4) the 'distribution' and 'data service' that provide e.g. download file and/or operations via programming interfaces, and (5) the 'catalogue record' primarily concerning the

registration information, such as who added the item and when. For example, Maelstrom could use DCAT to describe each cohort and network as a 'resource' (including details such as the id, title, landing page, issuer) and use 'datasets' to detail datasets within the cohorts (including details such as the title, keywords, contact point) including links how to access additional information such as lists of variables (as 'distribution'). In addition, we want to mention BioSchemas [https://bioschemas.org/] [40], an extension to schema.org used by Google and other search engines that provides concepts such as DataCatalog with properties such as description, keywords, provider, citation, licence, variableMeasured.

However, to be of real value more specific details would be needed as described in section 3. Therefore, ideally the catalogue community would come together to define common data elements and code systems for concepts. For example, we know of multiple consortia working on harmonisation of catalogue metadata (e.g., the EUCAN-connect project tries to harmonise Maelstrom and LifeCycle catalogues, in collaboration with the CINECA project). However, these efforts are not linked to existing standards. Instead of creating these common data elements in isolation, it is recommended to reuse existing metadata item definitions. A good starting point is the standard website FAIR sharing [37] and ontology publishing websites like BioPortal [https://bioportal.bioontology.org/] and Ontology Lookup Service [OLS, http://www.ebi.ac.uk/ols]. Ideally, the catalogue community will go through a rigorous ontology process to define all metadata types and instances while reusing existing ontologies, for example following examples from other domains [38], towards an ontology with high standards for quality control, for example using the standards of OBO foundry [39].

5.3 Syntaxes and Architecture for Metadata Interoperability

Standardisation of catalogue contents using common data elements is not sufficient to make data flow. In addition, the standardisation of file formats would be recommended to ease the download/upload of data between catalogues, and application programming interfaces (APIs) to automate such interactions using scripts and programs.

The gold standard for FAIR data is linked data or 'semantic web'. This for example includes RDF [resource description format, https://www.w3.org/RDF/], a data model and framework that can be represented in a file format similar to HTML files but with a specific structure to describe all data as 'triples' of {subject, predicate, object} using hyperlinks/URIs to denote cross references. However, in our experience, most catalogue developers have very limited information technology (IT) capacity, and therefore prefer using CSV files, Excel spreadsheets or JSON text files. Notably, JSON has a linked data extension called JSON-LD, which would still enable integration with the aforementioned linked data communities where desired.

In addition, we must consider the architectures for data exchange. In recent years, we have implemented both push interfaces – where data submitters would push metadata into the catalogue – and pull interfaces – where the catalogue developer would retrieve data from a web location. Push interfaces are typically used for submitting/updating catalogue records (e.g., from individual cohort into LifeCycle catalogue) while pull interfaces are more often used for moving data between catalogues (e.g. from the Dutch national catalogue, https://catalogue.bbmri.nl, to the BBMRI-ERIC EU biobank catalogue, https://directory.bbmri-eric.eu/). In most cases, simple interfaces using the HTTP protocol were preferred, using operations such as GET, POST, PUT/PATCH, DELETE to retrieve, add, update or remove contents respectively, ideally enabling 'batch' updates, i.e. affecting multiple catalogue records in one transaction.

With this in mind, we recommend embracing current practises. Therefore, first standardise on very simple CSV/JSON file formats for metadata exchange, and minimal set of 'batch' programmatic interfaces to retrieve/submit metadata in this file format. Only when desired by multiple catalogue developers, advanced query interfaces with proper (semantic) API documentation could be defined as standard to enable 'federated queries' (i.e., so that users could search multiple catalogues at once), for which an interesting emerging standard 'beacon' is being developed in the global alliance for genomics and health [GA4GH, https://beacon-project.io/].

5.4 Add Provenance and Attribution Records

Finally, it becomes essential that catalogues also start to describe and show the provenance of their records, i.e. where the original record originated, whether intermediate changes were made to the record, and what the procedure should be if one would find that the record would need to be updated. Examples for which information is necessary have been extensively developed within the scientific library community, notably in the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) and OpenAIRE project [http://www.openarchives.org/OAI/openarchivesprotocol.html]. For each copy made, a provenance record should be added that contains for example [https://openaire-guidelines-for-literature-repository-managers.readthedocs.io/en/v4.0.0/use_of_oai_pmh.html]: baseURL, identifier, datestamp, metadataNamespace, origin, and for each description, harvest date and whether data was altered. In addition, there should be information that explains how changes to the record should be submitted. For example, there might be a link to a website, programmer interface or simply email address to share updates.

6 Conclusion

We have summarised the landscape of data catalogues cataloguing, on the one hand, cohorts and on the other, RW data sources composed of multiple data banks, such as surveillance registries or administrative records.

While these catalogues are heterogeneous, of various scope, and use different terminologies, the underlying concepts seem potentially harmonizable. Our conclusion is that cohort catalogues can be nested in RW catalogues where: (1) cohorts are a special case of a data source with a single data bank,

having a set of research questions embedded in the data bank design, and having as population the set of persons whose data are included in the cohort, (2) collection events for cohorts are a special case of prompt for RW data, and (3) harmonisation of items is a special case of the conversion to a common data model, because in RW data sources additional information must be stored, that is, how to derive study variables to address research questions.

We recommend embracing the autonomy of current catalogue teams, and invest in their collaboration via minimal standardisation efforts such as clear data licensing, minimal metadata 'common data elements', symmetric architectures for data sharing (push/pull) with clear provenance tracks to process updates and acknowledge original contributors. Obviously, the implementation of such an ecosystem is a massive effort. Fortunately, many FAIRification efforts exist such as the research data alliance (RDA), ELIXIR and BBMRI. However, we still observe a gap between the domain-specific cataloguing groups which typically have vast domain knowledge. These gaps can be bridged by domain-specific efforts such as EU EUCAN-connect [http://www.eucan-connect.org], the metadata working group of European Human Exposome Network [https://ehp.niehs.nih.gov/doi/abs/10.1289/isee.2021.O-SY-127] and recently funded EU projects BeYond COVID [BY-COVID, https://by-covid.org/], IMI European Partnership for neurodegenerative diseases such as Alzheimer's and Parkinson's EPND [https://www.imi.europa.eu/projects-results/project-factsheets/epnd].

Even while many collaborations exist between these initiatives, these efforts can lead to silo-ed solutions if we do not invest in cross-domain and cross-consortium interactions. We therefore recommend creating practical training, documentation at entry level and intermediate level to ensure new catalogue efforts can converge on existing standards including: (a) common data elements for catalogue developers to pick from, (b) exchange formats and APIs that are easiest to implement, and (c) free-to-use systems for requesting persistent identifiers. Most importantly, environments for collaboration and resource sharing between catalogue developers should be facilitated.

Key messages:

- Catalogues are heterogeneous, of various scope, and use different terminologies, however the underlying concepts seem potentially harmonizable;
- The autonomy of current catalogue teams is valuable and cross-collaboration should be sustained via minimal standardisation efforts such as clear data licensing, minimal metadata 'common data elements', symmetric architectures for data sharing (push/pull) with clear provenance tracks;
- We recommend creating practical training, documentation at entry level and intermediate level to ensure new catalogue efforts can converge on existing standards.

Acknowledgements

We acknowledge contributions from BBMRI-ERIC (common services for IT), BBMRI-NL (NWO 184.021.007) European Commission (H2020, FP7), Innovative Medicine Initiative (IMI), European Medicine Agency and national scientific organisations (in particular Canadian Institutes of Health Research, CIHR) to this work, in particular collaborations with EUCAN-connect (H2020 and CIHR grant 824989), ConcePTION (IMI 821520), MINERVA, ReaCH (CIHR OCR-144561), LifeCycle (H2020 grant 733206), RECAP Preterm (H2020 grant 733280), LongITools (H2020 grant 874739), ATHLETE (H2020 874583), BioSHaRE (FP7 HEALTH 261433), CINECA (H2020 grant 825775), EOSC-Life (H2020 824087). We thank Susana Perez-Gutthann for useful discussions.

References

1. Daniel C, Kalra D; Section Editors for the IMIA Yearbook Section on Clinical Research Informatics. Clinical Research Informatics. Yearb Med Inform 2020 Aug;29(1):203-7.
2. Safran C. Update on Data Reuse in Health Care. Yearb Med Inform 2017 Aug;26(1):24-7.
3. Schlegel DR, Ficheur G. Secondary Use of Patient Data: Review of the Literature Published in 2016. Yearb Med Inform 2017 Aug;26(1):68-71.
4. Meystre SM, Lovis C, Bürkle T, Tognola G, Budrionis A, Lehmann CU. Clinical Data Reuse or Secondary Use: Current Status and Potential Future Progress. Yearb Med Inform 2017 Aug;26(1):38-52.
5. Schriml LM, Chuvochina M, Davies N, Eloe-Fadrosh EA, Finn RD, Hugenholtz P, et al. COVID-19 pandemic reveals the peril of ignoring metadata standards. Sci Data 2020 Jun 19;7(1):188.
6. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 2016 Mar 15;3:160018.
7. Fortier I, Raina P, Van den Heuvel ER, Griffith LE, Craig C, Saliba M, et al. Maelstrom Research guidelines for rigorous retrospective data harmonization. Int J Epidemiol 2017 Feb 1;46(1):103-5.
8. Hutchinson DM, Silins E, Mattick RP, Patton GC, Fergusson DM, Hayatbakhsh R, et al; Cannabis Cohorts Research Consortium. How can data harmonisation benefit mental health research? An example of The Cannabis Cohorts Research Consortium. Aust N Z J Psychiatry 2015 Apr;49(4):317-23.
9. Holub P, Swertz M, Reihs R, van Enckevort D, Müller H, Litton JE. BBMRI-ERIC Directory: 515 Biobanks with Over 60 Million Biological Samples. Biopreserv Biobank 2016 Dec;14(6):559-62.
10. Merino-Martinez R, Norlin L, van Enckevort D, Anton G, Schuffenhauer S, Silander K, et al. Toward Global Biobank Integration by Implementation of the Minimum Information About BIobank Data Sharing (MIABIS 2.0 Core). Biopreserv Biobank 2016 Aug;14(4):298-306.
11. Eklund N, Andrianarisoa NH, van Enckevort E, Anton G, Debucquoy A, Müller H, Zaharenko L, et al. Extending the Minimum Information About BIobank Data Sharing Terminology to Describe Samples, Sample Donors, and Events. Biopreserv Biobank 2020 Jun;18(3):155-64.
12. Bhattacharya S, Dunn P, Thomas CG, Smith B, Schaefer H, Chen J, et al. ImmPort, toward repurposing of open access immunological assay data for translational and clinical research. Sci Data 2018 Feb 27;5:180015.
13. Bergeron J, Doiron D, Marcon Y, Ferretti V, Fortier I. Fostering population-based cohort data discovery: The Maelstrom Research cataloguing toolkit. PLoS One 2018 Jul 24;13(7):e0200926.
14. Doiron D, Marcon Y, Fortier I, Burton P, Ferretti V. Software Application Profile: Opal and Mica: open-source software solutions for epidemiological data management, harmonization and dissemination. Int J Epidemiol 2017 Oct 1;46(5):1372-8.
15. van Vliet-Ostaptchouk JV, Nuotio ML, Slagter SN, Doiron D, Fischer K, Foco L, et al. The prevalence of metabolic syndrome and metabolically healthy obesity in Europe: a collaborative analysis of ten large cohort studies. BMC Endocr Disord 2014 Feb 1;14:9.
16. Doiron D, Burton P, Marcon Y, Gaye A, Wolffenbuttel BHR, Perola M, et al. Data harmonization and federated analysis of population-based studies: the BioSHaRE project. Emerg Themes Epidemiol 2013 Nov 21;10(1):12.
17. Pinot de Moira A, Haakma S, Strandberg-Larsen K, van Enckevort E, Kooijman M, Cadman T, et al; LifeCycle Project Group. The EU Child Cohort Network's core data: establishing a set of findable, accessible, interoperable and re-usable (FAIR) variables. Eur J Epidemiol 2021 May;36(5):565-80.

18. Ronkainen J, Nedelec R, Atehortua A, Balkhiyarova Z, Cascarano A, Ngoc Dang V, et al. LongITools: Dynamic longitudinal exposome trajectories in cardiovascular and metabolic noncommunicable diseases. Environ Epidemiol 2021 Dec 28;6(1):e184.
19. Vrijheid M, Basagaña X, Gonzalez JR, Jaddoe VWV, Jensen G, Keun HC, et al. Advancing tools for human early lifecourse exposome research and translation (ATHLETE): Project overview. Environ Epidemiol 2021 Oct 1;5(5):e166.
20. van der Velde KJ, Imhann F, Charbon B, Pang C, van Enckevort D, Slofstra M, et al. MOLGENIS research: advanced bioinformatics data software for non-bioinformaticians. Bioinformatics 2019 Mar 15;35(6):1076-8.
21. Bamber D, Collins HE, Powell C, Gonçalves GC, Johnson S, Manktelow B, et al. Development of a data classification system for preterm birth cohort studies: the RECAP Preterm project. BMC Med Res Methodol 2022 Jan 7;22(1):8.
22. Final set of metadata and definitions, process, and catalogue tool. https://www.encepp.eu/encepp/openAttachment/documentsLatest.otherDocument-0/43021. [cited 2022 February 4]
23. https://www.encepp.eu/encepp/viewResource.htm?id=39323. [cited 2022 February 4]
24. IMI EHDEN. D4.7 Yearly Progress Report on Technical Framework. 16 December 2020. Available from: https://ec.europa.eu/research/participants/documents/downloadPublic?documentIds=080166e5d7268def&appId=PPGMS
25. Oliveira JL, Trifan A, Bastião Silva LA. EMIF Catalogue: A collaborative platform for sharing and reusing biomedical data. Int J Med Inform 2019 Jun;126:35-45.
26. Final Good Practice Guide for Metadata Collection for Real-World Data Sources. https://www.encepp.eu/encepp/openAttachment/studyResult/45243
27. 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 2022 Jan;111(1):321-31.
28. Swertz M, Hyde E, Cunnington M, Gini R. 2021. Test report for FAIR data catalogue (1st) (D7.9). Zenodo. https://doi.org/10.5281/zenodo.5829454
29. Schneeweiss S, Patorno E. Conducting Real-world Evidence Studies on the Clinical Outcomes of Diabetes Treatments. Endocr Rev 2021 Sep 28;42(5):658-90. Erratum in: Endocr Rev. 2021 Nov 16;42(6):873.
30. Conway M, Berg RL, Carrell D, Denny JC, Kho AN, Kullo IJ, et al. Analyzing the heterogeneity and complexity of Electronic Health Record oriented phenotyping algorithms. AMIA Annu Symp Proc 2011;2011:274-83.
31. Roberto G, Leal I, Sattar N, Loomis AK, Avillach P, Egger P, et al. Identifying Cases of Type 2 Diabetes in Heterogeneous Data Sources: Strategy from the EMIF Project. PLoS One 2016 Aug 31;11(8):e0160648.
32. OHDSI Community. The OMOP CDM. In: OHDSI Book. https://ohdsi.github.io/TheBookOfOhdsi/CommonDataModel.html [cited 2022 February 4].
33. Gini R, Schuemie M, Brown J, Ryan P, Vacchi E, Coppola M, 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.
34. Swerdel JN, Hripcsak G, Ryan PB. PheValuator: Development and evaluation of a phenotype algorithm evaluator. J Biomed Inform 2019 Sep 1;97:103258.
35. Gini R, Sturkenboom MCJ, Sultana J, Cave A, Landi A, Pacurariu A, et al; Working Group 3 of ENCePP (Inventory of EU data sources and methodological approaches for multisource studies). Different Strategies to Execute Multi-Database Studies for Medicines Surveillance in Real-World Setting: A Reflection on the European Model. Clin Pharmacol Ther 2020 Aug;108(2):228-35.
36. David R, Mabile L, Specht A, Stryeck A, Thomsen M, Yahia M, et al. The Research Data Alliance – SHAring Reward and Credit (SHARC) Interest Group. FAIRness Literacy: The Achilles' Heel of Applying FAIR Principles. CODATA Data Science Journal 2020;19(32):1-11.
37. Sansone SA, McQuilton P, Rocca-Serra P, Gonzalez-Beltran A, Izzo M, Lister AL, et al; FAIRsharing Community. FAIRsharing as a community approach to standards, repositories and policies. Nat Biotechnol 2019 Apr;37(4):358-67.
38. Jonquet C, Coulet A, Dutta B, Emonet V. Harnessing the Power of Unified Metadata in an Ontology Repository: The Case of AgroPortal. J Data Semant 2018;(7):191–221.
39. Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol 2007 Nov;25(11):1251-5.
40. Gray AJ, Goble C, Jiménez RC. The Bioschemas Community Bioschemas: from potato salad to protein annotation. International Semantic Web Conference; Berlin. 2017. Available from: http://ceur-ws.org/Vol-1963/paper579.pdf [cited 2022 Jan 23]

Correspondence to:
Rosa Gini
Via Dazzi 1
55141 Florence
Italy
E-mail: rosa.gini@ars.toscana.it

Prof Morris Swertz
Department of Genetics, HPC CB50
University Medical Center Groningen
P.O. Box 30001
9700 RB Groningen
The Netherlands
E-mail: m.a.swertz@rug.nl