WorldFAIR

Project Title	Global cooperation on FAIR data policy and practice
Project Acronym	WorldFAIR
Grant Agreement No	101058393
Instrument	HORIZON-WIDERA-2021-ERA-01
Topic, type of action	HORIZON-WIDERA-2021-ERA-01-41 HORIZON Coordination and Support Actions
Start Date of Project	2022-06-01
Duration of Project	24 months
Project Website	http://worldfair-project.eu

D7.1 Implementation Guidelines for Annotating Health Research Data and Harmonisation: The INSPIRE Approach

Work Package	WP07 – Population Health
Lead Author (Org)	Arofan Gregory (CODATA)
Contributing Author(s) (Org)	David Amadi (LSHTM), Jay Greenfield (CODATA), Sylvia Muyingo (APHRC), Jim Todd (LSHTM), Keith Tomlin (LSHTM)
Due Date	31.05.2023

Date	25.05.2023
Version	1.0 DRAFT NOT YET APPROVED BY THE EUROPEAN COMMISSION
DOI	https://doi.org/10.5281/zenodo.7887385

Dissemination Level

- [X] PU: Public
- [ ] PP: Restricted to other programme participants (including the Commission)
- [ ] RE: Restricted to a group specified by the consortium (including the Commission)
- [ ] CO: Confidential, only for members of the consortium (including the Commission)

Versioning and contribution history

Version	Date	Authors	Notes
0.9	30.04.2023	David Amadi, Jay Greenfield, Arofan Gregory, Sylvia Muyingo, Jim Todd, Keith Tomlin	Draft for internal review
1.0	20.05.2023	Iseult Lynch, Ian Bruno	Internal Review
1.9	25.05.2023	David Amadi, Ian Bruno, Jay Greenfield, Arofan Gregory, Iseult Lynch, Laura Molloy, Sylvia Muyingo, Jim Todd, Keith Tomlin	Final document for proof reading
2.0	xx.05.2023	Arofan Gregory	Final document submitted

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

Disclaimer

WorldFAIR has received funding from the European Commission's WIDERA coordination and support programme under the Grant Agreement no. 101058393. The content of this document does not represent the opinion of the European Commission, and the European Commission is not responsible for any use that might be made of such content.

As LSHTM is a UK institution, the financial support for LSHTM's contribution to this work package comes from the UK Research and Innovation (UKRI) under the guarantee given by the UK government (Grant # IFS 10040401).

Abbreviations and Acronyms

ALPHA	Analysing Longitudinal Population-based HIV data from Africa
APHRC	African Population Health Research Centre
CBS	Case-based surveillance
CDIF	Cross Domain Interoperability Framework
CDM	Common Data Model
CODATA	Committee on Data (International Science Council)
CORDIS	Community Research and Development Information Service
CRF	Case-based (Clinical) Record Form
CSV	Comma Separated Values
DDI	Data Documentation Initiative
EHDEN	European Health Data Evidence Network
EOSC	European Open Science Cloud
ETL	Extraction, Transform, Load
FAIR	Findable, Accessible, Interoperable, Reusable
FIP	FAIR Implementation Profile
HDSS	Health and Demographic Surveillance System
HEIs	Higher Education Institutions
IDSR	Infectious Disease Surveillance and Response

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

INSPIRE	Implementation Network for Sharing Population Information with Research Entities
JSON-LD	Javascript Object Notation – Linked Data
LMIC	Low and Middle Income Countries
LSHTM	London School of Hygiene and Tropical Medicine
MOH	Ministries of Health
OMOP	Observational Medical Outcomes Partnership
OHDSI	Observational Health Data Sciences and Informatics
SDG	Sustainable Development Goals
SDMX	Statistical Data and Metadata Exchange
SQL	Structured Query Language
SSA	Sub-Saharan Africa
UN	United Nations
WHO	World Health Organisation

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

Executive Summary

One of the key requirements for FAIR data reuse is that the user of a FAIR data resource understands the exact nature of the data. The FAIR principles talk about the kinds of metadata needed to describe data, but it is necessary for implementers to understand how these metadata can be provided, to effectively realise FAIR within their systems. This implementation guide describes the way all aspects of the data are made available for use, both within and from outside the INSPIRE Network[1] community, using standard metadata to describe the data. This is an exploration of how generic standards can be used to express the agreed community metadata set. The INSPIRE platform supports network studies using population health data to stand up their own instances of a common data model called the OMOP CDM. The WorldFAIR project is an exploration to facilitate a better understanding of what is needed for data infrastructures to provide data in line with the FAIR principles within and across domains.

The types of metadata used in INSPIRE are aligned as much as possible with existing and popular models common in the public health domain. Primary among these are the standards (and tools) coming from OHDSI (Observational Health Data Sciences and Informatics)[2], notably their OMOP Common Data Model (CDM). This suite of products addresses the definition of specific concepts and their semantics, standard (primarily medical) classifications, and the mechanism for selecting data from among those available to produce a specific cohort for analysis. These standards are common within the public health domain internationally, and INSPIRE has chosen to use them to reduce the significant cost of developing tools for many aspects of data and metadata management and use.

FAIR demands that we provide data in a useful way to those who may not be familiar with the community tools and standards used by INSPIRE. More generic standards are thus needed to support this broader community. It is significant that members of the OHDSI community have already looked at how Schema.org - developed and supported by many popular search engines, Google foremost among them - can be used in combination with the OHDSI OMOP CDM to describe data resources. Here, INSPIRE builds on that work to describe how INSPIRE data resources, specifically, can be documented in a way which will be maximally accessible to users both within the community and external to it.

One critical part of the overall information set provided by standard FAIR metadata is a description of the experiment for which the data was used, and the protocol employed in the selection and analysis of the data. This aspect of the metadata description is a major focus of the implementation guide, and one for which Schema.org would seem to be well-suited.

WorldFAIR WP (Work Package) 07 is one of eleven domain-specific case studies being undertaken by the WorldFAIR project, with the domain-specific practices being analysed across these domains in WP02. Early indications from WP02 suggest that Schema.org is one of the standards which will be recommended as part of the Cross-Domain Interoperability Framework (CDIF). This implementation

[1] https://aphrc.org/inspire/
[2] https://www.ohdsi.org/

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

guide contributes to an understanding of exactly how Schema.org fits into the description of domain data.

While some open questions remain, the implementation guide has achieved its primary goal of showing how standards such as Schema.org can be used within the public health domain to provide a complete set of the information needed for FAIR data use across and within domain boundaries.

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

Table of contents

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

1. Introduction and scope

1.1 Background: INSPIRE and WorldFAIR

1.1.1 Organisational Context

The London School of Hygiene and Tropical Medical (LSHTM) leads WorldFAIR WP07, with support from CODATA. LSHTM is a world-leading centre for research and postgraduate education in public and global health. LSHTM runs population health studies in many countries throughout sub-Saharan Africa (SSA) and other low- and middle-income countries (LMIC). These activities put LSHTM in an ideal position to explore ways to align standards for sharing population health data, to support analyses, and to enable the production of FAIR population health data to benefit researchers and policy makers. This WorldFAIR case study (project) is nested under the INSPIRE network, which is a collaborative network for shared data in SSA[3].

As LSHTM is a UK institution, the financial support for the LSHTM effort for this work package comes from the UK Research and Innovation (UKRI) under the guarantee given by the UK government (Grant # IFS 10040401).

The work of WP07 falls into three categories. Firstly, we attend meetings to explore good practice. Secondly, we define FAIR Implementation Profiles and FAIR enabling resources which are used in population health. Thirdly, we develop and document the implementation guides which are being used to make specific population health data available to the international community. This third effort provides the basis for this implementation guide.

1.1.2 Meetings to explore good practice

We attended the workshop, 'Interoperability for Cross-Domain Research: Machine-Actionability & Scalability' at Dagstuhl Leibnitz Centre for Informatics in August 2022 where the concept of FAIR data in population health was compared to approaches and considerations in other domains. Further details of this event can be found at https://codata.org/initiatives/decadal-programme2/dagstuhl-workshops/dagstuhl-workshop-2022/

The WorldFAIR consortium also organised high-level meetings where LSHTM learned how to align the metadata from our studies with international norms and standards. We also used monthly progress meetings with INSPIRE to enable data collection institutions to see how they can align their data collection and use the same norms and standards. In February 2023 the INSPIRE AGM identified ways to devolve and support FAIR data in member institutions, building on the work from the WorldFAIR project.

1.1.3 Define FAIR Implementation Profiles and FAIR enabling resources

Several months (July to October 2022) were spent exploring FAIR Implementation Profiles (FIPs) and FAIR Enabling Resources (FERs) that can be used with INSPIRE data. In particular we identified use

[3] https://aphrc.org/inspire/

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

cases which demonstrate FIPs in use with the OHDSI common data model (CDM). The OHDSI CDM uses standard ontologies and vocabularies which can be reused in many settings, but these are mainly focused on clinical and hospital records. We are working to produce new vocabularies for population health data and to describe public health interventions which can be added to the standard vocabularies in OHDSI.

The OHDSI CDM is a common data model developed by the Observational Medical Outcomes Partnership (OMOP), which is gaining traction in the area of health and medical research and was selected by INSPIRE because it represents those aspects of observational data which are also meaningful in population health. It is becoming increasingly popular within the domain. In essence, it establishes a standard set of tables, populated with concepts from well-recognised classifications, which can be used as a reference point for sharing data in the form of standard observational databases. The OHDSI CDM standard is organised around the observation of persons, making it suitable for longitudinal population data even if this was not the original anticipated use, although this required some modification in the INSPIRE CDM (the implementation of the OHDSI CDM provided by the INSPIRE network).

The next step is to produce implementation guides to describe the data and metadata available in our INSPIRE implementation of the OHDSI CDM. These implementation guides can be found by others using the same data structures, and can be adapted and reused for other disease-specific data. The implementation guides are used to make the data Findable, Accessible, Interoperable and Reusable (FAIR) using standard resources. In the future we will add the standard analyses used in the INSPIRE CDM to the FAIR resources which would have machine-actionable code. All such FAIR resources could be exposed via the INSPIRE platform and further registered or catalogued as appropriate, in line with more general FAIR practice as it emerges from WorldFAIR and other initiatives.

1.2 The INSPIRE platform

The INSPIRE platform is designed to support network studies in which Health and Demographic Surveillance System (HDSS) sites and/or Ministries of Health (MoHs) set up their own instances of the OHDSI CDM and their own instance of an analysis workbench that runs on top of the OHDSI CDM called ATLAS (of which there is an online demo at https://atlas-demo.ohdsi.org/#/home). Depending upon needs and security requirements these databases and their analysis workbenches are deployed in a public cloud, a private cloud or locally using a turnkey deployment solution. Source microdata (i.e. person-level data, often responses to questionnaires, held in the systems of those who collected it from respondents) is extracted, transformed and loaded (ETL) into the OHDSI CDM. Source microdata and a loaded OHDSI CDM remain with the data owners and ATLAS executable analysis metadata are shared. Using the shared metadata, each site produces study results in parallel. A coordinating centre develops the study metadata iteratively with the sites in what amounts to a series of sensitivity analyses (Delaney and Seeger 2012). Study results may take the form of an analysis model with weights and/or aggregates that trace the progress of a target cohort through occurrences to one or more outcomes. Note that because the microdata stays at home (i.e. with the original data provider), case-level anonymisation is largely unnecessary.

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

Figure 1 shows how the Infectious Disease Surveillance and Response (IDSR) Case-Based Surveillance (CBS) Reporting Form, used to collect data on COVID-19, was implemented using synthetic test data during the system development. This is a typical example and would be similar for all the different streams of population health data being brought into INSPIRE.

Note that Figure 1 shows both the design-time and run-time aspects of this process. Once an "on-ramp" – that is, a data source or set of aligned sources – has been identified as a candidate for integration, the data are subjected to a "pipeline" ETL process which can be repeated to obtain data from that source in an ongoing fashion. Design-time activities are only needed for the first introduction of a new data source into the platform – thereafter, the pipeline process functions as designed in an automated fashion.

Data queries and analysis can then be performed against the standard OMOP CDM model, using tools created for this purpose. These include R libraries as well as point-and-click tools, and indeed a query against the OMOP CDM may be expressed in SQL, the ubiquitous relational query language.

The result is a distributed pool of harmonised data from a range of sources across different geographic areas in eastern and southern Africa, and across time as the longitudinal population surveys and other sources add to the central platform. The INSPIRE hub holds (meta)data on many topics of interest to population health researchers, including COVID-19 and HIV, building on the initial effort built on the ALPHA Network, which was focused on HIV data coming from HDSS sites.

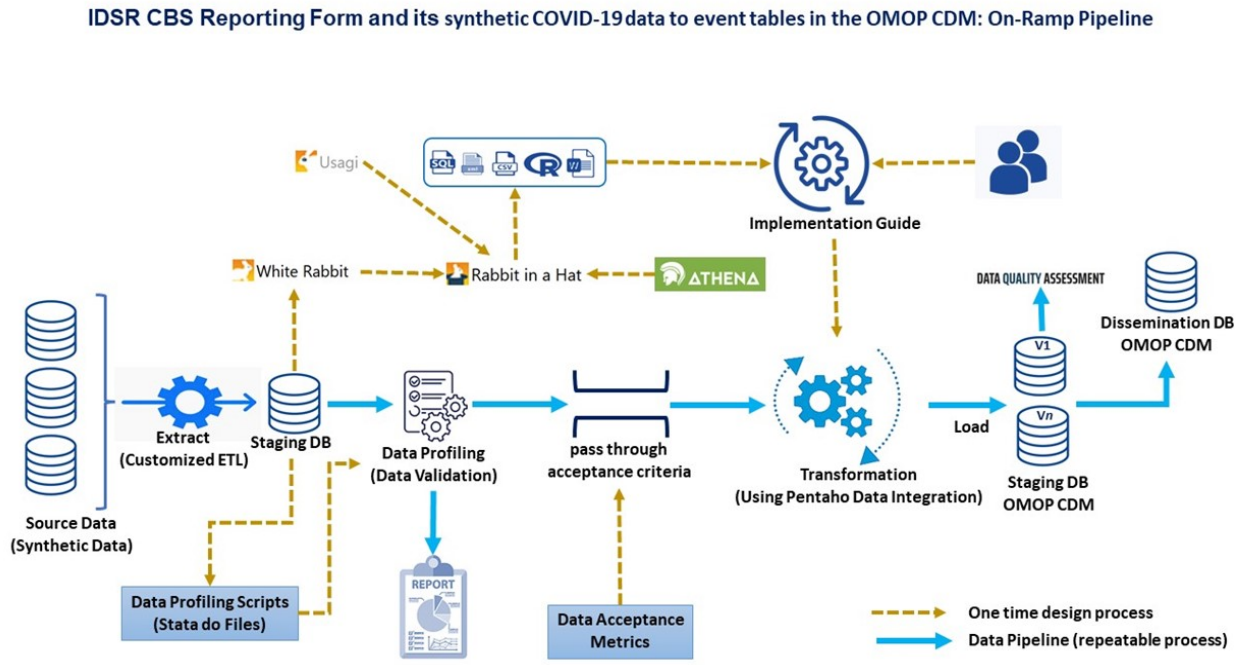

Figure 1. IDSR CBS Reporting Form Extract, Transform, Load (ETL) Pipeline

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

1.3 The research perspective

Researchers are ultimately concerned with understanding and reusing the data to be found on data sharing platforms, and are less interested in the technical process by which data was made available. The challenge here is in knowing what the data selected from the available pool are – where do they come from, and how were they produced?

There is a further complication: often, the shared data are the products resulting from analysis, because the source data are too confidential to be used by a general audience. While researchers often prefer to have access directly to the microdata for their own research, in many cases this is not possible even if these microdata have been pseudo-anonymised and otherwise processed to make them less disclosive (of identifying personal data). These restrictions are often imposed by national Governments through Data Protection acts.

The result is that researchers are often using data produced by an analysis of the microdata in the form of tabulations, other aggregations and/or weighted neural network models (Wikipedia 2023) depending on the analysis, rather than directly using the source microdata. Given the sensitive nature of population health data, this is a fairly common case.

However, having no access to the microdata does not mean that the information about the microdata – the metadata – is not useful and interesting. Knowing that the aggregate data available for use are based on microdata from a particular source, or collected in a particular way, can still be of considerable interest to the researcher, as this has a direct bearing on the quality and usefulness of the aggregates.

The OHDSI models and tools recognize this perspective and enable the sharing of aggregate data across different sources for distributed research. INSPIRE faces the same requirements as the OHDSI community more broadly and thus takes an aligned approach, using this aspect of the OHDSI standards and tools and extending it to support the specific needs of population health.

For a researcher who is looking for data relevant to their own work, it is important that they can search for and locate both microdata and aggregate data, but also information about *microdata that are only available in the form of the aggregates based on them*, included as metadata about the aggregated data set. Understanding this relationship between microdata and aggregates demands, however, that microdata be fully described and characterised. The nature of the study protocols used in the analysis for producing the shared aggregates becomes important metadata for understanding whether the microdata thus interpreted – as aggregate data - is suitable for other purposes, beyond the experiment for which the aggregation was performed.

As a further benefit, such metadata and documentation about experiments can also be critical in determining the reproducibility of findings of a given research output. The term often used for such information - required for reproduction of findings but also used for other purposes - is “provenance.” The ability to fully reproduce an experiment relies not only on a detailed description of the microdata inputs, but also the specifics of the processing environment and the code executed within it. While important (and mentioned in R1.2 in the FAIR Principles) this is not the primary

‘Global cooperation on FAIR data policy and practice’ (WorldFAIR) has received funding from the European Union’s Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

focus of FAIR, which is aimed at supporting data reuse. One of the major goals in making the INSPIRE data FAIR for cross-domain use is to identify ways to provide the full set of information researchers need about the data, both for discovery and reuse, but also to explain its provenance and context.

2. Describing research using INSPIRE (meta)data

2.1. An OHDSI-based Schema.org model

At the Innovative Medicines Initiative (IMI) Programme's EHDEN COVID-19 Study-a-thon in 2020 (OHDSI, 2020), a group looking at FAIR and interoperability in relation to studies employing the OHDSI OMOP CDM outlined an approach for the use of Schema.org, expressed as JSON-LD.

JSON-LD is a way to encode metadata and other information using a flavour of Javascript, especially popular for embedding such information into Web pages such that it is hidden from human users, but visible to computers.

This effort was aimed at the FAIRification of OHDSI, and was described in one of the examples as:

> "how to make observational health databases and observational studies more Findable (notably the FAIR principles F2 and F4) and [increase] interoperability (principles I1 and I2), by creating a metadata model and a human and machine readable website for dissemination of these resources." (EHDEN, 2020)

This work provided the starting point for our consideration of FAIR to describe INSPIRE data. It should be noted that INSPIRE has a particular focus on population health and describes data that are not the typical use case for the OMOP CDM - in large part, because many of the sources for INSPIRE are longitudinal population health surveys.

Figure 2 presents a diagram of the metadata model expressed in Schema.org. In this Figure, the term "literal" refers to a value such as a string or enumerated value taken from a list. The "class" is a more complex structure containing subfields. The colour coding seen in this diagram is reproduced in the INSPIRE diagram which extends it (see Figure 4).

In OHDSI terminology, the metadata description is a description of a network study (a study which uses resources across different nodes in a network – typically different institutional holdings) conducted on top of observational data modelled according to the OHDSI CDM. The microdata stays with the original data providers (at their home institutions) and the metadata needed to conduct the study is shared across study partners in a "study package". By having a standard description of the network study as proposed here, the study design and the several OHDSI CDM databases across which the study is executed. This enables the study as a whole and its multiple OHDSI databases to be discovered and the aggregate data the study produces across these databases to be more easily understood.

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

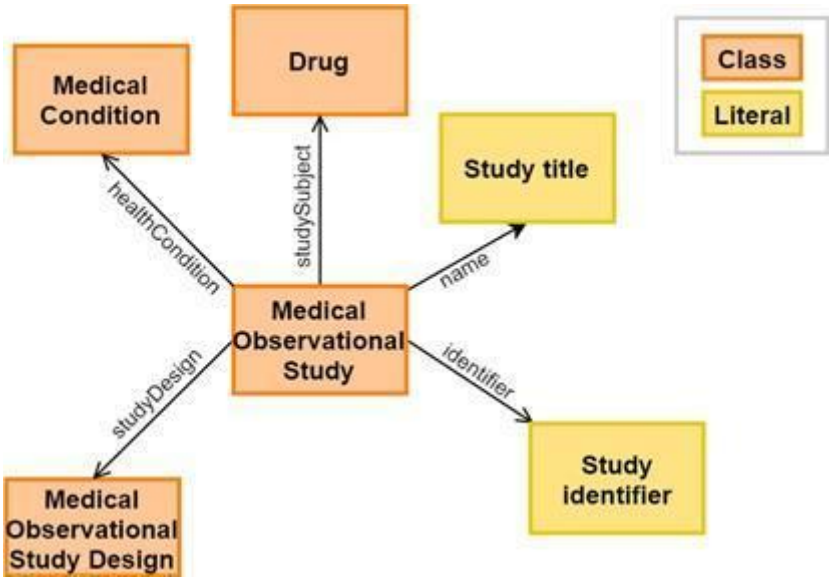

Figure 2. EHDEN COVID Study-a-thon Metadata Model

While FAIR is primarily concerned with data, the findings based on those data are also important for understanding it and reusing it. A description of the study is thus a significant topic for description within FAIR metadata. The connection to the research perspective described in the preceding section can be easily made: if we can discover the study which has been conducted, and also the data on which it was based and the methods and protocols used, we are much closer to fully understanding the research perspective that engendered the study. The selection of those data for our own research purposes would be better-informed as a direct result. In fact, using the study description, a potential partner in a network study can 'run' the study description on their own database to determine whether and how much that potential partner can contribute to the network study.

In order to express this metadata model in Schema.org, extensions were made to the base Schema.org model. This is a typical use of Schema.org, and is an intended use of that standard, which is designed to be specialised (i.e. tailored) for use within particular domains when describing domain-specific resources such as observational health data and observational studies.

It is important to note that the use of Schema.org expressed as JSON-LD explored in the Study-a-thon is a very common approach across domains, and that WorldFAIR WP02 has proposed this as a candidate for use as part of the Cross-Domain Interoperability Framework (CDIF). There are many other domains unrelated to public health which have adopted a similar approach to the implementation of Schema.org for the kind of metadata being exposed. (For example, this is the approach employed in the Ocean Infohub - see https://oceaninfohub.org/odis/, as focused upon in WorldFAIR WP11). Even though the goals of the Study-a-thon were specifically aimed at practitioners in the observational health space, they were employing techniques which are

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

commonly used in other domains as well. Because of this, their work provides a good foundation for further exploration in the more-specific INSPIRE case around population health.

One difference should be noted, as regards the goals of the Schema.org exploration conducted at the Study-a-thon: it explicitly aimed at making the metadata around observational studies "human and machine *readable*" [italics are ours]. While this may seem like a nuance, it is one which is significant in relation to WorldFAIR, which emphasises where possible that FAIR metadata should be not only readable by machines, but also *actionable* by machines, such as when (as described above) a program reviews a database against a description of a study in order to find and invite potential network study partners. The relevance of this will be discussed further in Section 2.2.

2.2 INSPIRE Requirements and Extensions

The approach taken towards FAIRification of OHDSI, as described above, provided a solid basis for further exploring the approach to FAIR which could be taken by INSPIRE. In general terms, INSPIRE tries to employ relevant standards and the OMOP CDM has been identified as the central domain standard for describing health-related observational data. It makes sense to use a similar approach for the FAIRification of INSPIRE. However, the INSPIRE focus on population health – and the differences in the kind of data being described – present us with some additional requirements.

This section summarises those additional requirements and presents a high-level view of the metadata model INSPIRE could use to describe its data, building on the work done at the aforementioned Study-a-thon. It should be remembered that the work done by EHDEN at that event is not a final, production-ready proposal. The Study-a-thon was itself an exploratory exercise, albeit one which was very successful, as we can see in the adoption of its results in the EHDEN interoperability deliverable[4]. The extension of that work by INSPIRE must be understood as a further exploration – none of the additions to the metadata model for INSPIRE should be taken to imply criticism of the work on which it was based. Indeed, the results of the effort in WP07 will be reported to OHDSI in the hope that a final proposal will support all of the identified requirements.

One significant requirement from the INSPIRE perspective is the need for the aggregate outputs of the study – the post-analysis form of the data on which findings are based – to also be described (as metadata). This is a consequence of the public-facing user needs from INSPIRE: some users may not be researchers who are qualified to see the microdata and will instead be restricted to using the less-disclosive aggregates. As described above, however, they will still benefit from being able to search on and understand the microdata on which the study was conducted. In the diagram below (Figure 3), everything outside the blue box is visible: researchers using the aggregates will benefit from information about the microdata, including information about the methods used to collect and process the microdata. Access to the microdata - even after being subject to de-identification and other processing to make it less disclosive - may not be permitted by any but qualified researchers, and gaining such access can be a lengthy process.

[4] https://zenodo.org/record/4474373

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

The audience for INSPIRE data drives other requirements as well. Many population health researchers are familiar with a research paradigm similar to that found in the social sciences and demography, where the typical data are not held in observational databases but are packaged in discrete data sets organised according to the waves of questionnaire-based data collection. In this paradigm, tools such as Stata are commonly employed for data analysis, and data are often stored and exchanged in the discrete files and formats used by these tools. Metadata standards such as the Data Documentation Initiative (DDI) (DDI, 2023) are used to describe such data, and the tools for managing them are based on the DDI model. Researchers familiar with this paradigm may not be able to easily understand the OHDSI model and tools for using observational data drawn from databases and will benefit from having a complete description of the study in which the specific observational data were drawn.

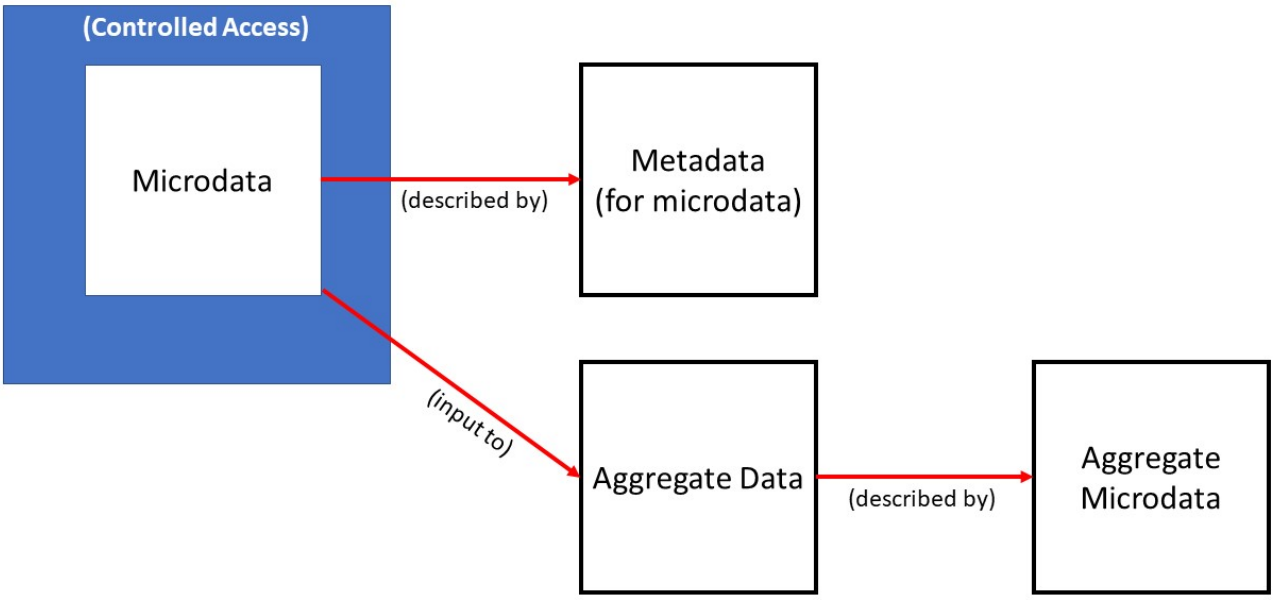

Figure 3. Controlled microdata and associated metadata and aggregate data

If we consider further the needs of researchers outside the population health domain, we can also see that they will have even less understanding of observational databases structured according to the OHDSI domain models. While they may not be familiar with the domain, these researchers will want access to the data – subjects such as COVID-19 show us that many important areas of research are, by their nature, multi-disciplinary and can only be effectively studied using data drawn from several domains. The kind of data held by INSPIRE is a clear candidate for such cross-domain use.

A further requirement relates to the growing volume of data available to researchers, especially when they are looking at data both inside and external to their traditional domains and disciplines. WorldFAIR WP02 has emphasised the need to support machine-actionable metadata descriptions where possible, as a means of dealing with the large volume of data. If we cannot use automation

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

to increase our capacity to document, manage, and disseminate data, then we may be unable to meet the resource demands placed on us by researchers in the future.

INSPIRE needs to consider the extent to which any FAIR solution for describing data is subject to machine-actionability, both in terms of how the data are exposed to users' machines, and in how the data are produced and described within the INSPIRE platform. It is recognized that full automation is likely not possible, at least in the medium term, but it nevertheless remains an important topic for investigation.

The requirements described formed the basis for the extension of the OHDSI Schema.org metadata model. The approach to expressing the model using JSON-LD has been adopted, and the extensions to Schema.org produced during the Study-a-thon have likewise been used.

The high-level diagram of the resulting INSPIRE model can be seen in Figure 4 below. The model will be explained more completely in the following section, but a general description will make it easier to understand what is shown here.

The central object in the diagram (Figure 4) is the Medical Observational Study. This is the experiment or study which is being described. On the right-hand side of the diagram, we see a number of descriptive elements: the Study Title, its identifier, status, and the medical or health use case.

Above the Medical Observation Study, we see the Medical Risk Factor. This describes the internal and external exposome – the exposure of the individuals being observed in the environment. This can be a complex description – different exposures define "pathways" which will be significant in understanding the data and how they can be analysed, where a pathway is a particular set of exposures among those being studied, as applied to an individual. Note that in the study model the Medical Risk Factor is a Schema.org MedicalEntity that can, depending on the study, be replaced - for example, by a Drug such as a vaccination. Drugs are often the study subject in OHDSI-emulated clinical trials. Emulated clinical trials are another type of medical study that is not shown in the diagram yet but will be added as the next planned step.

Other important aspects of the study are shown to the left: Medical Condition and Medical Study Type are self-explanatory, but the Medical Observational Study Design is both detailed and critical to understanding the data. Here, we see different types of analysis (the ones shown are Predictive Analysis and Incident Rate Analysis, but others are possible). In each of these, different types of metadata are needed. A lot of the standard definitions for the metadata have been agreed in the health domain, but further metadata will be needed as new concepts and types of analysis are developed or described.

Both forms of analysis include a definition of the relevant terms: for an Incident Rate Analysis, these are the criteria for stratification; for a Predictive Analysis, these would be the predictors. Each requires a dedicated element as they have different semantics and will need to be understood as distinct analyses (Stratify Criteria and Predictors, respectively, in the model as shown). They

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

become parameters in the selection of the microdata used in the study and are thus important in understanding the variables which will be included in the analysed data.

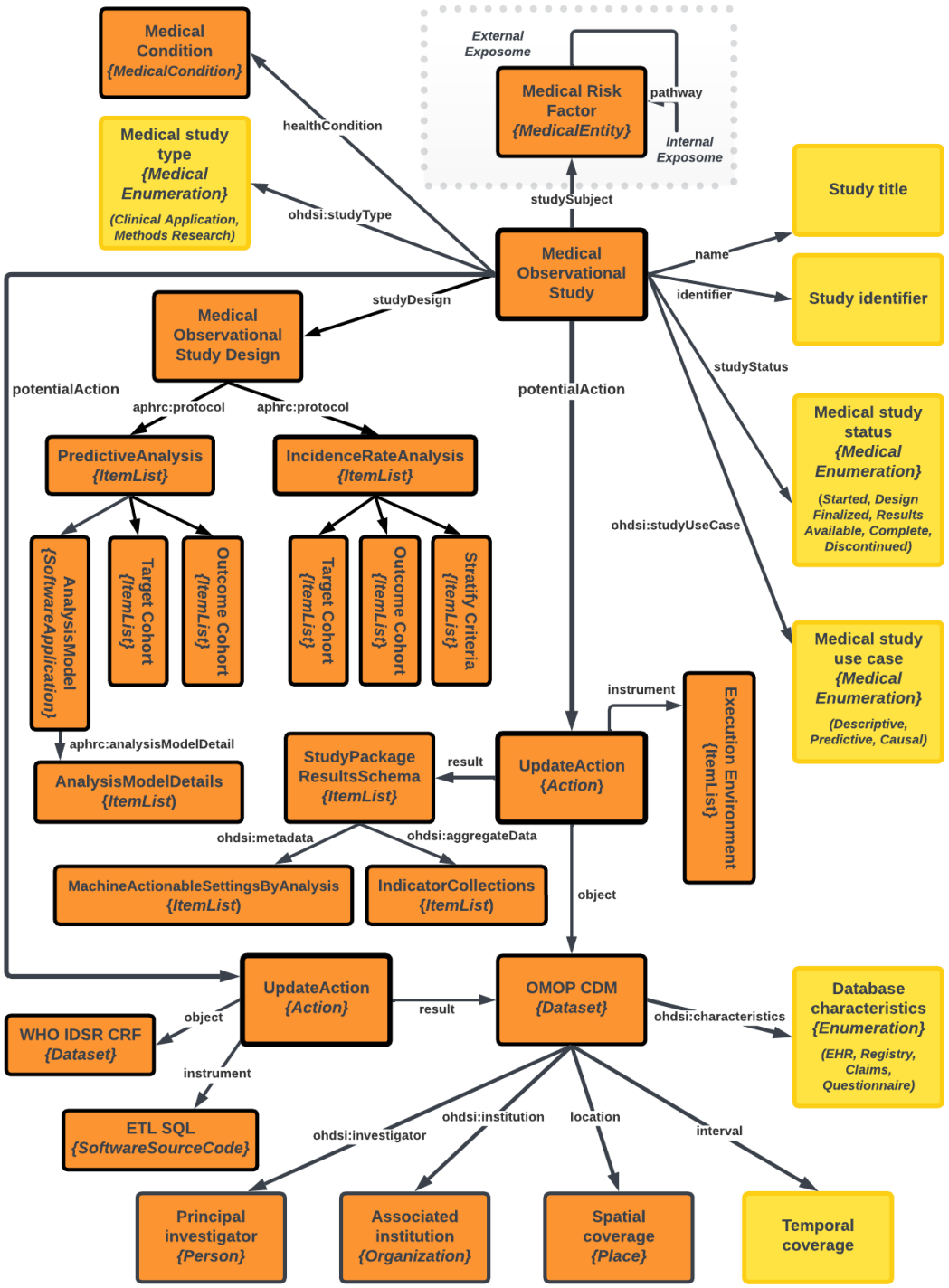

Figure 4. The INSPIRE Metadata Model for Schema.org

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

A predictive analysis will also employ some software which encodes the predictive methods employed, and this will need to be identified. This is described in Analysis Model and Analysis Model Details. Analysis Model Details include outcome labels, the algorithm(s) and hyper-parameters employed in a predictive analysis supervised learning experiment, and details about how the target population is split up between learning (model training) and testing (model validation). Eventually, this section will cover and support weights sharing in deep learning network studies. One other use case for this section of the metadata model is transfer learning.

Common to both Medical Observational Study Design types (i.e., Clinical Application or Methods Research, as shown in the top left in Figure 3) are the Target Cohort and the Outcome Cohort. These indicate the characteristics of the individuals who will be included in the analysis data set (the Target Cohort) and the outcomes reported for those individuals (the Outcome Cohort), based on their path through the stratification criteria or predictors, as described above.

It is typical to display the data associated with these cohorts as a sunburst diagram, with each of the cohorts depicted as a layer, as shown in Figure 4.

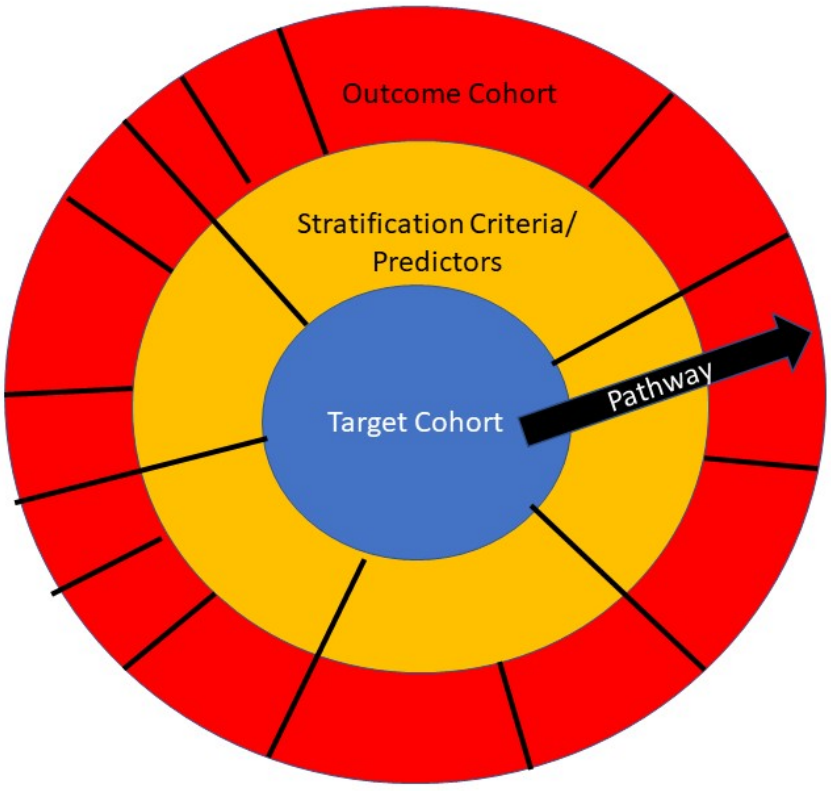

Figure 5. Cohorts organised as a sunburst diagram

Each cohort can be expressed as the results of a query – using the defined terms – against the observational database. Together, they form a group of linked data sets (or a "hierarchical" data

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

set). Each member of the target cohort in the inner layer will appear according to a predictor or stratification criteria in the second layer, and the outcomes associated with the members of this group will be recorded in the third layer. Working from the inside out, the cohorts can be understood as representing a hierarchy. For any given individual, this set of records is termed a "pathway" through the data, and it corresponds to the information in the Medical Risk Factor described above.

The organisation of the microdata used in the study as this type of hierarchical data set is a familiar one to almost all researchers, even if they are not used to thinking about this data in terms of the OMOP CDM. While metadata can be described according to the names given in the OMOP CDM, the "terms" and "concepts" can be understood as "variables", "categories", and so on in more general terminologies such as that found in the core Schema.org model, DDI, and elsewhere.

In the Medical Observational Study Design, we have the metadata which describes the organisation of the data used in the study according to their structural aspects. This includes the selection of cases from those available, and the set of fields (the "variables") which will be included in the record for each case. Further, metadata includes the hierarchical structure used in the analyses. This is not the data set itself, but a description of the data set's structural features – the metadata about the microdata are exactly what can be shown without disclosure risk, as the records themselves are not part of the information provided.

The lower part of the INSPIRE metadata model (in Figure 4) describes the details of the conduct of the study or experiment, which can be done on an on-going basis (as is often the case with longitudinal population surveys). This is an Update Action in our model, which again has several important elements. The source of data is a data set associated with a study that acts as a source of data (in our example, the WHO IDSR CRF). This will be brought into the OMOP CDM through an ETL process (here, the ETL SQL).

The OMOP CDM element represents the observational database(s) used in providing the actual microdata records, as described in the Medical Observational Study Design. The term "database" here must be understood generally, as in the case of INSPIRE the records themselves are the result of a questionnaire administered to the surveyed population, while in other cases these will be records coming from other sources, as per the OMOP CDM. Thus we can think of this – the database populated according to the cohort definitions – as the "data set" containing all microdata records. In fact, the OHDSI community in its study description has used the database extension of a Schema.org data set schema with JSON-LD on several website pages, and these "data sets" now appear in Google Dataset Search.

This data set (i.e. the data for analysis populated according to the cohort definitions) has several elements which describe it in useful ways, including the Database Characteristics, Principal Investigator(s), Associated Institution(s), and spatial and temporal coverages.

The Update Action – the "execution" of the study – takes place in a computational environment that can itself be described with the Execution Environment element.

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

Further, the Study Package Results Schema element describes the study results, including – significantly – the aggregate data that support the findings. These aggregate data are found in the Indicators Collection. Information regarding the settings used in the computation can also be described as part of the results, and these are seen in Machine Actionable Settings By Analysis.

The INSPIRE metadata model is thus an extension of the one produced in the Study-a-thon, in direct response to the additional requirements of public health cohort monitoring. As a result, there is more detail here, produced in a fashion which allows INSPIRE to provide detailed structural metadata about the hierarchical data set to be provided along with the description of the aggregate data resulting from the study, but without necessarily providing access to the microdata records themselves. The data are described according to a more general terminology than that used in the OMOP CDM for those INSPIRE audiences which may not be familiar with the OHDSI paradigm. This allows for a broader group of researchers to find the data they need. Further, the computational environment – which can be associated with specific so-called "devops" scripts – can be used to assure the comparability of execution environments across nodes in a network study and/or for the sake of subsequent reproducibility.

In the description of the OMOP CDM element – the "dataset" – we have the richness of the OMOP CDM model regarding the people and institutions associated with the data, the temporal and geographical aspects, and – through the "object" dataset and the associated ETL – the source variables from which those in the OMOP CDM data set are derived. (We will further explore this in our examples, below).

Thus, we have a rich metadata model for better describing – and enhancing the FAIRness of – the data held in the INSPIRE platform, according to the OMOP CDM model, and building on the work of the OHDSI community generally.

2.3 Considerations

OHDSI is currently working on a more detailed definition of the "study package" which will include a description of aggregate data. This is an area which is of considerable importance to INSPIRE for describing the data produced as the result of a study. At the time of this writing we cannot yet evaluate what OHDSI is proposing and determine whether it meets all of the INSPIRE requirements.

Some audiences for this data – such as organisations within the United Nations (UN) system – would prefer for such aggregates to be expressed in other standards such as SDMX (used for the UN Sustainable Development Goals indicators[5]). The WorldFAIR CDIF is likely to recommend the use of the DDI Cross Domain Integration (DDI-CDI) standard for the description of multi-dimensional data in cross-domain FAIR scenarios.

It seems possible that the aggregates in the OHDSI study package might be amenable to translation into these other standard formats programmatically. If such a step can be fully automated, then that

[5] https://unstats.un.org/sdgs/indicators/indicators-list/

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

will have implications for how INSPIRE addresses the recognized need to describe aggregates, potentially simplifying the picture. This remains an area where further exploration is needed.

3. Elements of the INSPIRE research package description

3.1 Introduction

This section provides a detailed look at some of the elements described above and provides examples of their JSON-LD implementation. This is not intended to be a comprehensive explanation of all of the elements, but should serve to illustrate how this approach will look at a concrete level, and also introduce the reasoning behind the selection of metadata being described in the model.

We will look at two specific areas: first, the description of the variables and their source. Researchers are often looking for specific types of data, and this means that a description of the fields in the data and the provenance of those fields can help researchers when made available to search tools. Secondly, we will consider the overall description of the research package, and highlight the fields described in Section 2.2.

3.2 Variable description

When researchers are looking for data, it is often the case that considerations of data quality are strongly related to the methods used in data collection. For COVID-19 research, a rich source of microdata from a population health perspective are the forms used to conduct surveillance activities. One of the most popular case reporting forms – and one which was used in east Africa – was the IDSR Immediate Case-Based Reporting Form (CRF), which can be seen in Figure 6 below. The use of this form was advocated by the World Health Organisation (WHO) and the US Centers for Disease Control, and is quite common and well-understood.

The IDSR CRF is used as the primary example here for examining FAIR and the way in which INSPIRE supports access to data, because some of the data in the INSPIRE platform comes from sources like the Ministry of Health in Malawi which use this form either exactly or some variation thereof.

As described above in the explanation of Figure 4, one of the Update Actions in our metadata model described in Schema.org is the loading of data through an ETL process. Such processes invariably involve a transformation from the native form of the raw data – in this case, the IDSR CRF – into the harmonised form, which, for INSPIRE, is an implementation of the OMOP CDM.

Researchers looking for data that have been collected using the IDSR CRF should be able to find the INSPIRE data, and to understand that these data are provided in a form which has been harmonised according to a recognized standard. Researchers looking for data in terms of the OHDSI standard, using the concepts held in the Athena system, should also be able to know the method of collection for the data held in the OMOP CDM-compliant form.

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

Often, searches will not be made on the level of the study or the "data set" but will be made for the specific variables or observations of interest to researchers. At this level, there are correspondences between the variables resulting from completion of the IDSR CRF and those observations held in (or described by) the INSPIRE hub. These correspondences need to be described if researchers are to be able to understand how the data has been produced.

IDSR Immediate Case-Based Reporting Form		
	Variables/Questions	Answers – Case n
X	Record's unique identifier (YYYY-WEEK-CCC-PPP-DDD-Case nnn)	
1	Reporting Country	
2	Reporting Province/Region	
3	Reporting District	
4	Reporting Site (Health Facility, Camp, Village…)	
5	Disease/Event (diagnosis): *	
6	Inpatient or Outpatient?	
7	Date seen at health facility (day/month/year)	_____\
8	Patient Name(s)	
9	Date of Birth (day/month/year)	_____\
10	Age (…Years/…Months/…Days).	
11	Sex: M=Male F=Female	
12	Patient's residence: Name of Community/ Neighbourhood	
13	Name of Town/City	
14	Name of District of residence	
15	Urban/Rural? (U=Urban R=Rural)	
16	Address, (cell)phone number … If applicable, name of mother and father if neonate or child	
17	Occupation	
18	Date of onset (day/month/year) of first symptoms	_____\
19	Travel history (Y or N), if Yes, state destination	
20	Number of vaccine doses received in the past against the disease being reported**	
21	Date of last vaccination	_____\
22	Date specimen collected	
23	Date specimen sent to lab	
24	Laboratory results	
25	Outcome: (Alive, Dead, transferred out, Lost to follow-up or unknown)	
26	Final Classification: Confirmed, Probable, Compatible, Discarded	
27	Date health facility notified District (day/month/year)	_____\
28	Date form sent to district (day/month/year)	_____\
29	Person completing form: name, function, signature	
*** Disease/Event (Diagnosis):** AFP, Anthrax, Cholera, Bloody Diarrhoea, Dracunculiasis (Guinea Worm Disease), Neonatal Tetanus, Non-neonatal Tetanus, Measles, Dengue, Chikungunya, Meningitis, Monkey Pox, Yellow Fever, SARS, SARI, Maternal death, Neonatal death, Viral Haemorrhagic Fever, Plague, Typhoid fever, Rabies (Human), Smallpox, death, Influenza due to new subtypes, Adverse Effects following immunization (AEFI), Any event or disease of public health importance (Specify)		
**** Measles, Neonatal Tetanus (TT in mother), Yellow Fever, and Meningitis, etc.** For cases of Measles, NT (TT in mother), Yellow Fever, and Meningitis; 9=unknown		

Figure 6. The IDSR Immediate Case-Based Reporting Form

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

The ETL process transforms the data corresponding to the IDSR CRF form into values which exist in the OMOP CDM, as specific fields in the different standard tables. Tables 1 and 2 show how these mappings were done for the INSPIRE platform:

Table 1: Descriptions of the CDM tables used in IDSR mapping.

Table Name	Table role	Primary key
person	Records that uniquely identify each person or patient. Requires completed fields relating to sex, date of birth, race and ethnicity for each individual.	person_id
location	A list of uniquely identified physical locations of persons and of care sites.	location_id
care_site	A list of uniquely identified institutional (physical or organizational) units where healthcare delivery is practiced.	care_site_id
provider	A list of uniquely identified individuals who provide hands-on healthcare.	provider_id
visit_occurrence	Records of an individual's engagement with a healthcare system for a duration of time. Records are bounded by a start and end date of a total encounter time. They are defined by the circumstances in which they occur i.e. how a patient interacts with a healthcare provider, what cadre of staff are involved in providing care and the duration of the visit.	visit_occurrence_id
visit_detail	Optional - nested within the visit_occurrence table to describe the details within a visit occurrence i.e. movement between hospital units during one period of admission or care from different providers. Each visit detail has its own start date and end date. The start date of the first record and end date of the last record correspond to the start and end dates of the visit occurrence record i.e. the visit detail should be contained entirely within the visit occurrence.	visit_detail_id
condition_occurrence	Records of the present of a disease or condition, including signs, symptoms and diagnoses. These can be observed by either the provider or the patient. Each condition occurrence has its own record.	condition_occurrence_id
observation	Data obtained from an individual during the course of an examination or procedure which cannot be captured in other domains or tables. These include social and lifestyle information, medical history, family history etc. It should not contain information which can be more precisely captured in other CDM tables.	
specimen	Records of individual biological samples	specimen_id
measurement	Records of the results derived from specimen samples. These are the structured values of results obtained from systematic and standardised examination or testing. It can include vital signs and quantitative results from examinations as well as laboratory results. Measurements differ from Observations in that they require a standardized test or some other activity to generate a quantitative or qualitative result.	measurement_id
drug_exposure	Records that capture the utilization of a drug, including vaccines	drug_exposure_id

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

Table 2: IDSR CRF questions and equivalent CDM tables and fields

Question #	IDSR variable name	Table in CDM	CDM variable
X	Record's unique identifier	person	person_id
1	Reporting Country	location	country_concept_id
2	Reporting Province/region	location	county *
3	Reporting District	location	state *
4	Reporting Site (facility, camp, village)	care_site	care_site_name
5	Disease/event (diagnosis)	condition_occurrence	condition_concept_id
6	Inpatient or outpatient?	visit_detail	visit_detail_concept_id
7	Date seen at facility (day/month/year)	visit_detail	visit_detail_start_date visit_detail_end_date
8	Patient Name(s)	---	---
9	Date of birth (day/month/year)	person	year_of_birth month_of_birth day_of_birth
10	Age (years, months, days)	---	---
11	Sex (M, F)	person	gender_concept_id
12	Patient's Residence: Name of Community/Neighbourhood	location	location_source_value
13	Name of town/city	location	city
14	Name of District of residence	location	County *
15	Urban or rural? (U = Urban R = Rural)	?	
16	Address	location	address_1 address_2
17	Occupation	observation	value_source_value
18	Date of onset (day/month/year) of first symptoms	condition_occurrence	condition_start_date
19	Travel history (Y or N), If Yes, state destination	observation	value_source_value
20	Number of vaccine doses received in the past	drug_exposure	drug_exposure_start_date drug_exposure_end_date
21	Date of last vaccination	drug_exposure	drug_exposure_start_date drug_exposure_end_date
22	Date specimen collected	specimen	specimen_date
23	Date specimen sent to lab	specimen	measurement_date
24	Laboratory results	measurement	value_as_number
25	Outcome: (Alive, Dead, transferred out, LTFU, unknown)	visit_detail or death	visit detail: discharged_to_source_value death: death_date
26	Final classification: Confirmed, Probable, Compatible, Discarded	?	
27	Date health facility notified District (day/month/year)	?	
28	Date form sent to district (day/month/year)	?	
29	Person completing form	provider	provider_name

Note that some fields are particularly disclosive (of personal information) and are not mapped, or are not yet completed in the raw data as submitted.

Figure 7 (below) provides a visual representation of how the IDSR CRF fields map onto the OMOP CDM.

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

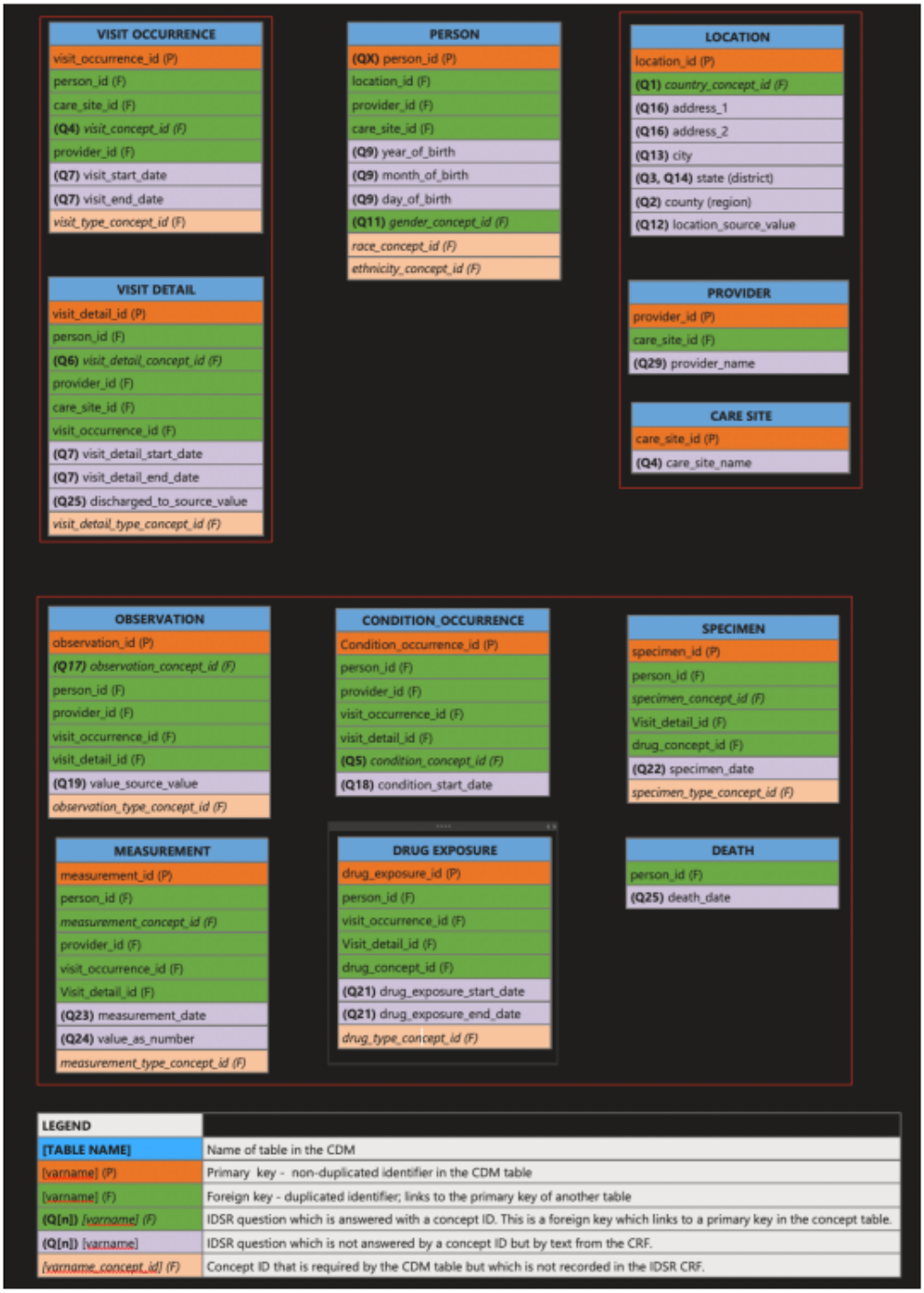

Figure 7. IDSR CBS CRF mapping to OMOP CDM

Below is a portion of this material rendered in JSON-LD, according to the Schema.org metadata model described in Section 2.2 above:

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

```
{
  "@graph": [
        "hasPart": [
            {
            "@type": "schema:Dataset",
            "name": "OMOP CDM Person table",
            "description": "OMOP CDM Person table for Malawi IDSR data",
            "variableMeasured": [
              {
                "@type": "PropertyValue",
                "name": "person id",
                "alternateName":"none",
                "description": "Unique person identifier",
                "value": "number",
                "minValue": "1",
                "maxValue": "51299"
              },
              {
                "@type": "PropertyValue",
                "name": "gender concept id",
                "propertyID": [
                    "https://athena.ohdsi.org/search-terms/terms/8532",
                    "https://athena.ohdsi.org/search-terms/terms/8507"
                ],
                "sameAs": [
                    "http://snomed.info/id/139867007",
                    "http://snomed.info/id/248153007"
                ],
                "alternateName": "none",
                "description": "Gender Concept ID (FEMALE=8532, MALE=8507)",
                "value": "number"
              },
              {
                "@type": "PropertyValue",
                "name": "gender_source_concept_id",
                "propertyID": [
                    "F",
                    "M"
                ],
                "subjectOf": {
                    "@type": "schema:CreativeWork",
                    "name": "WHO IDSR Africa",
                       "description": "Technical Guidelines for Integrated Disease
Surveillance and Response in the WHO African Region",
                                                                                "url":
"https://apps.who.int/iris/bitstream/handle/10665/312317/WHO-AF-WHE-CPI-01-2019-eng
.pdf",
                    "hasPart": {
                        "@type": "schema:CreativeWork",
                        "name": "Annex 2F",
                        "description": "IDSR immediate case-based reporting form."
                    }

                },
                "alternateName": "none",
```


'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

```
            "description": "Gender Concept ID (FEMALE=8532, MALE=8507)",
            "value": "number"
          },
          {
            "@type": "PropertyValue",
            "name": "year of birth",
            "alternateName": "none",
            "description": "year of birth",
            "value": "number"
          }
        ]
    ]
}
```


This is a part of the Schema.org graph which first establishes that this is a graph, and then explains the various parts of it (the entire instance of the metadata model is a “graph” in Schema.org terms). This is essentially packaging for the standard – the first three lines can be ignored for our purposes.

In our example, we have a standard table from the OMOP CDM – the Person Table – which is established as a Dataset in Schema.org terms. (Schema.org uses very general terms for things – a “data set” can be thought of as a two-dimensional layout in Schema.org, like a CSV file, and each table of the standard OMOP CDM model can thus be termed a “Dataset”.) This Dataset contains properties (Schema.org PropertyValue), which can be understood as the columns in the table. These are “person id”, “gender concept id”, and “year of birth”.

Note that while the description of the table says it is used for the OMOP CDM rendering of the IDSR data, it is in fact the Schema.org for the OMOP CDM tables – it is not a direct description of the IDSR CRF variables. Because the data are structured according to the OMOP CDM, this is how the description must be organised. That being said, the OMOP CDM doesn’t lose source (meta)data. Characteristics like gender in the PERSON table or, for example, the condition in a CONDITION_OCCURRENCE CDM table have a provenance. CDM tables capture this provenance in one or more CDM fields that host the IDSR CRF source variables such that every CDM table includes one or more “features of interest” and their provenance.

It should also be noted that fields can be associated with common data elements. In this example, we see the OMOP CDM gender id concept being equated with the corresponding terms in the SNOMED classification (https://www.snomed.org/), using a “sameAs” relationship. Because SNOMED is well-known, this correspondence may be useful to the researcher looking for or assessing the data.

While specific to Schema.org and the OMOP CDM, this is a reasonably intuitive construction: we have a table – a Dataset – which has columns (Properties). Someone searching this data would expect the columns in a person table to contain things like identifiers, gender, and year of birth (etc.). A description for each column is given, and the datatype of each property is also described (the “value”). While minimal, this is the information needed to work with the data in other systems, and may be of interest to those searching for and using those data.

‘Global cooperation on FAIR data policy and practice’ (WorldFAIR) has received funding from the European Union’s Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

Note that in the OMOP CDM there are two or three different types of tables used by the OHDSI analysis workbench (called ATLAS) that INSPIRE Schema.org JSON-LD describes. There are the occurrence tables where each occurrence corresponds to a health-related event like a round in a longitudinal study, the administration of a questionnaire, questions and answers (OBSERVATIONs) made in the course of the questionnaire; environmental exposures; drug exposures; specimens collected; lab tests ordered and lab results reported in connection with the specimens; and signs, symptoms and diagnoses received. Each of these tables contains a single fact and its context. A second type of table is person and person-related tables that enjoy a foreign key relationship linking them to the facts. Person-related tables include LOCATION and LOCATION HISTORY. Finally there are the vocabulary tables. Each fact table, as well as the person table, includes multiple concept identifiers that refer to concepts specified in the vocabulary tables.

3.3 The study package: protocols, execution, and outputs

3.3.1 The Medical Observational Study Design and related cohort definitions

One of the significant aspects of the study design – and one directly related to the data – are the cohorts which identify the hierarchical data set containing the microdata. Cohorts extract features from the occurrence and person-related tables. Cohorts are phenotypes – sets of observable traits in a set of individuals. All analyses include a target cohort and an outcome cohort. Depending on the analysis, there may be a comparator cohort or cohorts for a succession of occurrences that happen between the target and the outcome. Cohort definitions can be expressed as SQL statements, and are often executed in this form by applications because they include inclusion and exclusion criteria against a standard set of tables (the OMOP CDM), which correspond to the parameters of a relational query. Cohort definitions are potentially machine actionable for the purposes of understanding the data selected for the study, if fully described in the metadata. They are always of interest to those wishing to understand the analysis.

Below we show the JSON-LD in an annotated form, to illustrate how this part of the Medical Observational Study can be described in Schema.org according to the INSPIRE metadata model.

```
{
  "@id": "IDSRCovid19PrevalenceAndPredictionStudy/design",
  "@type": "schema:MedicalObservationalStudyDesign",
  "identifier": "<nil>",
  "url": "<nil>",
  "ohdsi:dateModified": "2020-03-28",
  "ohdsi:license": "open-source license",
  "ohdsi:designVersion": "5.4",
  "aphrc:protocol": [
      {
        "@type": "ohdsi:IncidenceRateAnalysis",
        "name": "<nil>",
```


'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

```
            "description": "In an incidence calculation, we describe: amongst the
persons in the target cohort, who experienced the outcome cohort during the time at
risk period",
            "timeAtRisk":{
                "@type": "ohdsi:TimeAtRisk",
                  "description": "The time at risk defines the time window relative
to the target cohort start or end date in which a person is at risk to an
outcome.",
                "startDate": "<nil>",
                "startDateOffset": "<nil>",
                "endDate": "<nil>",
                "endDateOffset": "<nil>"
            },
            "url": "http://51.105.33.160:9090/Atlas/#/iranalysis/2",
```

Up to this point, we see the high-level description of the study design, providing some basic information (id, date modified, url, etc.). The MedicalObservationalStudyDesign is a component of a Schema.org MedicalObservationalStudy schema (Schema.org, 2022).

We then see the description of the protocol – in this case, an incidence rate analysis. The time at risk is then established – an important aspect of the analysis – and the url of the protocol as described in the ATLAS application is provided.

In an incidence rate analysis, we can expect three types of cohorts: a target cohort, describing which persons will be included in the study, an outcome cohort, describing the medical conditions, tests, and procedures, and a cohort defining the stratification criteria. These are structured as lists of items in Schema.org.

The simplest way of describing a cohort is with a link to an external description. We see this here in our target cohort definition – it links to the web page which provides the ATLAS description of the cohort:

```
            "ohdsi:targetCohort":
                {
                  "@type": "schema:ItemList",
                  "name": "<nil>",
                  "url": "http://51.105.33.160:9090/Atlas/#/cohortdefinition/7",
                  "description": "Atlas webpage for Covid-19 patient cohort"
                },
```

If we want to provide a more-complete description, we can list the parameters as we see here in the outcome cohort. (Note that this is explanatory code, so some of the fields are left blank or contain only exemplary “placeholder” values.) We start with the link to the cohort definition in ATLAS, as above, but then itemise the medical conditions, tests, and procedures which describe the outcomes. These contain relevant details, such as the coded description of the condition according to a classification, sameAs relationships, guidance (e.g., evidence levels), and so on.

```
            "ohdsi:outcomeCohort":
```


‘Global cooperation on FAIR data policy and practice’ (WorldFAIR) has received funding from the European Union’s Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

```
                {
                  "@type": "schema:ItemList",
                  "name": "<nil>",
                  "url": "http://51.105.33.160:9090/Atlas/#/cohortdefinition/11",
                  "itemListElement": [
                      {
                        "@type": "schema:ListItem",
                        "identifier": "listItemMedicalCondition",
                        "item": {
                            "@type": "schema:ItemList",
                            "identifier": "itemListMedicalConditions",
                            "description": "This is an array of condition diagnoses
any one of which qualifies a person to move from the target cohort to the outcome
cohort.",
                            "itemListElement": [
                                {
                                    "@type": "schema:MedicalCondition",
                                    "identifier": "medicalCondition #1",
                                    "code": {
                                        "@type": "schema:MedicalCode",
                                        "codeValue": "",
                                        "codingSystem": "SNOMED",
                                        "sameAs": "http://snomed.info/id/"
                                    },
                                    "guideline": {
                                        "@type": "schema:MedicalGuideline",
                                        "evidenceLevel": ["EvidenceLevelC"],
                                        "evidenceOrigin": "<nil>"
                                    }
                                },
                                {
                                   "@type": "schema:MedicalCondition",
                                    "identifier": "medicalCondition #n",
                                    "code": {
                                        "@type": "schema:MedicalCode",
                                        "codeValue": "",
                                        "codingSystem": "SNOMED",
                                        "sameAs": "http://snomed.info/id/"
                                    },
                                    "guideline": {
                                        "@type": "schema:MedicalGuideline",
                                        "evidenceLevel": ["EvidenceLevelC"],
                                        "evidenceOrigin": "<nil>"
                                    }
                                }
                            ]
                        },
                        "nextItem": "listItemMedicalTest",
                        "previousItem": "<nil>"
                      },
                      {
                        "@type": "ListItem",
                        "identifier": "listItemMedicalTest",
                                "description": "This is a placeholder for work in
progress.",
```


'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

```
                "item": {"@type": "schema:ItemList"},
                "nextItem": "listItemMedicalProcedure",
                "previousItem": "listItemMedicalCondition"
            },
        ]
    },
```

The stratification criteria are described in a similar fashion, We start with a link to the ATLAS page which provides a description of the cohort, and then itemise its significant features. Although verbose, this description gives an exact idea of the stratification, by identifying the codes. Again, the OMOP CDM codes are provided, along with the sameAs relationships to SNOMED.

We can thus understand that the age breakdown is for persons younger than 40, those between 40 and 65, and those older than 65. The medical conditions of interest include asthma, diabetes (etc.), and the study describes both women and men, and covers those vaccinated for COVID-19. (This is not a complete example, but is indicative of what can be described.)

The JSON-LD description is sufficient to support both searches (in the terminology of the OMOP CDM or SNOMED) and further machine actionability, with reference to the standard tables of the OMOP CDM. Human-readable descriptions are provided, but the JSON-LD is designed for use by machines.

```
    "ohdsi:stratifyCriteria":
        {
          "@type": "schema:ItemList",
                                                                    "url":
"http://51.105.33.160:9090/Atlas/#/iranalysis/2/definition",
          "itemListElement": [
              {
                "@type": "schema:ListItem",
                "item": {
                    "@type": "schema:MedicalIndication",
                    "description": "Age < 40",
                    "code":
                    {
                        "@type": "schema:MedicalCode",
                        "codeValue": "<nil>",
                        "codingSystem": "<nil>",
                        "sameAs": "<nil>"
                    }
                }
              },
              {
                "@type": "schema:ListItem",
                "item": {
                    "@type": "schema:MedicalIndication",
                    "description": "Age between 40 and 65",
                    "code":
                    {
                        "@type": "schema:MedicalCode",
                        "codeValue": "<nil>",
                        "codingSystem": "<nil>",
```


'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

```
                    "sameAs": "<nil>"
                }
            }
        },
        {
            "@type": "schema:ListItem",
            "item": {
                "@type": "schema:MedicalIndication",
                "description": "Age > 65",
                "code":
                {
                    "@type": "schema:MedicalCode",
                    "codeValue": "<nil>",
                    "codingSystem": "<nil>",
                    "sameAs": "<nil>"
                }
            }
        },
        {
            "@type": "schema:ListItem",
            "item": {
                "@type": "schema:MedicalIndication",
                "description": "Pre-existing condition: asthma",
                "code":
                {
                    "@type": "schema:MedicalCode",
                    "codeValue": "195967001",
                    "codingSystem": "SNOMED",
                    "sameAs": "http://snomed.info/id/195967001"
                }
            }
        },
        {
            "@type": "schema:ListItem",
            "item": {
                "@type": "schema:MedicalIndication",
                "description": "Pre-existing condition: diabetes",
                "code":
                {
                    "@type": "schema:MedicalCode",
                    "codeValue": "195967001",
                    "codingSystem": "SNOMED",
                    "sameAs": "http://snomed.info/id/195967001"
                }
            }
        },
        {
            "@type": "schema:ListItem",
            "item": {
                "@type": "schema:MedicalIndication",
                    "description": "The biological classification of
individuals as female",
                "code":
                {
                    "@type": "schema:MedicalCode",
```


'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

```
                                "codeValue": "F",
                                "codingSystem": "OMOP Gender",
                                                                    "sameAs":
"https://athena.ohdsi.org/search-terms/terms/8532"
                            }
                        }
                    },
                    {
                        "@type": "schema:ListItem",
                        "item": {
                            "@type": "schema:MedicalIndication",
                                    "description": "The biological classification of
individuals as male",
                            "code":
                            {
                                "@type": "schema:MedicalCode",
                                "codeValue": "M",
                                "codingSystem": "OMOP Gender",
                                                                    "sameAs":
"https://athena.ohdsi.org/search-terms/terms/8507"
                            }
                        }
                    },
                    {
                        "@type": "schema:ListItem",
                        "item": {
                            "@type": "schema:MedicalIndication",
                            "description": "Vaccinated for COVID",
                            "code":
                            {
                                "@type": "schema:MedicalCode",
                                "codeValue": "85713-6",
                                "codingSystem": "LOINC",
                                "sameAs": "https://loinc.org/search/?t=1&s=85713-6"
                            }
                        }
                    }
                ]
            }
        }
```


3.3.2 Potential Actions

The second part of the study package we will look at describes the software and the results schema – the aggregates – which are used and produced.

The first part of this study description is omitted – it is a COVID-19 Prevalence and Prediction study, for which general information and cohorts have been defined, as appropriate to the type of study, similar to the technique described in the preceding section.

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

We begin here with the potential action – the study execution. We have a list of the software applications used in the execution of the study, along with the agent responsible for executing them (in this case, a human researcher, as opposed to a machine or an organisation):

```
"potentialAction": [
      {
        "@type": "schema:UpdateAction",
        "name": "MedicalObservationalStudy Execute Action",
        "agent": [
          "person:MuyingoS"
        ],
```

This is followed by a listing of the instruments used in the study, including the R libraries provided by OHDSI: ACHILLES (https://github.com/OHDSI/Achilles), HADES (https://ohdsi.github.io/Hades/), etc.

```
        "instrument": {
            "@type": "schema:ItemList",
            "name": "OHDSI R library",
                 "description": "The software stack by way of which the Action
produces the Result from the Object dataset(s) at the Target",
            "itemListElement": [
            {
              "@type": "schema:ListItem",
              "name": "ATLAS",
              "item": {
                  "@type": "schema:SoftwareApplication",
                  "softwareVersion": "2.13",
                  "applicationCategory": "R Package",
                       "description": "ATLAS is an open source software tool for
researchers to conduct scientific analyses on standardised observational data
converted to the OMOP Common Data Model V5. Researchers can create cohorts by
defining groups of people based on an exposure to a drug or diagnosis of a
particular condition using healthcare claims data. ATLAS has vocabulary searching
of medical concepts to identify people with specific conditions, drug exposures
etc. Patient profiles can be viewed within a specific cohort allowing visualisation
of a particular subject's health care records. Population effect level estimation
analyses allows for comparison of two different cohorts and leverages R packages."
                }
            },
            {
              "@type": "schema:ListItem",
              "name": "ACHILLES",
              "item": {
                  "@type": "schema:SoftwareApplication",
                  "softwareVersion": "1.7",
                  "applicationCategory": "R Package",
                           "description": "Automated Characterization of Health
Information at Large-scale Longitudinal Evidence Systems (ACHILLES). Achilles
```


'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

```
provides descriptive statistics on an OMOP CDM database. ACHILLES currently
supports CDM version 5.3 and 5.4."
                }
              },
              {
                "@type": "schema:ListItem",
                "name": "HADES",
                "item": {
                    "@type": "schema:SoftwareApplication",
                    "softwareVersion": "5.0.3",
                    "applicationCategory": "R Package",
                            "description": "HADES includes PatientLevelPrediction.
PatientLevelPrediction is an R package for building and validating patient-level
predictive models using data in the OMOP Common Data Model format."
                }
              },
              {
                "@type": "schema:ListItem",
                "name": "WebAPI",
                "item": {
                    "@type": "schema:SoftwareApplication",
                    "softwareVersion": "2.13",
                    "applicationCategory": "R Package",
                           "description": "OHDSI WebAPI contains all OHDSI RESTful
services that can be called from OHDSI applications."
                }
              }
            ]
          },
```


The target is then described – in this case, the ATLAS application function, as the platform for the execution.


```
          "target": [
            "http://51.105.33.160:9090/Atlas",
             {
               "@type": "EntryPoint",
               "actionPlatform": "ATLAS",
                      "description": "ATLAS is an open source software tool for
researchers to conduct scientific analyses on standardized observational data
converted to the OMOP Common Data Model V5. Researchers can create cohorts by
defining groups of people based on an exposure to a drug or diagnosis of a
particular condition using healthcare claims data. ATLAS has vocabulary searching
of medical concepts to identify people with specific conditions, drug exposures
etc. Patient profiles can be viewed within a specific cohort allowing visualization
of a particular subject's health care records. Population effect level estimation
analyses allows for comparison of two different cohorts and leverages R packages.",
               "urlTemplate": "http://51.105.33.160:9090/Atlas",
               "contentType": "text/plain"
             }
          ],
```


'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

The object of the execution is then documented: this is the OMOP CDM source database used in the execution.

```
"object": [
    "dataset:mubasCDM",
    "dataset:coordinatingcenterCDM"
],
```

The result set is then described, including the aggregate (meta)data which will be produced, along with the settings used and other relevant details.

```
"result": {
    "@type": "schema:ItemList",
    "name": "OHDSIStudyPackageResultsSchema",
       "description": "Hosts study settings (metadata) and study results
(aggregate data) by analysis from the study package",
    "itemListElement": [
    {
      "@type": "schema:ListItem",
      "name": "OHDSIStudyPackageMetadata",
      "item": {
          "@type": "schema:ItemList",
          "name": "MachineActionableSettingByAnalysis",
               "description": "OHDSI study package metadata by analysis.
Facilitates sharing of analysis settings across multiple CDMs at different
institutions"
      }
    },
    {
      "@type": "schema:ListItem",
      "name": "OHDSIStudyPackageAggregateData",
      "item": {
          "@type": "schema:ItemList",
          "name": "IndicatorCollections",
             "description": "OHDSI study package aggregate study data by
analysis formatted for synthesis at a network study coordinating center",
          "subjectOf": [{
              "@type": "schema:CreativeWork",
              "name": "EvidenceSynthesis",
                    "description": "This R package contains routines for
combining causal effect estimates and study diagnostics across multiple data sites
in a distributed study. This includes functions for performing meta-analysis and
forest plots.",
                 "disambiguatingDescription": "Not for use in synthesizing
predictive analysis network studies across CDMs. Achieving transportability of
predictive analysis models takes techniques like transfer learning, attention based
learning and/or fine turning.",
              "url": "https://ohdsi.github.io/EvidenceSynthesis/"
          }]
      }
    }
```


'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

```
        ]
    }
```

We have focused here on the parts of the metadata model that connect the data of interest to the study, both in defining the microdata and specifying the aggregates produced, but there is more information regarding the study package provided at the same level of detail.

While reading this kind of code may be tiresome, it is the details of the specific model and its implementation which are critical if we are to enable machine-to-machine interoperability. From a FAIR perspective, this touches both on the Findability and Interoperability/Reusability of the data. While machine-readability is critical, machine-actionability places additional demands on how we describe our studies, and these examples should give a flavour of the detail needed.

4. Implementation considerations

Several questions remain to be answered, to determine whether this use of Schema.org is the best one. In general, this approach seems sound but it may be that Schema.org has other fields which might be used instead. Agreement on the best use of Schema.org will ideally be reached with other users of the OHDSI standards and tools, in consultation with them.

One question here is whether the Schema.org “variableMeasured” should be used. While this is not necessarily an intuitive use for those familiar with the OMOP CDM, it is a construct which might be more effective with the more general research audience. This feature of Schema.org is a powerful one, and its use might prove to be advantageous.

Similarly, in describing the study package, we might also wish to employ Schema.org fields such as “measurementTechnique.”

Questions such as these are best addressed according to how the metadata model is leveraged by the relevant tools. Google Dataset Search is one important application which can be considered for the purposes of supporting Findability (in FAIR terms).

Another area which is of importance for implementing production systems is the source of the metadata. Typically, ETL processes such as the INSPIRE “on-ramps” take data from one database and transform it to agree with the standard, harmonised OMOP CDM platform. Some of the description won’t be found, for example, in the study package ETL but will need to be collected from the data producers in narrative form. The workflow needed to support this process at scale is one which will need to be explored and answers to the following questions identified: what tools and resources are required to capture this information in a suitable form?

This is a part of a larger question: how best to produce the Schema.org JSON-LD metadata, and how to deploy it on the Web. The OMOP CDM is a standard structure which describes large numbers of observational health databases. The cohorts defined in the protocols act as queries which, when applied to such databases, produce result sets which can be used as data sets for research. How do

‘Global cooperation on FAIR data policy and practice’ (WorldFAIR) has received funding from the European Union’s Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

we describe the data sets and protocols so that a 'hit' in a tool like Google Dataset Search will bring the researcher to the right place? The data in a database is not a static data set, but Google Dataset Search treats the listings it finds as such.

Yet another consideration is the style of JSON-LD employed. The examples here are nested, which means that properties are found inside the descriptions of the classes to which they correspond. It is also possible to use a "flatter" style of JSON-LD to provide the same payload. The relative advantages of each style should be more fully explored, and any impact they might have on the usability of the metadata determined.

5. Conclusions and next steps

This INSPIRE implementation guide outlines how to expose (meta)data as a FAIR resource, and brings several important themes into focus each of which will need further exploration.

First, we emphasise the importance of aligning and building on existing work within the domain. INSPIRE itself has benefitted from employing the OMOP CDM, through the availability of a range of software tools which would be extremely expensive or impossible to develop and maintain using project resources. When we consider the FAIRification of INSPIRE resources, it continues to make sense to pursue this course: if interoperability is the goal, then aligning with other approaches within observational health makes sense. We are more likely to practically achieve interoperability if we base our implementations on the same fundamental models and technology approaches as used in contiguous communities. The results of the Study-a-thon provided an excellent basis on which to base our work.

One significant difference, however, is between the stated goals of Study-a-thon – an exploration of improved machine-readability – and that of INSPIRE, which is also interested in improved machine-actionability. The Study-a-thon employs a smaller metadata model than INSPIRE, because of these different goals. Machine-actionability requires additional metadata. Both, however, can be described in a FAIR way using Schema.org and JSON-LD. The detail needed for supporting machine-actionability expands what is required in our descriptions by a large margin, and this expansion demands that we do further exploration to determine what features require support, and must be covered by the metadata model. This is an area of FAIR which remains unclear, and which deserves on-going exploration as it is more fully defined within the FAIR community. (For example, is there a benefit to adopting the FAIR Digital Object Framework when the FDOF specification is published? Questions such as these cannot be fully answered at this time.)

It is also important to consider how users - both human and machine - will interact with the metadata we publish. One obvious use case is Google Dataset Search, but there are other Schema.org-compliant applications which we can imagine. Ultimately, we would want to optimise the metadata schema to work with these types of end-user applications. More work remains to be done in this area.

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

INSPIRE serves a specific group of population health researchers, and they are not necessarily identical to the researchers working in other areas of observational health research. The users of INSPIRE data will include policy makers as well as researchers, and these audiences must also be considered. Are there other standards which might be more appropriate for users in some segments of this target audience? For microdata, DDI standards are often used by researchers, and that metadata description might be something which would enable them to more easily reuse our data. For governmental users – as well as those at the international level – standards like SDMX are often used to describe aggregate data. This audience includes the UN organisations such as the WHO. Can these descriptions of the data be derived from the OMOP CDM, and provided in a coordinated fashion with the Schema.org descriptions? The specific needs of the target audience for our metadata should be fully considered. Indeed, INSPIRE is participating in an ongoing initiative along these lines. As part of the OHDSI Africa Chapter, INSPIRE is exploring the use of ATLAS and OHDSI to produce PEPFAR indicators[6] (US Department of State, 2023). Historically, PEPFAR indicators have been represented in an SDMX compliant indicator exchange format (UNAIDS, 2022).

FAIR itself can be understood at two levels: FAIR within a community of use – such as observational health – and FAIR across domains. When we consider the early findings of WorldFAIR WP02, we know that Schema.org, implemented using JSON-LD, is a likely candidate for cross-domain use. But open questions remain: will an OHDSI-based description of aggregate data be useful in this context? Will it align with CDIF? Which other domains will want the INSPIRE data, and what are their requirements? These remain open questions.

The importance of describing not just the data, but the study which produced them, is important to making the data more FAIR. While this may be especially true in a cross-domain context, where the practice within observational health may not be well-understood, it is also true within observational health. An explicit description of the study design and protocols is beneficial regardless of the audience. A more complete understanding of the study, its protocols and methods, and the results of the consequent analysis can be described using Schema.org in the way we have explored, and this information provides a useful basis for others to find and reuse the data.

As this work within INSPIRE proceeds, it will be communicated to relevant people within the OHDSI community, in the hopes that the requirements identified – and possibly the technical approaches developed and evaluated – can benefit others as well. This work on the implementation guide for INSPIRE has highlighted the importance of alignment with the OHDSI standards, especially their Common data Model and related FAIR expressions of it, and also the degree of precision which machine-actionability demands beyond that needed to achieve machine-readability. Further development of these ideas, and their incorporation into the OHDSI toolkit, will not only benefit INSPIRE but potentially enhance the FAIRness of the data holdings of other members of the community as well.

[6] See https://www.measureevaluation.org/resources/publications/sr-16-133.html

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

Bibliography

DDI 2023, Document, Discover and Interoperate, 24 April 2023, Available at https://ddialliance.org" https://ddialliance.org [Last accessed 29 April 2023].

Delaney, J & Seeger, J 2012 Chapter 11 Sensitivity Analysis, 2012, Available at https://effectivehealthcare.ahrq.gov/sites/default/files/ch_11-user-guide-to-ocer_130129.pdf [Last accessed 28 April 2023].

EHDEN 2020, FAIRification of observational studies and databases: the EHDEN and OHDSI use case, Available at https://faircookbook.elixir-europe.org/content/recipes/applied-examples/ehden-ohdsi.html [Last accessed 28 April 2023].

OHDSI 2020 OHDSI Kicks Off International Collaborative to Generate Real-World Evidence on COVID-19 with Virtual Study-a-thon, 2020, Available at https://www.ohdsi.org/ohdsi-news-updates/covid19-studyathon/ [Last accessed 28 April 2023].

UNAIDS 2022 IXF version 3.0, Available at https://data.unaids.org/pub/manual/2008/sdmx_ixf3_en.pdf [Last accessed 29 April 2023].

US Department of State 2023, PEPFAR Fiscal Year 2023 Monitoring, Evaluation, and Reporting (MER) Indicators, Available at https://www.state.gov/pepfar-fy-2023-mer-indicators/" https://www.state.gov/pepfar-fy-2023-mer-indicators/ [Last accessed 28 April 2023].

Schema.org 2022 MedicalObservationalStudy: A Schema.org Type, 25 October 2022, Available at https://schema.org/MedicalObservationalStudy" https://schema.org/MedicalObservationalStudy Last accessed 29 April 2023].

Wikipedia 2023 Neural network, 7 April 2023, Available at https://en.wikipedia.org/wiki/Neural_network [Last accessed 28 April 2023].

Wilkinson, M.D., Dumontier, M., Aalbersberg, Ij.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L.B., Bourne, P.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., Gonzalez-Beltran, A., Gray, A.J.G., Groth, P., Goble, C., Grethe, J.S., Heringa, J., 't Hoen, P.A.C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S.J., Martone, M.E., Mons, A., Packer, A.L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M.A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., Mons, B., 2016. The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data 3, 160018. https://doi.org/10.1038/sdata.2016.18

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.