EHDEN

EUROPEAN HEALTH DATA & EVIDENCE NETWORK

806968 – EHDEN

European Health Data & Evidence Network

WP4 – Technical implementations

D4.5 Roadmap for interoperability solutions

Lead contributor	Kees van Bochove (The Hyve)
Lead contributor email	kees@thehyve.nl
Other contributors	Emma Vos (The Hyve) Maxim Moinat (The Hyve) Sebastiaan van Sandijk (Odysseus) Tess Korthout (The Hyve) Peyman Mohtashani (The Hyve)
Due date	31/10/2020
Delivery date	16/12/2020
Deliverable type	R
Dissemination level	PU
DoA - Version	V1
Date	12/11/2018

This project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking (JU) under grant agreement No 806968. The JU receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA.

TABLE OF CONTENTS

DOCUMENT HISTORY

Version	Date	Description
V0.1	09 Sep 2020	First Draft
V0.2	28 Oct 2020	Revised draft including feedback from WP4
V0.3	29 Oct 2020	Version submitted for formal review
V1.0	30 Nov 2020	Finalized version for consortium review
V1.0f	16 Dec 2020	Final version for submission

DEFINITIONS

Participants of the EHDEN Consortium are referred to herein according to the following codes:

EMC	Erasmus Universitair Medisch Centrum Rotterdam- The Netherlands **(Project Coordinator)**
Synapse	Synapse Research Management Partners S.L. - Spain
UOXF	The Chancellor, Masters and Scholars of the University of Oxford - United Kingdom
UTARTU	Tartu Ulikool - Estonia
UAVR	Universidade de Aveiro – Portugal
The Hyve	The Hyve BV – the Netherlands
Odysseus	Odysseus Data Services SRO – Czech Republique
EPF	Forum Europeen des Patients (FPE) - Belgium
NICE	National Institute for Health and Care Excellence – United Kingdom
UMC	Stiftelsen WHO Collaborating Centre for International Drug Monitoring - Sweden
ICHOM	International Consortium for Health Outcomes measurement LTD - United Kingdom
Janssen	Janssen Pharmaceutica NV - Belgium **(Project Lead)**
Pfizer	Pfizer Limited – United Kingdom
Abbvie	AbbVie Inc - United States
IRIS	Institut De Recherches Internationales Servier - France
SARD	Sanofi Aventis Recherche & Developpement - France
Bayer	Bayer Aktiengesellschaft - Germany
Lilly	Eli Lilly and Company Limited – United Kingdom
AZ	AstraZeneca AB - Sweden
Novartis	Novartis Pharma AG - Switzerland
UCB	UCB Biopharma SPRL - Belgium
Celgene	Celgene Management SARL - Switzerland

Grant agreement	The agreement signed between the beneficiaries and the IMI JU for the undertaking of the EHDEN project (806968).
Project	The sum of all activities carried out in the framework of the Grant Agreement.
Consortium	The EHDEN Consortium, comprising the above-mentioned legal entities.
Consortium agreement	Agreement concluded amongst EHDEN participants for the implementation of the Grant Agreement. Such an agreement shall not affect the parties' obligations to the Community and/or to one another arising from the Grant Agreement.

PUBLISHABLE SUMMARY

The EHDEN project aims to make healthcare data in Europe Findable, Accessible, Interoperable and Reusable (FAIR) at an unprecedented level. This deliverable explores what this entails technically, and how the technical framework of EHDEN, including the OMOP model and the EHDEN Portal, could maximize the FAIRness of its data artefacts and their interoperability. It does so by delineating **digital objects** defined and exchanged in the context of EHDEN and OHDSI, such as model versions, vocabulary versions, database profiles, cohort definitions, query definitions, query results, analytics packages, study protocols, study results, as well as publications, authors, organizations, projects, studies, study-a-thons etc. The deliverable also goes into related **biomedical data standards** and communities that exist and how they relate to each other. The EHDEN project has explicitly chosen to standardize healthcare data for analysis using the OMOP Common Data Model, but other interoperability solutions, such as i2b2, CDISC, FHIR, openEHR, GA4GH etc. also play an important role. This document addresses the relevance of those standards for the data (re)use goals of EHDEN, and points out how OMOP and OHDSI relate to them.

1. INTRODUCTION

The EHDEN project aims to make healthcare data in Europe Findable, Accessible, Interoperable and Reusable (FAIR) at an unprecedented level. The EHDEN technical framework contains a number of important applications to support this objective. For example, the EHDEN catalogue supports Findability by providing metadata (F2) and a data source index (F4), the EHDEN Portal improves Accessibility (A1), the OMOP model, standardized vocabularies and mappings cater to Interoperability (I1, I2), and the Data Quality Dashboard output improves reusability (R1). However, to benefit EHDEN and the OHDSI community long term, it would be beneficial to have a standardized way to publish a number of key data resources as digital objects using globally defined, software-independent data standards in the spirit of GO-FAIR, which can contribute to the Internet of FAIR Data and Services (IFDS).

In this document, we aim to lay out a roadmap for **improving the FAIRness** (which includes interoperability) of **digital objects** which we could define and exchange in the context of EHDEN and OHDSI, such as model versions, vocabulary versions, database profiles, cohort definitions, query definitions, query results, analytics packages, study protocols, study results, as well as publications, authors, organizations, projects, studies, study-a-thons etc.

Another key aspect of the interoperability of healthcare data is an understanding of the various **biomedical data standards** that exist and how they relate to each other. The EHDEN project has explicitly chosen to standardize healthcare data for analysis using the OMOP Common Data Model. The OMOP model is supported by the aforementioned global OHDSI (Observational Health Data Sciences and Informatics, pronounced "Odyssey") community, a multi-stakeholder, interdisciplinary collaborative to bring out the value of health data through large-scale analytics, which is an important community for EHDEN to build on. But other interoperability solutions and communities, such as i2b2, CDISC, FHIR, openEHR, GA4GH etc. also play an important role. This document also aims to address the relevance of those standards for the data (re)use goals of EHDEN, and point out how OMOP and OHDSI relate to them.

From the discussion of related standards, it will become clear if and how there is overlap between OMOP/OHDSI and other interoperability solutions. In the concluding chapter, we will further elaborate on opportunities that may be relevant to the EHDEN goals and aspirations, and provide some suggestions for activities and efforts that would fit on the EHDEN roadmap.

2. METHODS AND APPROACH

This chapter provides a brief overview of the main methods and techniques used in the creation and preparation of the roadmap.

Prioritization of digital objects

In order to prioritize the roadmap for interoperability, we have solicited feedback from the EHDEN and OHDSI community on what digital objects would be most important to start with for improvements of their FAIRness. This was done via the EHDEN WP4 call, an OHDSI community call, and a post on the OHDSI forums. Based on this feedback (see fig. 1), the highest urgency seemed to be around improving the FAIRness of source databases mapped to (patient-level) OMOP databases, which are available to conduct research in, as well as the (network) studies that are executed using these databases. As a next step, we conducted a detailed assessment of the FAIRness of these two important digital objects, which is provided in this roadmap (Annex 1: FAIR assessment of OHDSI studies and databases), and ran a pilot focused on the OHDSI COVID-19 study-a-thon to improve the interoperability of those digital objects, both of which are briefly described below.

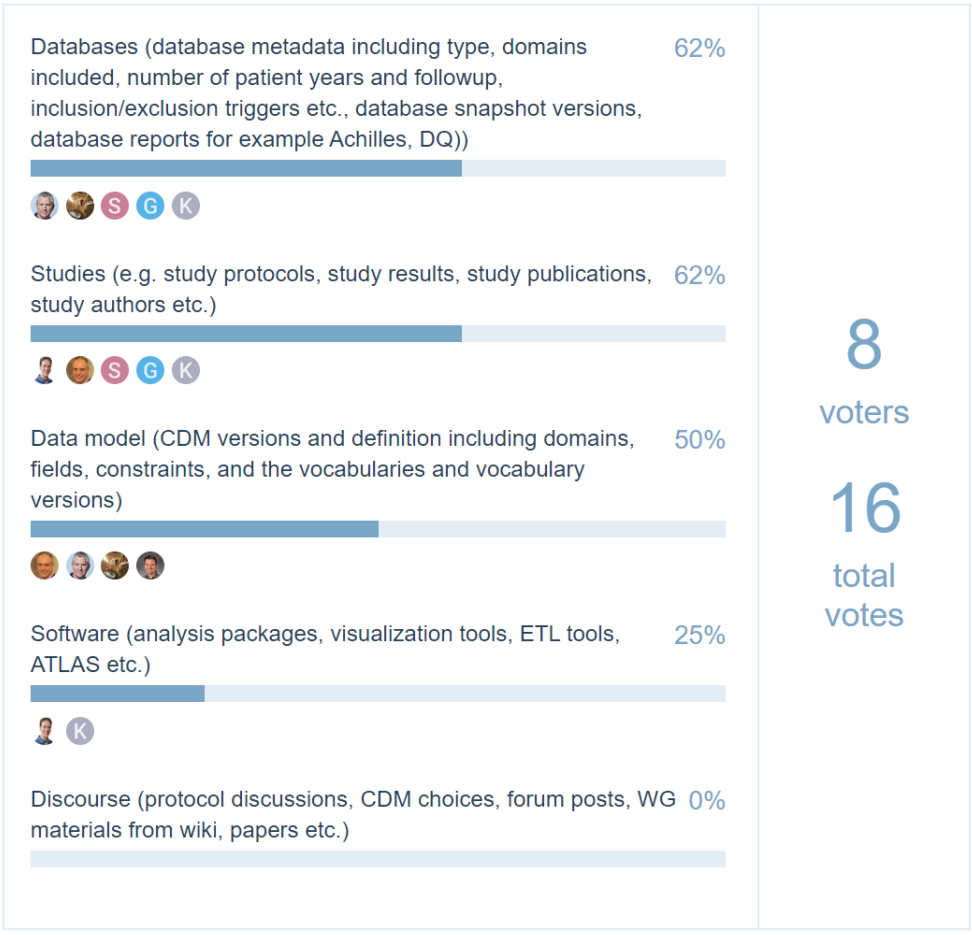


Figure 1. Feedback on the OHDSI forums on the relative importance of the interoperability of various digital objects in OHDSI (https://forums.ohdsi.org/t/implementing-the-fair-principles-in-the-ohdsi-approach-and-tools/10387)

FAIR assessment of digital resources in OHDSI

OHDSI studies and databases (the digital objects prioritized using the method above) were assessed against each of the 15 FAIR principles, as published by GO-FAIR. For each of these principles, it is investigated whether the digital resources and corresponding metadata complies with the principle. If the FAIR principle is met, a score of 2 is assigned. A partially met principle gets a score of 1, and when a principle is not met, a score of 0 is assigned. A total score of 30 indicates a completely FAIR digital resource. Whenever the score is below 30, there is room for improvement for Findability, Accessibility, Interoperability and/or Reusability of the digital resource. The result of this assessment can be found in Annex 1.

Pilot project around improving the FAIRness of OHDSI COVID-19 studies

In order to explore the latest recommendations and standards around FAIRification of open research data, a pilot project was conducted to generate structured metadata about studies and databases, in line with the above described prioritization and taking into account some of the findings from the FAIR assessment. Because of the urgency of the current Covid-19 pandemic, we focussed on FAIRifying the data from the Covid-19 study-a-thon. Our approach for the pilot was to create a website that provides at the same time an easy overview of the studies that have been run and their results for humans, with embedded machine-readable metadata in JSON-LD to make this overview FAIR for machines. This project is described in detail in the Metadata section below.

EHDEN's use of and relation to biomedical data standards

Finally, as explained in the Introduction, it is important for the purpose of this roadmap to explore how the EHDEN goals of making healthcare data and evidence available for research and health policy goals relates to the use of biomedical data standards and interoperability solutions. This includes the OMOP Common Data Model for observational healthcare data, but also other standards such as the Fast Healthcare Interoperability standard (FHIR) for exchanging healthcare data between systems, the Clinical Data Interchange Standards Consortium (CDISC) standards for encoding clinical trial data, the open Electronic Health Record (openEHR) standard for architecting electronic healthcare record systems, the i2b2/tranSMART (for Informatics for Integrating Biology & the Bedside) data models for creating networks of clinical data, and the Global Alliance for Genomics and Health (GA4GH) standards for representing genomics data and patient consent. The relation with these standards is explored in the chapter on Standards and Code. In order to describe this relation, we have relied on expertise from project partners and the authors in using these standards.

3. DIGITAL OBJECTS

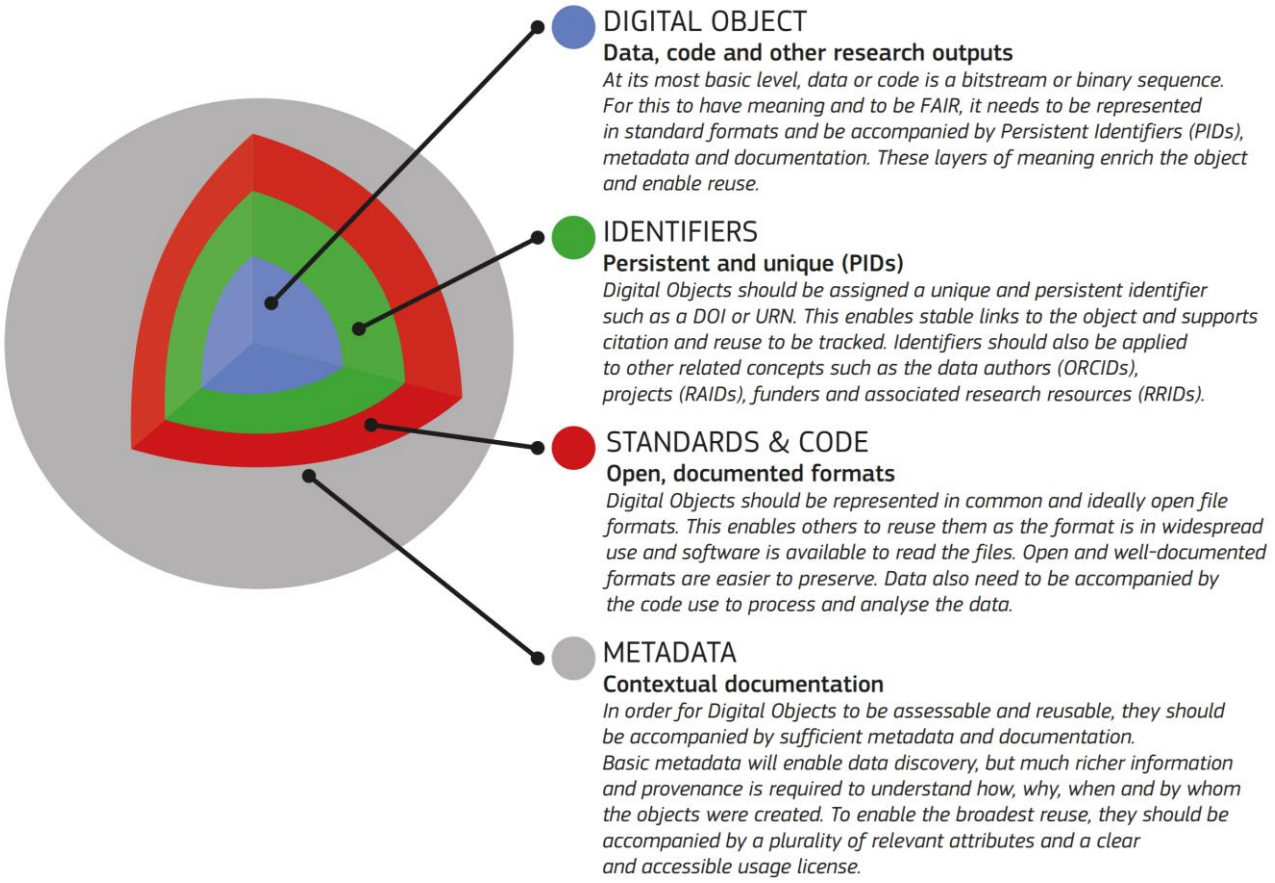


Figure 2: A model for FAIR Digital Objects, noting the elements that need to be in place for data to be Findable, Accessible, Interoperable and Reusable (source: Turning FAIR into Reality, European Commission report, https://doi.org/10.2777/1524, figure 8)

In the report of the European Commission expert group on FAIR data, the second priority recommendation for implementing the FAIR principles and improving the interoperability of data is to implement a model for FAIR digital objects (see figure 2). These objects represent data, software, protocols or other research resources. By making them FAIR, we enable these objects to be properly indexed and found in the wider academic community, and we make them easy to access, use and cite in the scientific and health policy discourse. This is an important objective to support the fundamental goals of EHDEN, to make health data in Europe accessible on an unprecedented level to support better health decisions, outcomes and care.

Digital objects in OHDSI and EHDEN could be delineated at **different levels**: for example, a research study contains a study protocol, which in turn contains cohort definitions, which in turn contain concept sets, etc. But at its highest level, the research study, including its protocol, the software used, information about the participating databases, and of course the study results, is arguably one of the most important digital objects in the scientific and health policy discourse. We can imagine the study in this case as a **digital object** that contains these elements (protocol, software etc.) as smaller embedded (or referenced) digital objects, like a virtual 'Russian doll' (see figure 3). But as we open the 'study' doll and find the smaller objects embedded inside, on each level the same details define the FAIRness (or the beauty and detail of the doll, to stretch this

analogy): the **identifiers** assigned to and used within the objects, the data **standards and code** used to represent the information in them, and the **metadata** that provides contextual information about the object. These 3 important aspects of the digital objects used in EHDEN are worked out in the following chapters, and a forward-looking roadmap is proposed to implement these aspects in the project.

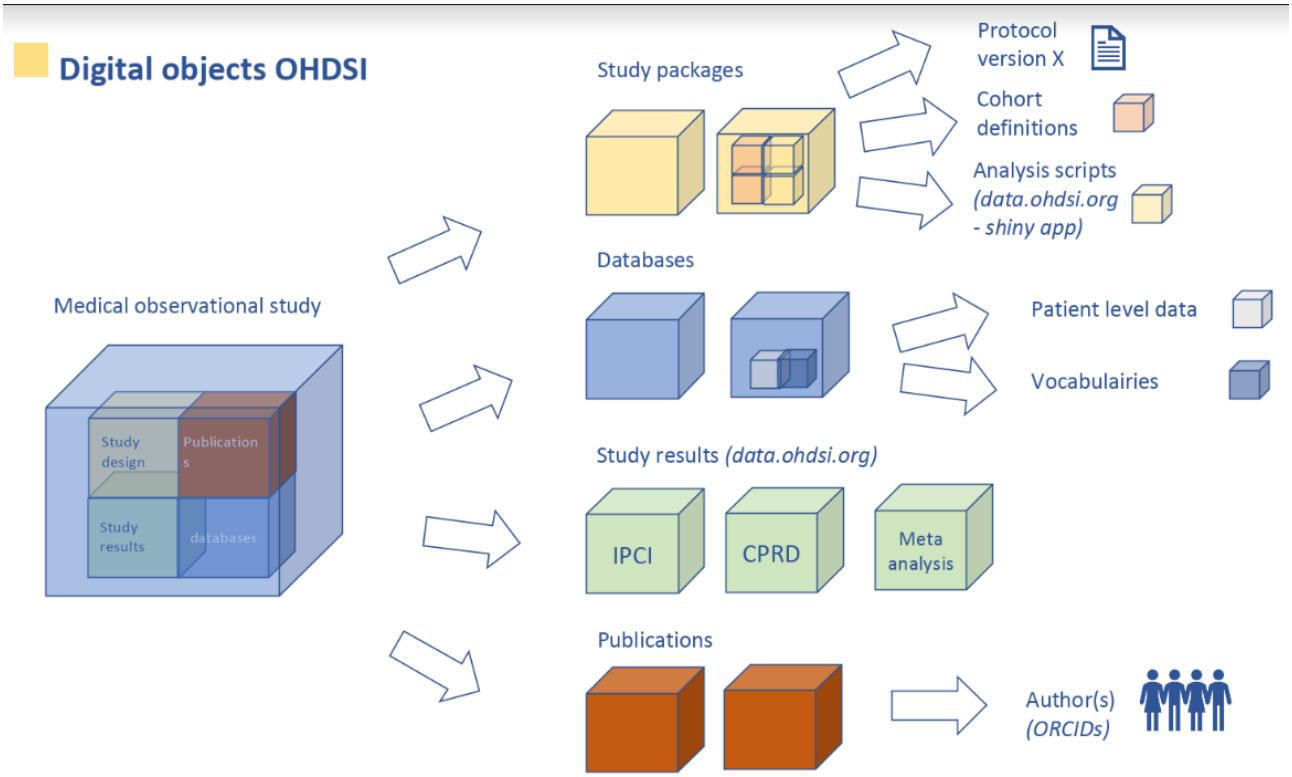


Figure 3. A visual representation of the various digital objects that play a role in an OHDSI study, and how they can be broken down into smaller embedded objects such as the study protocol, databases, results etc.

4. IDENTIFIERS

Throughout the 'digital object cascade', identifiers play an important role, to identify the objects themselves, but also to populate metadata and link to other digital objects of interest. In this chapter, we will focus on a number of important persistent identifier schemes, and where and how they are currently used and could be used in the future to identify digital objects in the EHDEN and OHDSI context. Three key properties of FAIR identifiers are:

- **Globally unique:** the identifier refers exclusively to the digital object in question
- **Resolvable:** there is a service that can take the identifier and return the metadata of the object, and the object or information on how to retrieve the object itself
- **Persistent:** the identifier does not change over time, and any resolver services are guaranteed to be in operation for at least the economic lifetime of the digital object

We will now discuss a number of specific identifiers that we should consider to improve the interoperability of OHDSI and EHDEN data assets.

What is a DOI?

A well-known and mature identifier scheme to use is the Digital Object Identifier (DOI) system. This system allows for assigning a persistent identifier (a so-called DOI name) to digital objects. This digital object can be a physical object (eg. a sample), a digital object (eg. a file) or even an abstract object (eg. a virtual compound). The DOI name resolves to the digital object itself as well as the most actual information about that object, such as the location where it can be found on the internet, contact information, and descriptive metadata. Note that this information can change over time, whereas the DOI itself will never change.

The DOI system was founded by the International DOI foundation in 1998, and it has been used extensively by many different communities since then. Each community is represented by a so-called Registration Agency. These agencies are responsible for handling the registration of new DOI names within their community. The number of registered DOI names is still growing. Examples of large communities are the Crossref application, the DataCite international federation of data centers and the Entertainment Identifier Registry.

Where should we use them in EHDEN/OHDSI

The findability of digital objects within EHDEN and OHDSI could be improved by implementing persistent identifiers. Study results that are published in journals will automatically have a DOI assigned to them. For other objects such as protocols or vocabulary versions, this could be achieved by uploading them to Zenodo. See also the EHDEN forum discussion on this topic.

Recently, study results have also been published in a more interactive manner by creating Shiny apps, which can be found on data.ohdsi.org. Although this URL can be used as an ID itself, there is no guarantee that it will remain the same over time. Furthermore, the URL does not provide a method to attach metadata about the results. We conclude that study results will also benefit from a persistent identifier.

Because the DOI system is mostly used for data in a publication-like format, the question is whether it will be suitable for study results. Registering a new DOI requires a minimum set of metadata, which might not fit perfectly with all elements in study results. Here, the Handle system offers a more flexible approach. This is part of the DOI system but with less strict requirements on the metadata elements.

The role of semantic web identifiers and namespaces

Finally, it could be useful to identify commonly occurring entities and properties in OHDSI using a permanent semantic web identifier. For instance, it would be very helpful to have globally unique, persistent ad resolvable identifiers for all databases that are mapped to OMOP and used for EHDEN and OHDSI network studies. Having such identifiers would make it easier to:

- refer to a datasource throughout a study (both in code and in reading) in a consistent and recognizable manner
- quickly look up metadata on the database in question, such as the target population, data custodian, data access mechanisms etc.
- quickly find all other studies that were performed in the same databases (provided that studies also have clear identifiers, and the definition of the relationship is standardized)

Examples of such identifiers and namespaces are provided in the ohdsi-schemas RDF files produced during the research for this deliverables, as detailed further in the Metadata chapter below.

5. STANDARDS & CODE

5.1 Introduction

As explained in the Introduction chapter, EHDEN builds on the OHDSI community and the standard it develops and builds on, the OMOP data standard (see also the Book of OHDSI). Data in OMOP represents longitudinal health data for one or more persons, and is designed to be queried and analyzed for temporal relations in this data, as visualized in figure 4. Originally, the efforts mainly focussed on digitization and analysis of health insurance claims data, but in recent years the model has been used to represent patient data from a wide variety of sources, including patient electronic health records (EHR), cohort studies and registries and even clinical trial data.

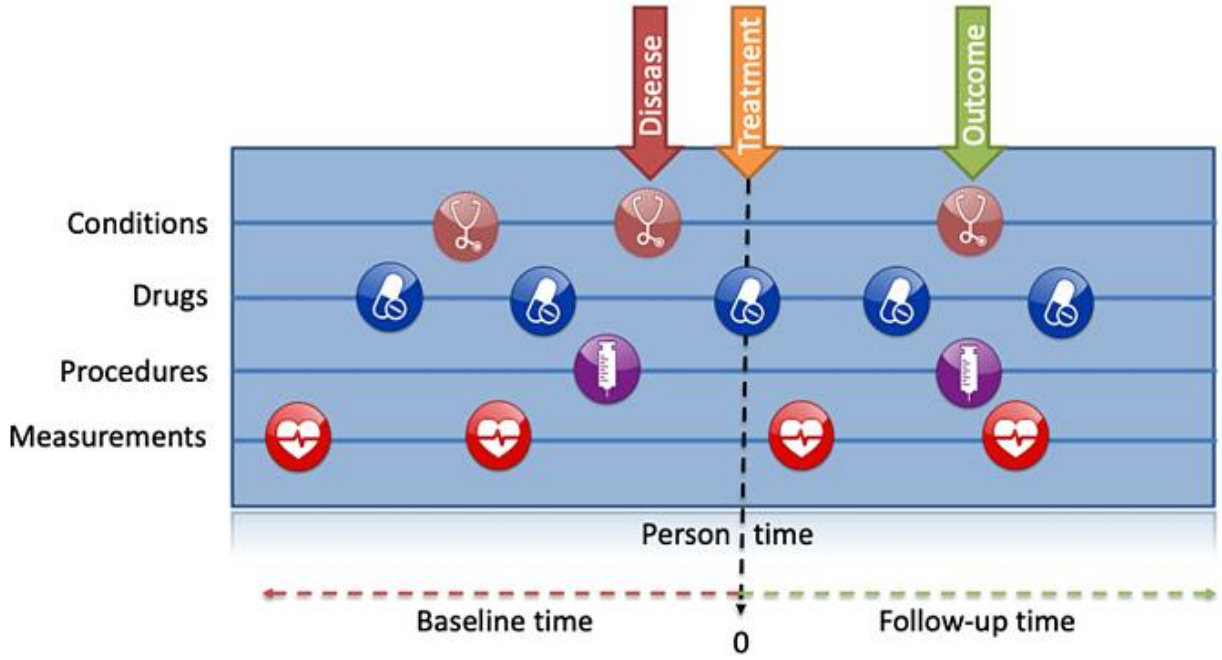


Figure 4. A schematic view on data in the OMOP CDM and how to query them

This raises the question where does OHDSI and the OMOP CDM fit in the wider health informatics ecosystem? In this section we compare OMOP to a number of other biomedical data standards. Although each of these platforms operates in its own domain and has different focus points, there is in some cases considerable overlap in use cases.

5.2 OMOP and openEHR

openEHR is an e-health technology, consisting of open specifications, clinical models and software to create standards in clinics. It has an active open source community that produces and manages standards used to model clinical information in detail. openEHR has been designed to be used in the clinic to collect and exchange electronic health records (EHR) data, focussing on an extensive open standard including hundreds of 'archetypes', a key element of the standard, which can be viewed online in the Clinical Knowledge Manager. The openEHR Foundation is an independent, non-profit foundation, facilitating the sharing of

health records by consumers and clinicians via open specifications, clinical models and open platform implementations.

Although scientific research can be conducted with the data stored in openEHR, the platform is not specifically designed for exploration and analysis of observational data. OHDSI on the other hand is not designed for operational healthcare processes, but does provide scientists with a comprehensive tool set for observational research. The tooling and underlying data model (OMOP CDM) have been designed to generate high quality and reproducible real-world evidence. Because of its strong semantic basis in the clinical archetype definitions, healthcare systems that are built using openEHR design principles would be great sources for a comprehensive mapping to OMOP. Also, the stated goal of the openEHR community ("to fundamentally change the quality of information technology in the service of medicine, so as to improve outcomes in clinical healthcare, public health and the value of secondary data use"[1]) is very complementary to the goals of EHDEN and OHDSI. However, the adoption of openEHR in EMRs and EHR systems is still limited (the openEHR website maintains a registry of known implementations).

When mapping to OMOP, the data from openEHR, which is focused on representing the healthcare and clinical context as accurately as possible, will typically be adapted slightly to account for the analytics focus of the OMOP CDM. This can be illustrated using the observation period as an example, which in the OMOP CDM represents the person-specific time span in which healthcare encounters occur for this person. In research, this determines how many years of follow-up time are available for analysis, both in a cohort and per person. An absence of events in a particular period means that this person did not visit the clinic. For clinical researchers this is important information with regard to cohort inclusion criteria (e.g. minimum follow-up period). In clinical practice this information may be represented in different ways, and will need to be standardized to the observation period paradigm in OMOP during the ETL process.

Overall, openEHR-based health information systems would be a great basis to work with for the purpose of observational research, and broader adoption of openEHR would likely make it easier for EHDEN to obtain high quality data sources for this purpose. It is worth considering to reach out to some of the healthcare organizations that have adopted openEHR to explore this further.

5.3 OMOP and FHIR

FHIR (Fast Healthcare Interoperability Resources) is a next generation standards framework published by Health Level Seven International (HL7). It is a relatively new interoperability solution that combines features of HL7's existing standards and the latest web standard (XML, JSON, HTTP, Atom, OAuth, RDF and many more). This makes FHIR a fast and rather easy to implement standard: it supports other HL7 standards that are currently in use and developers can leverage state-of-the-art web technologies for their solutions.

As an interoperability standard, FHIR is primarily intended to facilitate the exchange of healthcare information between healthcare providers, patients, caregivers, payers, researchers, and anyone else involved in the healthcare ecosystem. It has two main parts – a content model in the form of modular components called 'resources' (see figure below), and a specification for the exchange of these resources; FHIR supports REST and Service Oriented Architectures as well as message and document exchanges. There are multiple sandboxes to explore implementations of the standard, e.g. **Synthetic Mass**, where a simulated population of Massachusetts is exposed through FHIR.

[1] https://www.openehr.org/about/what_is_openehr

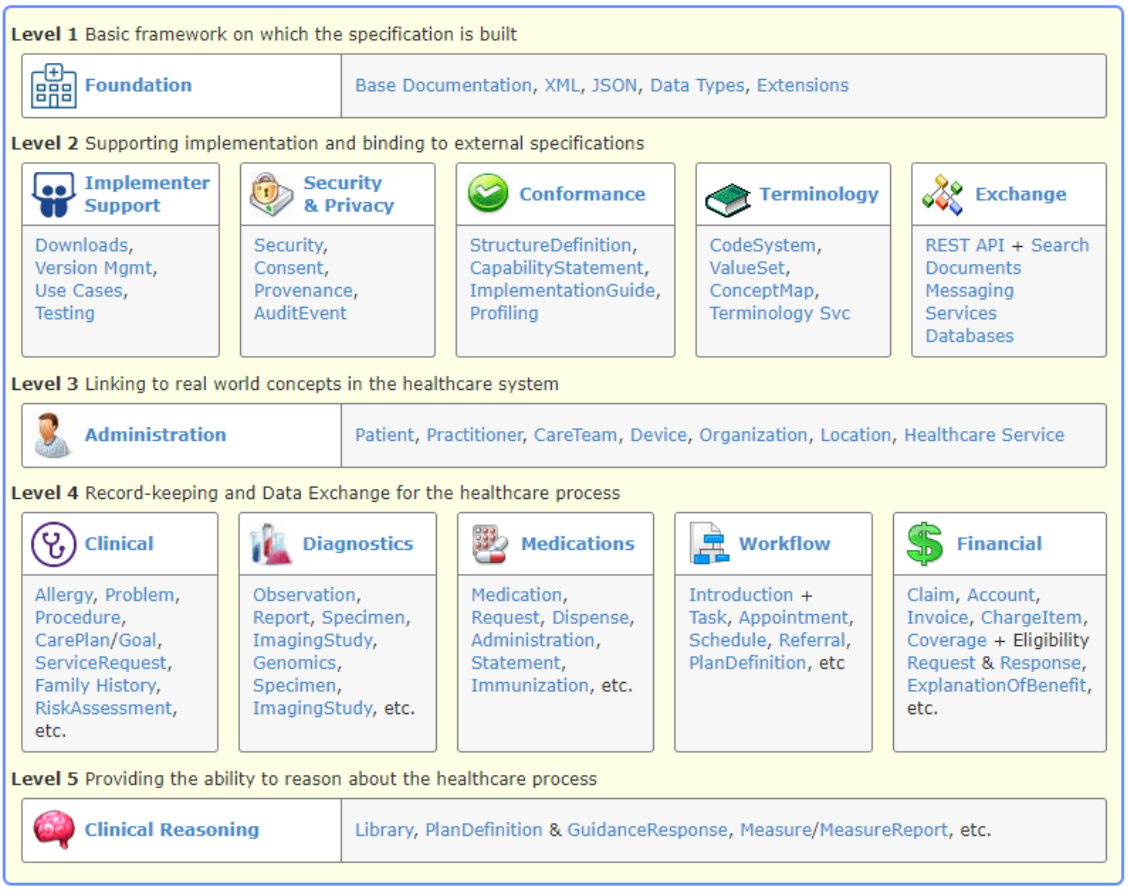


Figure 5: An overview of the Resource framework (content model) in the FHIR specification website.

With this content model, FHIR is suitable for use in a wide variety of contexts, including uses for mobile devices, cloud communications, EHR-based data sharing, server communication in large institutions, and much more. The FHIR platform can serve as a clinical model for analytics and machine learning. FHIR resources model healthcare practice (and related practices) on the information level as well as the process level. It defines which relations exist between resources and provides mechanisms to specify how these are to be used in specific contexts and settings. The FHIR specification also includes workflow definitions and -patterns underlying different categories of resources.

Part of the focus on implementation is the choice to make FHIR a flexible and extensible standard. As healthcare is rather diverse and constantly evolving, standards always face the challenge of coping with an enormous variability in healthcare processes. FHIR addressed this by defining a simple framework for extending and adapting FHIR resources. The 'content model' mentioned above, now encompassing ca 120 FHIR Resources, is not meant to capture all and everything. Rather, the primary resources are meant to represent 80% of what needs to be captured in a specific use case. For the remaining 20%, FHIR solutions can work with specifically defined extensions that use the same framework as the primary resources. Because they are an intrinsic part of the standard, all FHIR systems, on whatever platform, support these extensions.

Further, FHIR works with a flexible maturity model: in a sense, each resource is individually versioned, flagged with a Maturity Level indicating how reliable a given resource is for use. In the current, normative FHIR

version Release 4, the Patient resource has maturity level N (Normative), as well as the Observation. Resources like Medication, Procedure and Allergy, however, currently have a maturity level of 3 (Trial Use): relatively stable and good to use, but not yet part of the normative ANSI standard with the corresponding stability and (guaranteed) backward compatibility. Resources like RiskAssessment (maturity level 1) and ClinicalImpression (maturity level 0) are clearly still underdeveloped. Implementers and other community members are actively requested to provide feedback on scope and usage of such resources.

5.4 Relation between FHIR and OMOP

Important to note is that only a number of the FHIR resources would have an overlap with the data of interest for OMOP and EHDEN. For instance, resources that are related to the operational delivery of healthcare are not necessarily interesting for the secondary analysis purposes of EHDEN. At the same time, such resources typically provide (process) context and meta data that can be relevant for ETL design and -implementation or provenance purposes.

Also important to note is that although OMOP is a universal standard with one opinionated community view on how health data should be modeled for analysis, FHIR can be better compared to a set of building blocks that can be used to facilitate interoperability between healthcare systems in a wide range of use cases. There can be considerable differences between implementations of FHIR in different organizations. When discussing interoperability between OMOP and FHIR, it can be helpful to focus on specific FHIR implementation guides which provide more granular guidance on how FHIR resources are implemented and used. This being said, there are clear opportunities for interoperability between FHIR and OMOP, in particular when healthcare systems have implemented FHIR to expose healthcare data for analysis.

As of 2020, there is a fast growing number of hospitals which support FHIR in some way, as evidenced by the large number of hospitals that support Apple Health Kit (see also next section on the adoption of FHIR). However, it seems that many of those implementations are actually routed through the EHR vendors, and especially in the European context it is also important to consider (groups of) healthcare organizations that implement FHIR endpoints under their own control, as these would be great starting points for potentially mapping to OMOP. As adoption is expected to grow significantly in the coming years, having a way to easily map FHIR resources to OMOP would be very beneficial, and is already currently explored in EHDEN WP4.

FHIR and OMOP complement each other to a great extent, both in terms of focus (exchange resp. RWE) and of typical use cases (healthcare practice workflows resp. research and analytics). Although FHIR does come with a content model that can be interpreted as a data model, and although FHIR is used to set up Clinical Data Repositories and other data stores, FHIR is not about data storage and can work fine with data stores that conform to OMOP-CDM. As to the analytical focus of OMOP and OHDSI, FHIR originally has been focusing on use cases in primary processes and actual healthcare practice. However, in the last couple of years, secondary processes and analytical use cases have grown more and more within scope of FHIR – not in the least because these use cases – both administrative and research driven – are important to various stakeholders in the healthcare/FHIR community.

In practice, this results in growing overlap between OMOP/OHDSI and FHIR – a development that is not expected to diminish in the coming years. In practice also, it has resulted in a number of initiatives around the world that combine FHIR and OMOP. For example, the Sync for Science / AllofUs programme in the US incorporates both FHIR and OMOP-CDM. Also the Synthea/Synthetic Mass initiative is a good example; developed on FHIR, this synthetic data generator is being used quite a lot in the OHDSI community.

5.5 Adoption of FHIR

FHIR has been adopted rather quickly in the last five years (its origins being around 2012). In the context of interoperability and exchange of healthcare data, FHIR is quickly becoming the lingua franca of choice in health and care around the world. Part of its popularity is due to the adoption by the HL7 standards organisation, which provides robust structures and processes for governance and development of the specification according to HL7 guidelines. Another big part is the extensive support provided by a very active community and the adoption of FHIR in the last couple of years by industry (including the big 5) and regulatory stakeholders.

Summarizing the current rate of adoption, it is safe to say that hospitals and (some) other healthcare providers are generally in the process of developing FHIR capabilities, deploying FHIR-based CDRs, APIs and/or servers for specific purposes. As EHR vendors in the US are required to support FHIR-based APIs and other parts of the specification, a majority of the hospitals and clinicians in the US make use of FHIR. In other countries, including European countries, vendors are slower but increasingly supporting FHIR.

Most widespread use of FHIR at the moment in Europe would be in exchange use cases like referrals, interactions between professional and patient (EHR – PHR), and new EHR-EHR exchanges, for instance around medication administration and cross-country patient record sharing. The MedMij programme in The Netherlands is a good example of the usage of FHIR to specify the exchange of healthcare records for the purpose of providing data directly to patients in personal health applications. For these use cases, FHIR is increasingly commonly chosen as the mandatory interoperability standard by regulatory bodies and standardization organizations. In such use cases, vendors and service providers to healthcare organizations typically implement and deploy their FHIR solutions without the healthcare providers necessarily being aware.

An important driver behind the rapid adoption of FHIR, especially in the US but with wider impact, are the various accelerator and development programs. There have been a few of these, such as the Argonaut project (which greatly fostered initial development and adoption), the Da Vinci project (focusing on a.o. value-based care), the CodeX project (for oncology), and the recently launched Vulcan programme dedicated to connecting clinical research and healthcare, which is supported by TransCelerate. These programs usually involve many if not all relevant stakeholders: industry, healthcare, government, research as well as patients and citizens. The coordinated efforts of these parties bring clusters of FHIR resources to maturity and result in Implementation Guides and even new standards for specific use cases, typically brought from inception to market adoption in only a couple of years.

5.5 Three example interactions between FHIR and OMOP in other projects

A good example of interoperability between the two standards is the FHIR-to-OMOP project of the MIRACUM network (Medical Informatics in Research and Care in University Medicine). This network of German University Hospitals brings together their separate healthcare data sets in a federated network of OMOP-CDM instances across the country. The mapping from the respective hospital databases onto OMOP is facilitated via FHIR. The basis of their ETL is a FHIR-to-OMOP job built upon the open source HAPI FHIR library (HAPI is the Java FHIR reference server). The project currently focuses on the study recruitment use case and

works on improving/validating the mappings. For MIRACUM, future work will probably involve amongst other things making use of the OHDSI DQD (Data Quality Dashboard).

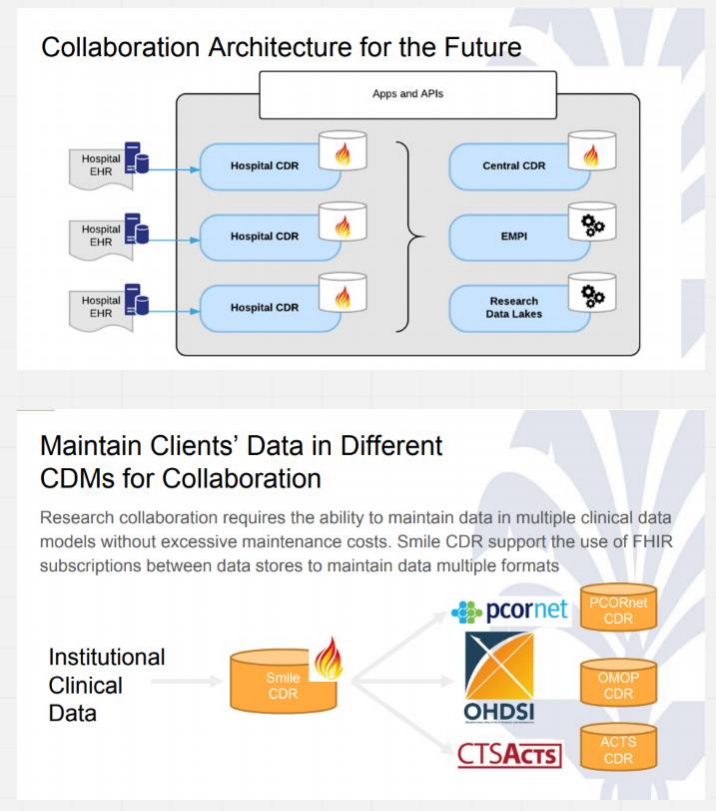


Figure 6. Potential FHIR collaboration architecture proposed in HSSC.

Whereas Miracum works on converting a more or less common university hospital data model onto OMOP, the HSSC (Health Sciences South Carolina) organization captures its source data from standardized (HL7v2) messages. This state wide public/private health data and research collaborative enables multi-institutional health research through its Clinical Data Warehouse and associated governance and research tools. The HL7v2 data are converted in (near) real-time making use of a FHIR-based pipeline (implemented on SmileCDR, which is a commercially supported version of the HAPI server) simultaneously to different analytic CDMs: OMOP, PCORNET and CTSACTS (see figure 6). This infrastructure makes use of FHIR Resources like Subscription (brought to Maturity Level 3 in the Argonaut project) and implements an MPI (Master Patient Index) solution to facilitate multi-institutional and longitudinal analyses.

Given the widespread use of HL7v2, this is an example that can be replicated in many different organizations and collaboratives.

Another way of interoperability between OMOP and FHIR would be the reverse direction, so exposing data that is already in OMOP as FHIR resources. This could be facilitated by the OMOPonFHIR project from Georgia Tech. For the goals of EHDEN this is probably less relevant than FHIR to OMOP to facilitate mapping, but could still be useful in some cases. For instance in case people have analysis software that is already written

to consume FHIR resources, we could support this as an alternative way to run network studies next to the native OHDSI methods library and analytics packages.

The OMOPonFHIR server is set up to directly connect to an OMOP-CDM database. Currently, OMOPonFHIR supports FHIR DSTU2, STU3 and (partially) R4. It uses the HAPI FHIR libraries to map FHIR-to-OMOP and OMOP-to-FHIR for OMOP v5 and v6. Mapping information can be found at OHDSI FHIR Workgroup wiki page (not complete). Like the previously described projects, also Georgia Tech has found ways to circumvent issues where FHIR and OMOP cannot be mapped one-to-one: some of data elements are set statically. The focus of the team at Georgia Tech is twofold:

- Provide bi-directional mappings between FHIR and OMOP. Currently working on AllergyIntolerance and the like and implementing R4 support, their goal is on the one side to support people that have FHIR data to set up an OMOP-CDM instance more easily, and on the other side to provide people that have an OMOP CDM with a FHIR server.
- In 2015, GeorgiaTech has published in Clinical Predictive Modeling Development and Deployment through FHIR Web Services, how OMOP-based patient-level prediction models can be fed back to healthcare practice in the form of patient-level clinical decision support via FHIR. Currently, this would be done via the (more recent) CDS hooks specification (see figure 7 below) but the principle is clear: OMOP-based evidence can be put to practice using FHIR-based exchange and workflow support mechanisms.

The GT team works intensively on the patient validation / study recruitment use case. “Do you have an OMOP phenotype? Check a patient match via ‘OMOP to FHIR’.” Again, OMOP-driven evidence or information needs here are linked to the relevant point of practice using FHIR.

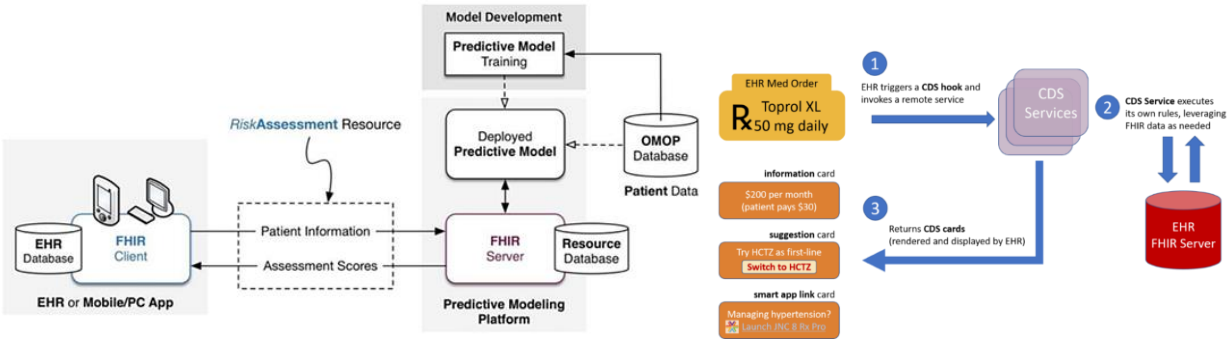


Figure 7. Interaction between Prediction models/CDS services and EHR/healthcare practice a) using the FHIR RiskAssessment resource (left) and b) applying the CDS hooks standard (right).

5.6 Future expectations and developments – relevance of FHIR to EHDEN

As mentioned before, the focus and scope of FHIR has been shifting or rather expanding. In terms of exchange patterns, Message-based exchange (FHIR~HL7v2) and Document (FHIR~HL7 CDA) are increasingly replaced by RESTfull and/or API exchange architectures. In the future, FHIR will also be used to support and strengthen Pub/Sub and Event-based exchange patterns, which are currently being developed and deployed in more experimental settings.

Where typical use cases in the past were mostly within healthcare practice, FHIR solutions are increasingly used to model, capture and/or support analytic, research and other secondary use cases. For OMOP/OHDSI

and EHDEN quite relevant are initiatives working on **FHIR-based protocol definitions and phenotypes**, and examples where FHIR is used to model and **support the clinical trial workflow**, for instance – see this presentation.

The examples described above indicate that also less directly linked FHIR developments may be relevant to OHDSI's RWE journey. Resources like Subscription and Provenance can help **improving OMOP-ETL design** and implementation procedures, whereas developments around CDS hooks (Clinical Decision Support hooks), ShinyFHIR and CQL (Clinical Quality Language) will help the OHDSI community to put the results obtained from its analyses **to use in healthcare practice** more directly. Also, FHIR solutions and mechanisms can readily be put to use in programmes like EHDEN. For instance when deciding on the **metadatamodel for OMOP/observational studies**, phenotypes and outcomes, with further development of the Outcome Driven Healthcare use case (using results of the Value Based Care-oriented Da Vinci Project), and with implementation of more closed-loop solutions for PLP or PLE results in healthcare practice.

Given activities and objectives of the program, it is recommended that EHDEN

1. Sets up a library of relevant FHIR / FHIR-OMOP projects and examples (EHDEN Academy knowledge resource?) for data partners and/or SMEs in Europe. Data partners with FHIR capabilities or aspirations thus will be able to make use of lessons learned and improve their ETL-design and implementation.
    - Given availability of HL7v2 => OMOP and other HL7/FHIR => OMOP mappings: EHDEN could explore with one or more Data Partners if and how ETLs can be designed and implemented more efficiently using these standardized sources cq 'available pipelines'.
    - Follow-up with the Miracum consortium on more information on their implemented solution. Explore the possibility to generalize their solution for EHDEN data partners.

2. Identify FHIR-OMOP projects and initiatives that are relevant to and overlap with the (new) use cases EHDEN works on. Depending on opportunities for collaboration and usefulness, EHDEN could work with some of these initiatives to strengthen interoperability and FAIRness of its approach, methods, and solutions.
    - Would it be possible to improve data provenance tracking and documentation during an ETL making use of FHIR Provenance Resource and operations?
    - To what extent would logical and workflow models being developed in the FHIR community for clinical research, trials and other relevant use cases overlap and/or complement with (meta)data models being developed in EHDEN?
    - Georgia Tech and Miracum both work on the enrollment/recruitment/patient validation use case – is this relevant to EHDEN?
    - OHDSI Rshiny apps and FHIRShiny? Other ways to share study results in a machine and human readily manner?
    - Link up CDM or source data with Ontologies/OWL/RDF?
    - Data quality metadata: fitness for use annotations?
3. Connects with relevant FHIR groups or programs to align future developments. The Vulcan project in particular is highly relevant for EHDEN and its use cases. If so desired, it would be an option to participate in one or more of the Vulcan connectathons or put joint activities on the EHDEN roadmap. http://www.hl7.org/vulcan/

5.7 OMOP and CDISC

CDISC (Clinical Data Interchange Standards Consortium) is a community that develops standards for data emerging in the clinical research process. The CDISC standard encompasses data models, data exchange standards as well as a controlled terminology. Two data models that capture clinical study data are Study Data Tabular Model (SDTM) and Analysis Data Model (ADaM). SDTM contains data collected during a clinical study, which are mainly observations of subjects participating in said study. SDTM consists of multiple domains, e.g. a Demographics (DM) domain, a Adverse Events (AE) domain, and an Exposure (EX). ADaM on the other hand contains data derivations of the clinical study data useful for statistical analysis, and contains a clear lineage of data collection to data analysis.

The CDISC models SDTM and ADaM and the OMOP CDM have overlapping domains with overlapping variables, e.g. the Demographics domain (DM) is fairly similar to the PERSON domain in OMOP, and the Exposure domain (EX) has overlapping variables with the DRUG EXPOSURE domain. Despite this overlap, SDTM/ADaM and OMOP serve different data types. Where SDTM and ADaM are suited for interventional trial data, is OMOP CDM suitable for observational data. SDTM therefore also contains planned events, like planned visit dates, and planned drug doses. This is not something one would normally capture in OMOP, as only events that actually happened are stored.

There are efforts mapping SDTM to OMOP. ADaM is less suitable to map to OMOP, as ADaM mainly contains derived data, which is not suitable for the OMOP CDM as OMOP actual real world observations and no derivations of observations. The OHDSI Clinical Trial Work Group, with major contributions from The Hyve and Odysseus as EHDEN members, is working on conventions to map SDTM to OMOP. SDTM can be mapped to OMOP CDM version 6 plus oncology extensions, without need for further extensions.

It is important to note that even though it is possible and in some cases desirable to map clinical trials data to OMOP, caution is advisable when using this data in analyses, for the obvious reason that interventional data is different from observational healthcare data in nature.

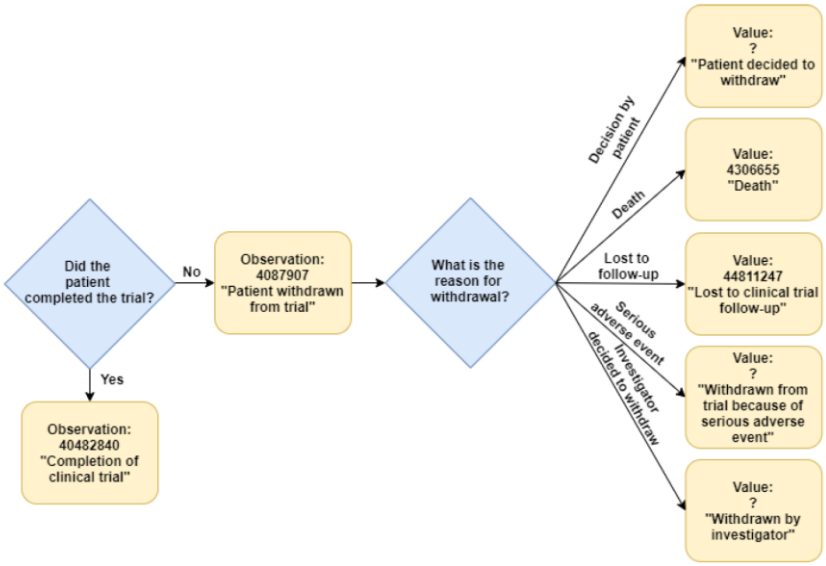


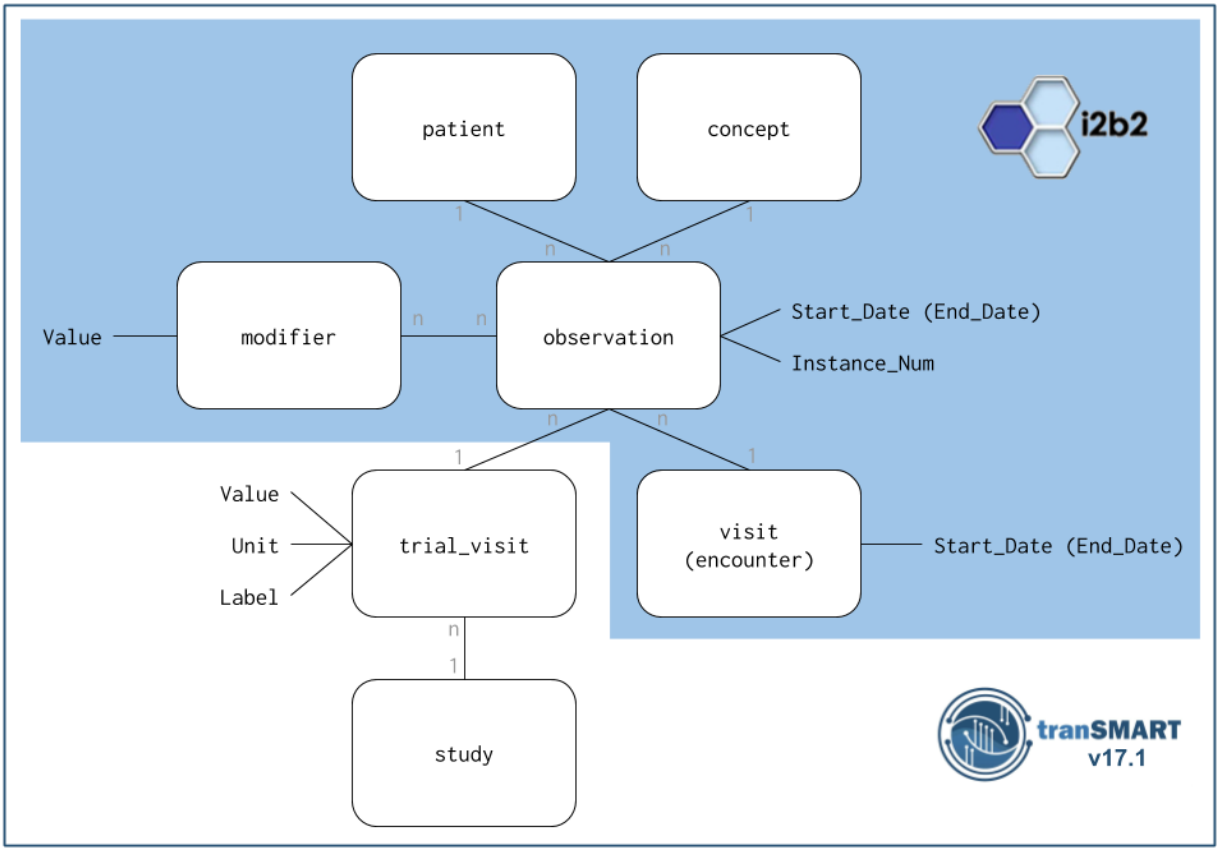
Figure 8. Illustration of ways to map critical trial status information to OMOP, from the working document of the OHDSI Clinical Trials conventions working group

5.8 OMOP and i2b2/tranSMART

i2b2 tranSMART is an open source software platform for query and exploration of clinical, translational and genomics data. There exist a number of flavours of i2b2/tranSMART today, but they all go back to two parent projects: the i2b2 project from Partners/Harvard and the tranSMART project started by J&J.

The Informatics for Integrating Biology and the Bedside (i2b2) platform has been developed at Partners HealthCare System in Boston, Massachusetts since 2004, with funding from the NIH. It is a data warehouse based on a star schema data model, having all observations in a large fact table and having dedicated tables for all observation dimensions.

tranSMART was developed by Johnson & Johnson and Recombinant, extending i2b2 with support for clinical trials and omics data, and has seen a number of subsequent releases, including various versions and variants of the original i2b2 data models, API's and/or code bases. In 2017, a project was funded by tranSMART Foundation and a number of pharmaceutical companies and executed by The Hyve to rewrite and update parts of the tranSMART codebase. The design document for this project contains a detailed discussion of the i2b2 and tranSMART data models and their relations, and a short summary is included below to illustrate the possible relation to the OMOP model:



▲*Figure 9. The tranSMART 17.1 data model. The underlying i2b2 data model is indicated with a blue background.*

The observations have the following domains:

- Patient: For uniquely capturing the study subject
- Concept: For capturing the variable in a tree hierarchy structure. Concepts have a concept_path property that can be used to relate concepts to each other and provide clues for the user interface to represent the data in a hierarchical way to the user.
- Start/End Date: Absolute timestamps linked to the specific observation
- Instance Num: A follow number allowing to differentiate otherwise identical observations, like technical replicates
- (Patient) Visit: The patient specific visit, like for example a hospital visit, bundling multiple observations from the same patient. It can have it's own absolute start and end date.
- Trial Visit: The visit shared over multiple patients, for example specific visits in a clinical trial, like 'Baseline' and 'Week 1'.
    - The trial visit can have:
        - A human readable Label, for example 'Week 1'
        - A Value and Unit for capturing structured information on the visit, like '7' and 'Days'
    - The trial visit is also the link to the Study, uniquely identifying clinical trials or other studies or dataset groupings. This dimension is also used for defining data access per user.
- Modifier: This dimension allows for capturing any other metadata on the observation, like the dosage or way of administering. It allows theoretically for adding an infinite amount of extra dimensions.

Since there are no hard constraints on which concepts are used in the i2b2 data model, this data model is very flexible and allows for many different use cases: sharing EHR data, sharing data from biobanks, clinical trials etc. This flexibility is likely one of the main reasons why the platform has been adopted by more than 200 institutions worldwide.

i2b2/tranSMART to OMOP

If institutions currently have their data in i2b2/tranSMART and they would like to map their data to OMOP, for instance to become EHDEN data partner, this would require a case-by-case investigation of how i2b2 is used. In principle, it should be possible for data in scope of i2b2 to map this to OMOP, and the fact that both are designed as patient centric data marts or warehouses is helpful for this purpose. But because the i2b2 and tranSMART data models are rather flexible, when it comes to mapping the actually populated i2b2 concepts to OMOP, likely custom mapping logic will be needed to determine to which OMOP domains the referenced concepts will need to be mapped. See the figure below for a potential high level mapping from the i2b2 and tranSMART data models to OMOP.

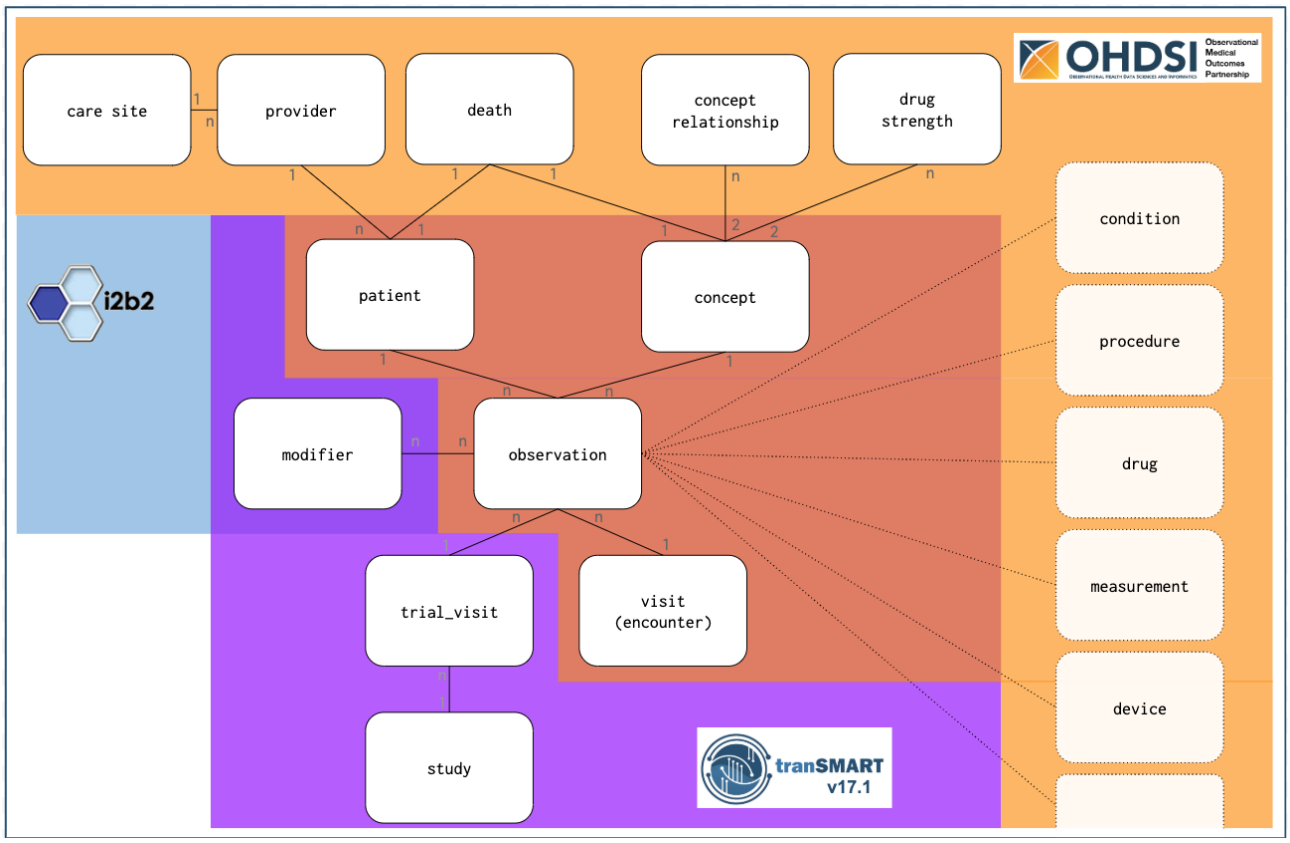


Figure 10. Potential syntactic mapping from i2b2/tranSMART to OMOP domains, from supplementary documentation to the presentation of Maxim Moinat at the i2b2/tranSMART meeting in Tubingen in 2019.

There are a number of ongoing initiatives for interoperability between OMOP, i2b2 and other standards such as PCORnet, such as in the All of Us Research Project and the National COVID Cohort Collaborative (N3C).

5.9 BRIDG

The Biomedical Research Integrated Domain Group (BRIDG) Model is a collaborative effort engaging stakeholders from the Clinical Data Interchange Standards Consortium (CDISC), the HL7 BRIDG Work Group, the International Organization for Standardization (ISO), the US National Cancer Institute (NCI), and the US Food and Drug Administration (FDA). The goal of the BRIDG Model is to produce a shared view of the dynamic and static semantics for the domain of basic, pre-clinical, clinical, and translational research and its associated regulatory artifacts.

BRIDG is potentially interesting because it aspires to semantically describe the domain of biomedical and clinical research, rather than focusing on one specific application. It can and is used to translate between some of the models mentioned above, for instance in the Common Data Model Harmonization (CDMH) and Open Standards for Evidence Generation project (see their recent report). For practical purposes such as ETL the model is likely too complex and detailed to use, but it can be a good reference model when semantic discussions arise or when interoperability between various standards needs to be formalized.

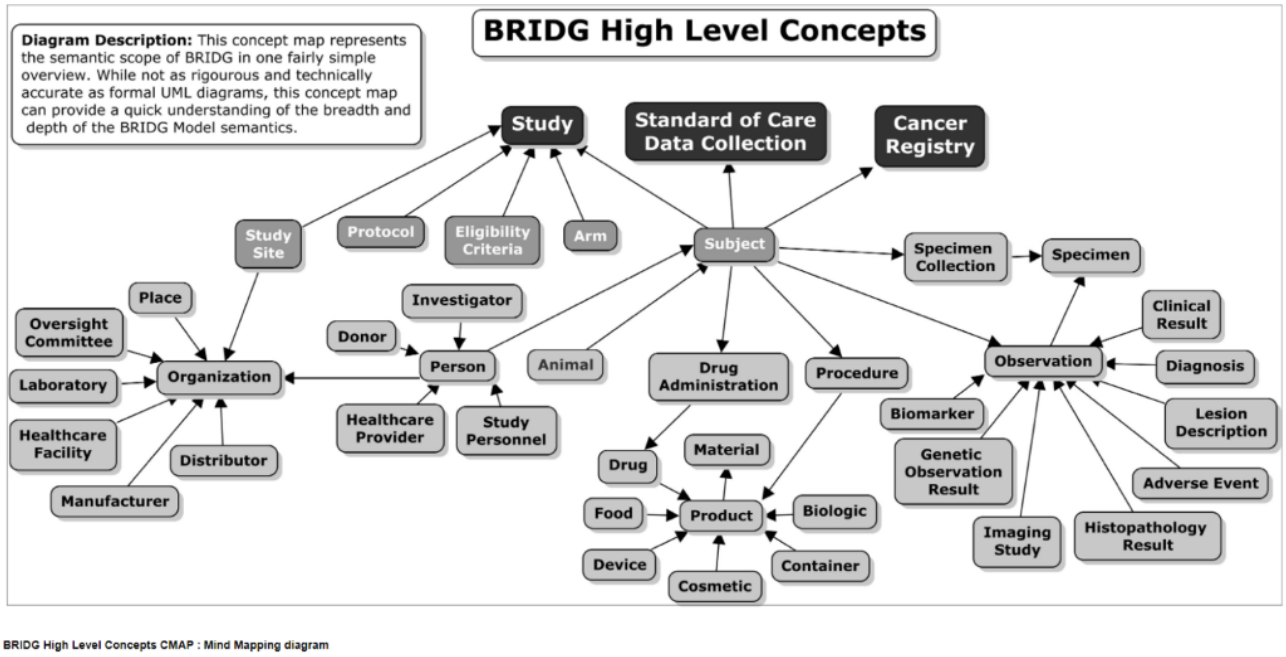


Figure 11. High level concept map, loosely based on the BRIDG model. (source: CBIIT github)

5.10 Conclusion

As we have shown, openEHR, FHIR, CDISC, OMOP and i2b2/tranSMART are not exclusive standards, but each has its place in the health care and medical research data environment. Side by side, they can be combined to serve a particular purpose. In summary:

- openEHR focuses on comprehensive clinical information models and can be used to implement systems that collect and persist medical records using open standards. These standards enable the exchange of patient information between clinicians and/or to other standards.
- FHIR is an important standard to work with because of its fast growing adoption in healthcare. Medical apps can be built on top of FHIR to aid/support patients, health care providers and researchers. Implementing FHIR to OMOP mappings is a growing interest for EHDEN as more and more health data providers implement FHIR in production.
- Various CDISC standards such as SDTM focus on representing data from clinical trials, and there is ongoing work through OHDSI and EHDEN to map these data to the OMOP standard for analysis.
- OMOP and the OHDSI tools have been built for scientists, providing them with the tools to generate high quality and reproducible evidence from observational health data. The data can be **extracted, transformed and loaded** directly from medical claims data or EHR-systems, and exchange standards like FHIR can facilitate this process.

6. METADATA

Recently, the OHDSI community organized a so-called **study-a-thon around COVID-19**, which was aimed at addressing a number of key open medical questions (e.g. real world safety of hydroxychloroquine, effectiveness of ACE inhibitors, prediction of hospitalization etc.) in a large collection of healthcare databases

around the world. The EHDEN team played a key role in organizing this study-a-thon, and many important digital assets were produced, such as study protocols, database metadata and characterizations, study results, publications etc. In this use case, we describe a mechanism for making these digital assets FAIR in a real-time manner using widely adopted web standards such as JSON-LD.

Two key goals for these solutions were:

- Provide visibility of the ongoing collaborative research to the rest of the world, both to humans and machines
- Create an overview and future reference for study-a-thon participants on the various digital assets created in the study process.

6.1 Process

FAIR Assessment: establishing the baseline

The FAIRification of study-a-thon data began with a FAIR assessment in order to better understand the current state of the data. The assessment covered two aspects: firstly, the studies themselves, including elements such as protocol and related publications, and secondly, the study's data sources. The complete assessment can be found in Annex A.

The FAIR assessment of the studies revealed that there was room for improvement in both Findability and Reusability, while the score was comparatively higher for Accessibility and Interoperability. Methods to improve Findability could include the implementation of an identifier scheme for both data and metadata, as well as using richer and more extensive metadata. Metadata, especially provenance metadata, would also increase the Reusability of the data. Data sources for a study are currently listed in a table, with a few columns of metadata for each entry. Here we saw improvement possibilities in Findability and Interoperability due to the lack of identifiers and structured data. Adding structure to the metadata, such as through external vocabularies, would greatly increase both the Interoperability and Findability of the data sources.

JSON-LD Schema.org: defining the semantic model

Since increasing Findability was a common denominator for both studies and data sources, we have been developing structured metadata elements tailored for studies and data sources. These elements are encoded using JSON-LD, which will be embedded in the study-a-thon website, thus creating a common definition of the data and metadata to both humans and machines. Representing the metadata in JSON-LD also increases the exposure of the study-a-thons to common search engines.

Metadata elements are based on schema.org, which is a broad and widely used vocabulary to make data on the web discoverable. The drawback of schema.org is that it is not extremely detailed, a compromise we settled on given the advantage of its popularity for web-based data. For each study and data source, a controlled set of metadata elements are chosen to describe an instance of each. For study, metadata covers concepts such as the type of study, the drug studied and information on the medical condition (Table 1). Similarly, for data sources, metadata is focused on the data source characteristics, such as the population covered in the data source, as well as provenance information. For a small percentage of the metadata, elements did not already exist in schema.org. In these cases, the vocabulary will be extended with custom concepts and relationships.

The resulting proposed metadata elements are published on github, and below are a few examples of schema.org concepts that were leveraged to provide structured metadata for studies.

Study metadata element	Corresponding schema.org concept
Study	schema.org/MedicalObservationalStudy
Study ID	schema.org/identifier
Study Title	schema.org/name
Study Description	schema.org/description
Medical Condition Studied	schema.org/MedicalCondition
Drug Studied	schema.org/Drug
Researchers/Authors	schema.org/Researcher schema.org/author

Table 1: An example of a few study metadata elements with the corresponding concept from schema.org. In the case of 'Study Type', which describes if the study is an estimation, prediction or characterization study, there is no equivalent concept in schema.org. 'Study Type' will map to a custom concept that will be encoded in the website's JSON-LD, along with all the schema.org concepts.

Site generator: Publishing the digital assets for humans and machines

Once all the metadata for a study and data source has been defined and, where possible, mapped to its corresponding schema.org concept, these can be expressed in a machine-interoperable JSON-LD format to be implemented in the study-a-thon website. For this particular project, we chose a static site generator named Hugo, which already had a nice theme that we could leverage for this purpose. The exact technology for the website is not very important with respect to the FAIR principles, but the advantage of choosing a static site generator is that it is very easy to take structured content (in this case expressed in YAML) and map that to both rendering to HTML and CSS for human readability as well as structured data rendering in the form of JSON-LD to enable machine readability. The Hugo project already has a site theme named academic that we adapted for this purpose.

6.1.1 Outcome

This case study describes work in progress, so we may update it once we have finalized the approach and made it reusable for future study-a-thons and hackathons that involve real-time production of research assets such as study protocols, results, preprints etc. However, currently the result consists of a website that both highlights the key outcomes of the study-a-thon as human readable data, as well as exposes structured JSON-LD data to facilitate machine readability for search engines and web crawlers. The source code for this website is on github, and the end result can be viewed live on https://covid19.ohdsi.app/. The proposed JSON-LD and RDF standard for representing OHDSI metadata are also published on github. This standard can be used in the EHDEN portal for instance to provide structured metadata on data sources and studies.

The end goal of EHDEN and also of this mini-project around the OHDSI COVID-19 study-a-thon is to create reusable frameworks for making (research in) healthcare data FAIR for the benefit of medical practice. This will ensure that once we have the next study-a-thon, there is an easy-to-use framework to quickly instantiate a website that covers both goals described in the introduction. Future work also needs to happen to make for example the generation of persistent identifiers easier, especially for auxiliary digital assets such as study protocols, cohort definitions, interactive study results etc. that are not scientific papers or pre-prints. Providing these services is the goal of projects such as EOSC.

6.1.2 Technical insights

During the creation of this pilot, we encountered a number of problems that are illustrative for why it is useful to strive for FAIR data and the usage of data standards. These are detailed here as potential input for the EHDEN portal and the implementation of above proposed standards.

The static website generator that we used to build the COVID-19 study-a-thon showcase website, Hugo has a templating system using “archetypes”. These are blueprints used when creating new content. Archetypes contain preconfigured front matter, including all metadata needed for the website. The editor of the website can create content for each post based on an archetype and fill in them. However, this method is completely manual, and our goal was to write a script to automatically collect data by scraping the source website so that it reads the content of the original pages and extracts the information needed for our website. In this particular case, the original website is the readme page of OHDSI Studies’ repositories in Github. The script has been written specifically for this website according to the structure of the original pages.

While testing and optimizing this script, there are problems in extracting information from the original pages, which mentioned below:
1) Some data are linked together, but in original pages, these linked attributes are not specified. For example, the link between a study author and his/her OHDSI forum name is missed.
2) The same format is not used for all dates. For example, we encountered different date formats such as Dec 16, January 11, July 7th, Mid June, etc.
3) There are typos in writing markdown elements which break the scraping process. For example, links should be in the format of [link title](url) but we got [[link title]]((url)) instead.
4) The description part of the original pages is not separated from other parts specifically.

The above are some of the problems for scraping the original pages which cause more tests and evaluations. Other properties needed for studies which couldn’t be fetched from Github readme files are in the table below.

Property	Extractable from the readme file	Comments
Title	Yes	
Tags	Yes	
Identifier	No	

Study use cases	Yes	Predefined study use cases on the website.
Authors	Yes/No	OHDSI forum name or name (both of them couldn't be extracted)
Description	Yes/No	If the description is separated from other parts of the readme file specifically.
Discussion URL	No	
Has Parts (Software Source Codes and Software Applications)	No	
Health Conditions	No	
Main Entity of Page	Yes	Url of Github repository
Publications	Yes	
Results	Yes	
Study Status	Yes	
Study Design	No	
Drugs	No	
Event	No	
Start Date	Yes	The date format (YYY-MM-DD) should be preserved
End Date	Yes	The date format (YYY-MM-DD) should be preserved

These learnings can be incorporated into any semantic-aware versions of the EHDEN portal and also for other means of disseminating metadata for studies, for example in the OHDSI github directly by adding a JSON-LD metadata file rather than a textual readme file.

7. Conclusion / Roadmap Summary

This document contains a large number of lessons learned and potential interoperability touchpoints for EHDEN with other biomedical standards and initiatives. The goal of this chapter is to propose a few main action points, which we can consider for interoperability in the coming years of the project.

1) Implement structured metadata for at least databases and studies in the EHDEN Portal. The proposed JSON-LD standard that we developed in preparation of this deliverable can be used as a blueprint for this.
2) Either adopt a community project (e.g. OMOPonFHIR, Miracum etc.) or start an EHDEN project for mapping existing FHIR resources to OMOP. Ideally this is guided by a concrete use case. This could be a

(potential) EHDEN Data Partner which already implemented FHIR and is interested in converting FHIR Resources to OMOP for the purpose of participating in network studies. Alternatively, it could concern a regulatory or governance use case where FHIR (meta)data and/or process flow definitions are used to accommodate specific requirements in the chain of evidence.

3) Especially for potential prediction models developed in EHDEN, explore the potential of deploying these directly on FHIR to facilitate clinical decision making in hospitals and clinics. From EHDEN, we can facilitate this technically, but it is important that a clinical organization is in the lead for this, and we could explore potential synergies with for example IMI H2O for this purpose.

4) Monitor (and perhaps inquire?) with (potential) EHDEN data partners which data standards and interoperability solutions they currently use and which could be leveraged for lowering the barriers to OMOP mapping. Very likely, these will be one of the standards already described in this document, and the corresponding sections offer insights in technical details and potential steps that can be taken to take advantage of these standards for the purpose of OMOP mapping.

ANNEXES

Annex 1: FAIR Assessment of OHDSI studies and databases

Source databases

Study-a-thon data sources

This is a FAIR assessment of the representation of reported data sources on page 4 of the supplemental appendix of the HCQ study, i.e. “safety of hydroxychloroquine, alone or in combination with azithromycin, in light of rapid wide-spread use for COVID-19: a multinational, network cohort and self-controlled case series study”.

Overall score: 5/30

Principle	Assessment	How to improve	Score
F1	Databases do not have unique and persistent identifiers. Databases are only represented by their free-text name and abbreviation. Metadata is all free-text and is not associated with any identifiers.	Assign to each database its own unique and persistent identifier. Such an identifier could be the URL to the database profile. Using metadata concepts from ontologies such as Provo-O, DCAT or DC Terms, would at least provide the metadata fields with identifiers that resolve to pages with further information on the concept.	0
F2	Each database has the following metadata: • Name • Abbreviation • Population • Patients (number) • Data history (date range) • Data capture process and short database description	Aside from name and abbreviation, there are only 4 fields that provide some metadata on the database. Metadata on the following concepts could be included to improve F2: • Population characteristics (age, gender, specific indications, etc.) • Vocabularies and languages • Data format • Contact person(s)	0
F3	A database’s metadata is all captured in the same row, but identifiers are not used to link metadata to the database it describes.	Once a database profile is created, metadata can then reference the database profile ID.	1
F4	Database information is captured in a table as free text. Aside from a text-search within the document, this data is not indexed and searchable.	The database profile should also be machine-readable and searchable. Using common ontologies and vocabularies can help with this process.	0
A1	Neither the databases nor the metadata can be retrieved as they do not have an associated identifier and are not indexed.	Databases and metadata should have associated IDs.	0
A1.1	The data sources table cannot be accessed through HTTP.	Creating a database profile accessible through HTTP would improve this score.	0
A1.2	There is no protocol in place to access the data.	If needed, access to database profiles could be limited to a community (as is the case in the EMIF catalogue). Although I imagine keeping open access is more in line with the spirit of OHDSI.	0
A2	The data sources table persists even if the	The database profile should persist even if the database is no	2

	databases don't exist anymore.	longer available.	
I1	Metadata is in the form of free text written in English. This is human-readable, but only to English speakers, and furthermore it is not machine-readable.	A more generic knowledge representation might benefit the community by encouraging a greater number of users to contribute data. Examples of some broadly-applicable knowledge representations are RDF, OWL and JSON-LD, as well as including options in multiple languages.	0
I2	There is no documentation on the metadata "vocabulary". The metadata itself is not FAIR, as it is not machine-readable and is not findable through a search engine. Additionally, there are no descriptions of the metadata.	Using a common ontology to represent the metadata could already improve the score. Metadata concepts should be represented in a machine-readable format (see principle I1) and there should be some documentation on each concept. For example, is 'Population' referring to geographical location or selection criteria or both?	0
I3	Data does not include references to other data or metadata.	Linking metadata concepts out to existing ontologies would improve this score.	0
R1	Metadata fields are free text, which means some entries are not consistent (see 'Population'). In addition, two metadata entries are captured as one in the 'Data capture process and short database description'. All metadata fields are complete for each database.	Adding more metadata concepts to each database would increase the richness of the database description. These concepts should have a clear description, and where applicable, a list of controlled vocabulary terms a user can choose from.	1
R1.1	No usage licence is provided.	Such information could be provided for each database profile.	0
R1.2	Provenance information for the database or the metadata is not provided. The study-a-thon will include participants and authors, although if these can be considered as the 'provenance information' is dubious.	For each database include a person of contact. Additionally, include provenance for the metadata, detailing who filled in the metadata and when.	0
R1.3	Metadata are not linked to standard ontologies, so this is hard to say. However, the data sources were gathered and described by the OHDSI community that participated in the study-a-thon, so there is community-backing.		1

EMIF Catalogue

This is a high level assessment of database profiles as reported in the EMIF Catalogue (the catalogue in the EHDEN Portal contained only test data at the time of assessment, but is built on the same foundation). This assessment is not done with respect to a particular profile but as a general assessment of the EMIF Catalogue's database profile capability.

A database in EMIF has the following information, divided into 4 tabs:

- Fingerprint (this contains the metadata of the database)
- Discussion (EMIF users can post questions or comments relating to the database)
- Suitability
- Literature (list of publications showing how data has contributed to research)

The database fingerprint contains the following information categories:

1. Database General Information (x/27)
2. Database characteristics (x/10)
3. Database population (x/11)
4. Data held (x/35)
5. Ability to gather supplementary data and biological samples (x/4)
6. Ethical issues (x/9)
7. De-identification, linkage (x/9)
8. Data set description, metadata (x/12)
9. Examples of accessing events and outcomes (x/66)
10. Documents and publications (x/3)
11. Comments (x/1)
12. EMIF database fingerprint provenance (x/6)

For each category there is a subset of information to be filled in by the user. The completeness of each category is expressed as a fraction of filled entries over total possible entries. This is also expressed as a percentage.

Overall score: 17/30

Principle	Assessment	How to improve	Score
F1	Each database profile is assigned a URL. See this example. Metadata (i.e. data that is part of the database fingerprint) does not have an identifier.	Metadata could be assigned an ID which links out to a resource with a definition of the metadata term. Where possible, metadata concepts could be linked to common ontologies. E.g. DCAT or Dublin Core could be used for concepts such as, title and contact.	1
F2	The database profile allows the user to describe the database with extensive metadata. The data collected for a database profile is rich and well-specified. Such data can include: • Generic information on database name, country, contact and population • Frequency of database updates • Qualified information on the population captured in the database • Vocabularies used in database	Same as F1 -- the profile could benefit from reusing metadata concepts from common and standard ontologies to eliminate any potential confusion.	1
F3	Metadata are collected as part of a database's profile, and so are directly associated with the database.		2
F4	Metadata of the database is collected via multiple choice questions, yes/no questions and free text descriptions. The EMIF catalogue has a free text search function. Within the EHR catalogue, a user can filter the entries based on quantitative data such as population size, cohort start date and last update.	Fuzzy matching is not enabled for the search. Although not necessary, it would enhance the search capabilities. Especially given the free-text nature of most of the metadata entries, which leaves room for inconsistencies. Furthermore, resources cannot be found outside of the EMIF environment. I.e. a database registered and profiled in EMIF does not appear in a Google search.	1

A1	Database profiles can be accessed through a URL, but the database's metadata does not have an associated ID.	Metadata could be assigned an ID which links out to a resource with a definition of the metadata term and potential links to similar terms in other ontologies.	1
A1.1	The database profile can be accessed through HTTP.		2
A1.2	Registering with email and password is necessary to access the EMIF Catalogue. Communities within EMIF require special access, which a user can request directly on the portal.		2
A2	The database profile exists independently of the database itself. Thus, the metadata persists.		2
I1	Metadata is in the form of free text questions written in English. This is human-readable, but only to English speakers, and furthermore it is not machine-readable.	Many of the databases in EMIF are from non-English speaking countries. A more generic knowledge representation might benefit the community by encouraging a greater number of users to contribute data. Examples of some broadly-applicable knowledge representations are RDF, OWL and JSON-LD.	0
I2	The vocabulary to describe the databases does not assign IDs to its terms. Documentation on the metadata can be found by hovering over entries, but this is not machine-readable.	The controlled vocabulary used needs to be documented and resolvable using globally unique and persistent identifiers. This documentation needs to be machine readable and easily findable and accessible by anyone who uses EMIF.	1
I3	Data does not include references to other data or metadata.		0
R1	The metadata collected for a database profile is rich and well-specified. Descriptions of the metadata collected appear when hovering over entries in the database fingerprint, providing further clarity to the user what type of information should be filled. However, many metadata fields are free text, which may lead to poor data if the user misinterprets the question, or leaves an answer that is hard to interpret. In addition, most fields are not required, meaning a database profile can contain very little data. Users can leave comments on particular metadata entries, but the response rate from the database contact may vary.	Data stewards could be assigned to review database entries to implement quality control and ensure entries are as complete as possible, and as clear as possible.	1
R1.1	A usage license is not provided but for the EMIF EHR catalogue information on the code of practice can be found in the 'Documentation' tab in the side panel.	Information on the re-use of data reported in the catalogue could be presented more clearly to the user. For example, specifying a data license in the footer of the database profile would be beneficial. Furthermore, the information in the code of practice could be made machine readable.	0
R1.2	Each database fingerprint contains a section on 'EMIF database fingerprint provenance' and specifies the contact details of the person responsible for filling in the questionnaire. Under the 'All' section it's possible to see when a		2

	database profile was last updated.		
R1.3	Metadata categories are not linked to external vocabularies, so it is not easy to say whether this meets the domain-relevant community standard. However, it has to be said that EMIF is a joint multidisciplinary research project that spans many communities. One could make the assumption that such a project would not be as widely adopted if it did not, to some extent, meet community standards.		1

Studies

The following elements are related to a study:

- Study protocol
- Source code
- Study results
- Publication(s)
- Researchers/Authors
- Study-a-thon (event)

OHDSI studies can be found at https://data.ohdsi.org/. Here an overview of all the studies with some metadata can be found: https://data.ohdsi.org/OhdsiStudies/.

This assessment is a high level FAIR assessment of the dissemination of studies in OHDSI. We take the Covid19EstimationHydroxychloroquine study, one of the studies performed during the COVID-19 study-a-thon, as an example. Information about this study can be mainly found in the GitHub repository: https://data.ohdsi.org/Covid19EstimationHydroxychloroquine/. Study related resources can be found as follows:

- ENCePP registration: http://www.encepp.eu/encepp/viewResource.htm?id=34498
- Study protocol: https://github.com/ohdsi-studies/Covid19EstimationHydroxychloroquine/blob/master/documents/OHDSI%20COVID-19%20Studyathon_PLE_HCQ_Protocol_v1.4.pdf
- Source code: https://github.com/ohdsi-studies/Covid19EstimationHydroxychloroquine
- Study results: https://data.ohdsi.org/Covid19EstimationHydroxychloroquine/
- Researchers/Authors: https://github.com/ohdsi-studies/Covid19EstimationHydroxychloroquine/blob/master/documents/OHDSI%20COVID-19%20Studyathon_PLE_HCQ_Protocol_v1.4.pdf (part of study protocol)

The following aspects are currently lacking in studies:

- Identifiers, e.g. a study identifier, protocol identifier
- Rich, findable and structured metadata:
    - General, structured and comprehensive information about source databases.
    - Metadata about and references to OMOP CDM version, vocabulary version, version of software used.
    - Author metadata (currently buried in protocols, which is not findable)
    - Date of study dissemination
- In the case of the Covid19EstimationHydroxychloroquine study, a lot of this metadata can be found at the ENCePP registration, but this is not the case for all OHDSI studies.
- Licenses

Overall score: 11/30

Principle	Assessment	How to improve	Score
F1	Studies and related protocols, source code, results, publications, researchers and study-a-thons do not have identifiers assigned.	Assign studies as well as all related resources globally unique and persistent identifiers. (identifiers could be created using https://www.doi.org/ or similar services)	0
F2	The following data about a study is present: description of the study, study protocol, source code, results. This data is not comprehensive. For example, relevant metadata about researchers involved, date of dissemination, source data and OMOP CDM database information is missing. This metadata is present when diving deeper in the associated files.	Capture metadata in a structured, clear and comprehensive way. Add metadata on researchers, dates, source data, protocols, results, publications, OMOP CDM database (version), a potentially related study-a-thon event.	1
F3	As there are no identifiers assigned, this is also not included in the metadata.	Create identifiers for all different study elements, and include these in the metadata.	0
F4	Study related information is disseminated at data.ohdsi.org. Data.ohdsi.org is not searchable, i.e. no keywords could be used to search for certain studies and associated metadata. https://data.ohdsi.org/OhdsiStudies is searchable.	Add rich metadata to data.ohdsi.org, or register studies at other repositories, such as http://www.encepp.eu/.	1
A1	Since no identifiers are associated, (meta)data is not retrievable by their identifier.	Associate globally unique and persistent identifiers	0
A1.1	All study-related elements are accessible through HTTP.		2
A1.2	No authentication or authorization is needed.	Since OHDSI is an open community, this is not necessary. Although the names of researchers and their email addresses are freely accessible as well (in the protocol). That kind of information could be privacy sensitive and might need an authentication/authorization layer.	2
A2	There is no protocol available that describes longevity of the data/metadata.	Provide a protocol ensuring longevity of data/metadata.	0
I1	Since controlled, standardized vocabularies are part of the OMOP CDM, study results	Use JSON-LD schema.org for metadata on protocol, source code, source data, study results, researchers, etc.	1

	automatically leverage controlled vocabularies. Other (metadata) elements related to a study (e.g. protocol, source data, source code) do not make use of a formal language for knowledge representation.		
I2	Since the OMOP CDM has standardized vocabularies incorporated, the datasets and results are described by unique and persistent identifiers (i.e. concept_ids).	None.	2
I3	Within study metadata, there are references to the protocol, source code and publication.	Additional references could be included, i.e. references to the source data, OMOP CDM version, standardized vocabulary version, software versions, etc.	1
R1	Metadata is present in a non-standardized manner and is not complete.	Metadata that could be included: date of generation of the study, who prepared the data, name and version of the OMOP CDM + standardized vocabularies + software used, etc.	0
R1.1	There is only a licence available for the study package (Apache License 2.0)	Release studies with a data usage license, such as the creative commons 4.0 license.	1

R1.2	No (detailed) provenance information is present.	Standardized and machine-readable (meta)data about provenance should consistently be given, when disseminating a study.
		One could think about:
		- origin of data
		- who generated/collected the data
		  - who generated the OMOP CDM mappings
		  - who performed the study
		- when did that happen
			0

R1.3	See I1. A FAIR way of documenting study metadata is currently not happening within OHDSI.	Use JSON-LD/schema.org to document metadata in a standardized way.	0