FAIRification of observational studies and databases:   
the EHDEN and OHDSI use case

# Main objectives

In this recipe, an example will be described on how to make observational health databases and observational studies more Findable, by creating a metadata model and a human and machine readable website for dissemination of these resources.

The output of the COVID-19 study-a-thon (a OHDSI community initiative, involving the IMI EHDEN project) held in March 2020, is used as proof-of-concept.

# Abbreviations

* EHDEN: European Health Data & Evidence Network
* ETL: Extract Transform Load
* JSON-LD: JSON - Linked Data
* OHDSI: Observational Health Data Sciences and Informatics
* OMOP CDM: Observational Medical Outcomes Partnership Common Data Model
* RDF: Resource Description Framework
* YAML: YAML Ain’t Markup Language (human-readable data-serialization language)

# Introduction

This recipe is based on work done on OMOP Common Data Model data as used during the Covid 19 Study-a-thon, a OHDSI community initiative, organised in collaboration with the IMI EHDEN project. The background information was previously described in this [use case](https://fairtoolkit.pistoiaalliance.org/use-cases/fairifying-collaborative-research-on-real-world-data-the-hyve/) on the Pistoia Alliance FAIR toolkit website, and also published in an EHDEN Deliverable, [D4.5 - Roadmap for interoperability](https://zenodo.org/record/4474373). Briefly, the IMI2 EHDEN project's main aim is to reduce the time to produce medical evidence for health outcomes research and to do so, EHDEN has prioritized data harmonization at source.

However, data standardization and harmonization does not necessarily equate with FAIR data (see for example this [article](https://www.thehyve.nl/articles/data-quality-with-fair-principles)). This is where the interaction with the IMI2 FAIRplus project can bring benefits.

The purpose of the present content is to highlight how to improve dataset Findability after the data harmonization has taken place.

The methods described in this recipe are however generally applicable for observational health databases and studies, not just data in the OMOP CDM format.

In March 2020 the OHDSI community organized a [study-a-thon focussed on COVID-19](https://www.ohdsi.org/ohdsi-news-updates/covid19-studyathon/), which was aimed at addressing a number of key medical questions around COVID-19, such as predicting hospitalization and drug safety and effectiveness. For this purpose, a large collection of healthcare databases around the world was investigated. Many important digital assets were produced during the [study-a-thon](https://ohdsi.org/ohdsi-kicks-off-international-collaborative-to-generate-real-world-evidence-on-covid-19-with-virtual-study-a-thon/), such as study protocols, database metadata and characterizations, study results and publications.

From the FAIR assessment performed in order to better understand the current state of FAIRness of studies and databases used in OHDSI and in the COVID-19 study-a-thon in particular, it was found that there was room for improvement on both Findability and Reusability. Since increasing Findability was a common denominator for both studies and data sources, the main focus of the FAIRification was on developing structured metadata, specifically tailored for studies and data sources

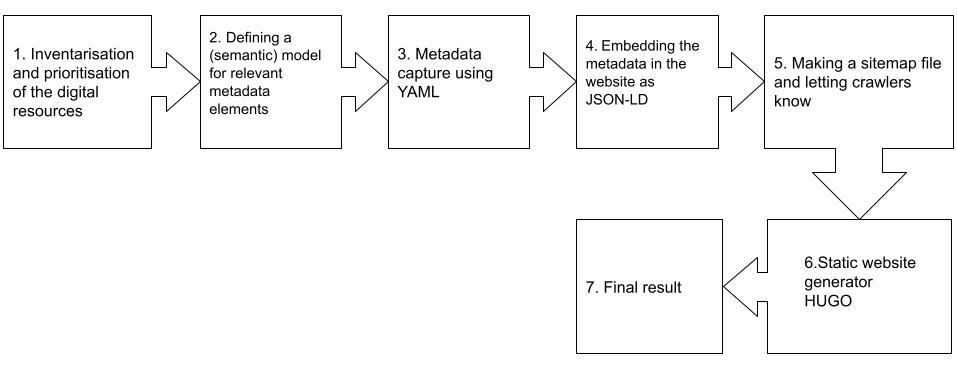
This recipe will describe a proof-of-concept that shows the step-by-step process of making resources more FAIR:

* inventorization and prioritisation of digital resources
* adding metadata to the resources using existing ontologies and defining new entities to make them findable by humans and machines
* creating a website for FAIR dissemination of selected digital resources.

# Ingredients

* Data formats
  + RDF
    - Turtle
    - JSON-LD
  + YAML
* RDF format validators:
  + [Apache Jena RDF validator](https://jena.apache.org/)
  + [Rich Results Test for JSON-LD](https://search.google.com/test/rich-results)
* Databases:
  + [EMIF Catalogue](https://emif-catalogue.eu/)
* Vocabularies
  + [Schema.org](https://schema.org/)
  + [Custom OHDSI extension](https://github.com/thehyve/ohdsi-schemas/blob/master/models/ohdsi_model/ohdsi_semantic_model.ttl) to schema.org
* Static website generator
  + [HUGO](https://gohugo.io/)

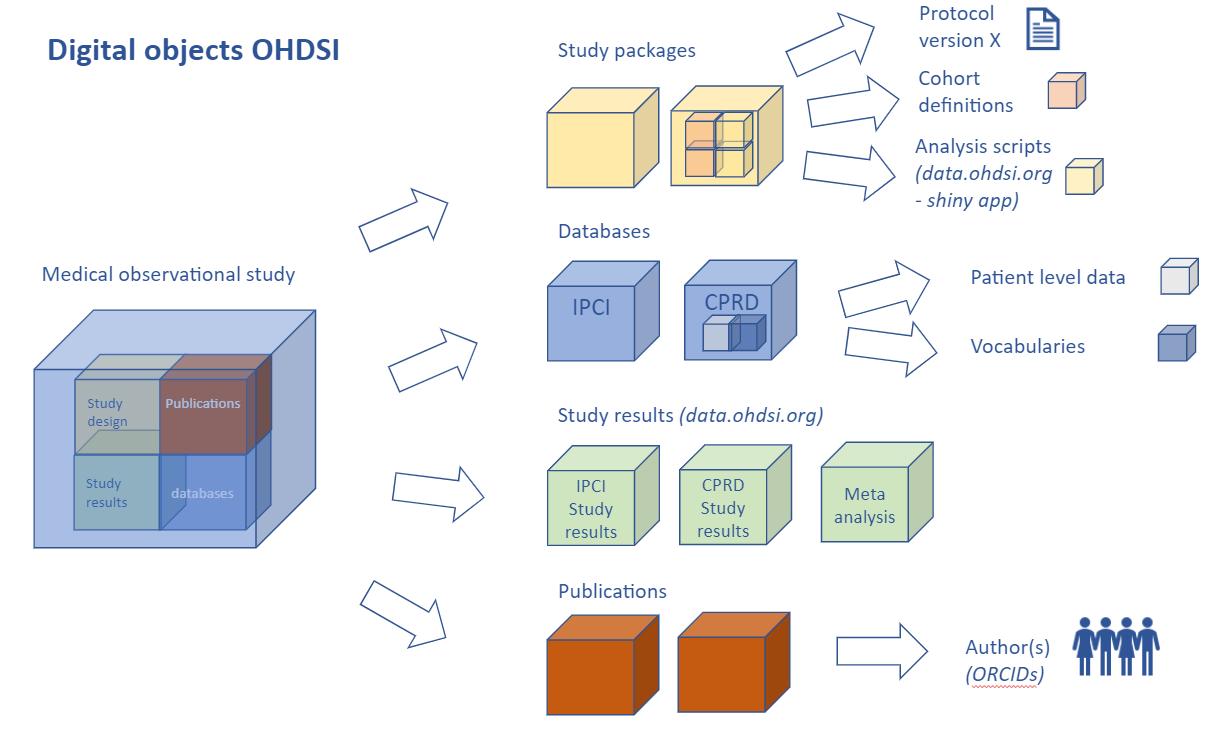
# Graphical Overview of the FAIRification Recipe Objectives



# Step-by-Step process

## Inventorization and prioritization of digital resources

In order to start the FAIRification process of digital resources in OHDSI, it was important to inventorize which kind of digital resources reside in the OHDSI ecosystem. Digital resources in OHDSI include for example the OMOP CDM itself, the OMOP Standardized Vocabularies, studies, cohort definitions and many more. Each digital resource itself can consist of multiple (sub) digital resources, as shown in figure 1.



**Figure 1.** Example of digital resources that are part of an OHDSI study (from [EHDEN D4.5](https://zenodo.org/record/4474373))

Prioritization of initial digital resources to FAIRify was important, as the OHDSI ecosystem is quite extensive. In light of the study-a-thon on COVID-19 held in March 2020 by OHDSI and after a [community inventorization](https://forums.ohdsi.org/t/implementing-the-fair-principles-in-the-ohdsi-approach-and-tools/10387), studies and databases were selected as relevant digital resources to FAIRify, as these are important resources for the researchers studying observational health data on COVID-19 and related topics.

### Current status of database and study dissemination in OHDSI

**Databases in OHDSI**

General sources that provide information on observational health databases that are conformed to the OMOP CDM are the [data network page](https://www.ohdsi.org/web/wiki/doku.php?id=resources:2020_data_network), the [EMIF catalogue](https://emif-catalogue.eu/) and the [EHDEN Portal](http://portal.ehden.eu) (under development).

The data network page of OHDSI shows some metadata on databases converted to the OMOP CDM.

The [EMIF catalogue](https://emif-catalogue.eu/) is developed by the European Medical Information Framework (EMIF) and enables publishing and sharing of metadata on health data sources. Different consortia implementing the OMOP CDM disseminate metadata on databases in this catalogue. The EHDEN Portal builds on technology from the EMIF catalogue and further expands this with functionality to compare databases and run studies in the network.

**Studies in OHDSI**

Studies are currently disseminated through [data.ohdsi.org](https://data.ohdsi.org/). This web page links to the study results and [GitHub repositories of studies](https://github.com/ohdsi-studies) that are performed in the OHDSI community. The web application [data.ohdsi.org/OhdsiStudies](https://data.ohdsi.org/OhdsiStudies/) gives an overview of all studies in OHDSI with some associated metadata, such as study status, study leads and relevant dates.

## Defining a (semantic) model for relevant metadata elements

To increase the findability we captured the structured and rich metadata, by developing a semantic model in RDF for the observational studies and databases. The model contained all relevant metadata elements for studies, databases, as well as persons associated with those studies and databases.

Modelling metadata elements in RDF has the advantage that the metadata elements are properly semantically described, structured and linked. In addition, RDF is easily expandable and modifiable.

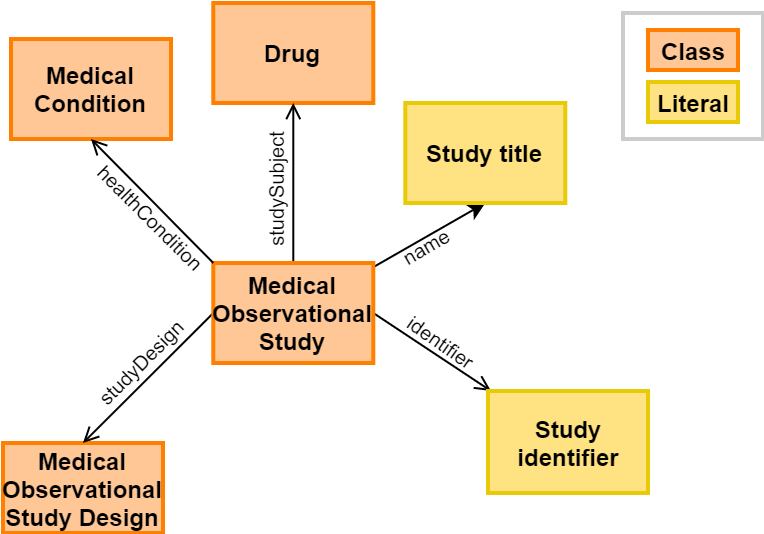
The Turtle serialization format was chosen as RDF format. Metadata elements were defined and mapped to [schema.org](http://schema.org) types and properties. Schema.org is a vocabulary that can be used to create structured data on the internet and to make the data discoverable.

Relevant schema.org types for studies and databases included:

* [MedicalObservationalStudy](https://schema.org/MedicalObservationalStudy)
* [MedicalObservationalStudyDesign](https://schema.org/MedicalObservationalStudyDesign)
* [MedicalStudyStatus](https://schema.org/MedicalStudyStatus)
* [MedicalCondition](https://schema.org/MedicalCondition)
* [Drug](https://schema.org/Drug)
* [Organization](https://schema.org/Organization)
* [Person](https://schema.org/Person)

Figure 2 shows part of the relevant metadata elements in the model.

**A.**

****

**B.**

schema:MedicalObservationalStudy a owl:Class ;

rdfs:isDefinedBy <https://health-lifesci.schema.org/MedicalObservationalStudy> ;

rdfs:label "MedicalObservationalStudy"@en ;

rdfs:subClassOf schema:MedicalStudy .

schema:Drug a owl:Class ;

rdfs:isDefinedBy <https://health-lifesci.schema.org/Drug> ;

rdfs:label "Drug"@en ;

rdfs:subClassOf schema:Substance .

**Figure 2. A.** Part of the metadata model. Here, each class (orange) and relationship is represented as a schema.org concept. Literals (yellow) are data types, e.g integers or strings.   
**B.** RDF representation of two schema.org types, written in Turtle.

### OHDSI extension to schema.org vocabulary

[Schema.org](https://schema.org/) is a general purpose vocabulary for searching entities on the internet, so metadata elements specific for some OHDSI assets were not present. To be able to capture metadata as extensive as necessary, custom types and properties had to be created. These included:

* properties specific for OHDSI studies(e.g. study types and study use cases)
* database properties (e.g. database characteristics and population size)
* author properties (e.g. GitHub handle and forum profile).

For OHDSI specific types and properties the namespace <http://data.ohdsi.org/> is used, using the “ohdsi” prefix.

The RDF model was validated using Apache Jena RIOT (version 3.15.0), using the following command:

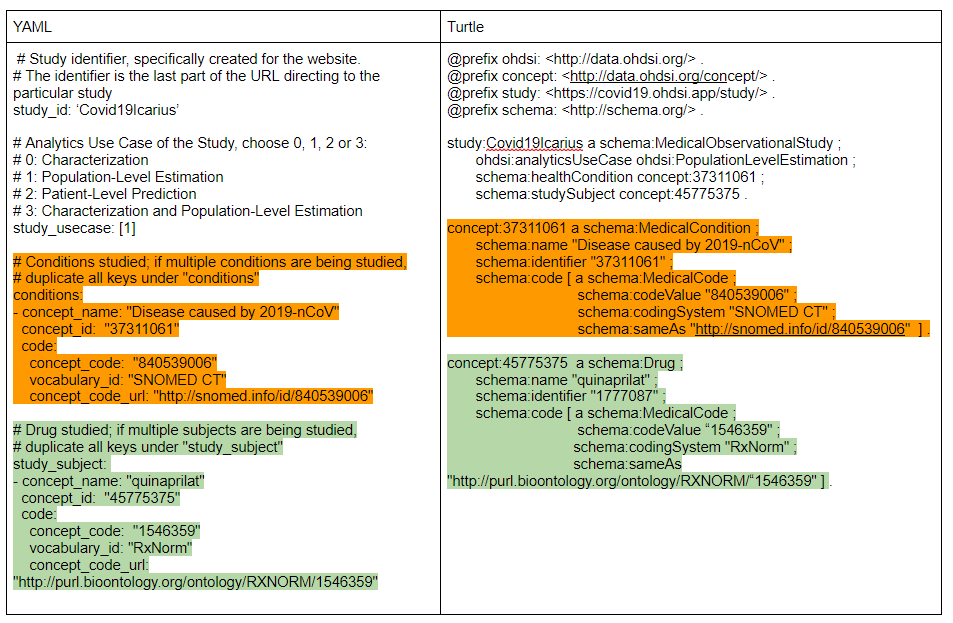
riot.bat --validate ohdsi\_semantic\_model.ttl

The metadata model is shared in a public [GitHub repository](https://github.com/thehyve/ohdsi-schemas/blob/master/models/ohdsi_model/ohdsi_semantic_model.ttl) [8].

## Metadata capture using YAML

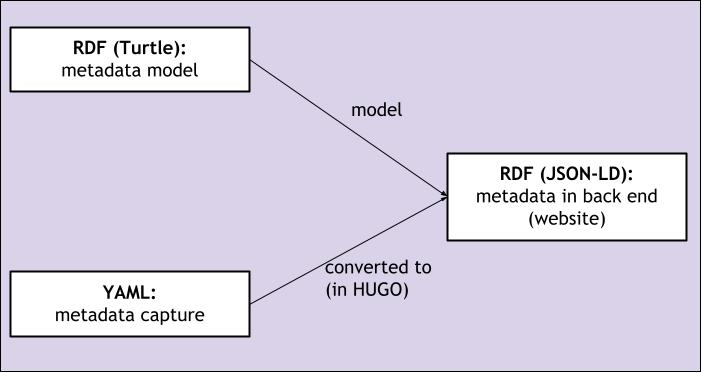
In principle, the RDF classes defined in the above Turtle file can directly be used to set up the JSON-LD file needed for the website metadata. But in this example, we used an extra step by first capturing the metadata of studies, databases and authors, in the more data entry friendly YAML format. The YAML format can be annotated with comments to guide the person submitting their metadata.

Figure 3 shows the metadata in Turtle and in YAML. YAML uses slightly different wording and syntax, to make it easier for end users / science bloggers to understand what kind of metadata is expected to fill in.



**Figure 3.** Left: Example Turtle representation of metadata instantiation of the OHDSI Covid19Icarius study. Right: YAML representation of the same metadata as displayed on the left.

The Hugo backend that we created, reads in the study articles, including the YAML metadata, and produces a HTML output from that which contains the article text but also embeds the entered metadata as JSON-LD (see figure 4 and step 4). In order to do this, some custom processing logic is written that takes the YAML fields as input and formats this into proper RDF in accordance with the metadata model.

****

**Figure 4.** Relationship between the use of Turtle, JSON-LD and YAML in this proof-of-concept project.

## Embedding the metadata in the website as JSON-LD

JSON-LD is a format for structured and linked data and, like Turtle, a serialization format of RDF.

It can be embedded in the HTML code in the back end of a website. Search engines like Google (including Google dataset search) use this structured data to find and understand the data. Understanding the data goes hand in hand with using the [schema.org](http://schema.org) vocabulary, because annotating metadata with schema.org types and properties gives meaning to the data. This search engine optimization (SEO, see the [Search Engine optimization](https://fairplus.github.io/the-fair-cookbook/content/recipes/findability/seo.html) recipe) enables better Findability of the data.   
  
 For this proof-of-concept project, a website was created for dissemination of metadata on studies, databases and authors, generated and used during the OHDSI COVID-19 study-a-thon. Metadata on studies, databases and authors is displayed on this [website](https://covid19.ohdsi.app/), and is embedded as JSON-LD in the HTML code as well.

Embedding metadata in JSON-LD in HTML will enable web crawlers to scrape the metadata and will increase Findability by search engines. When for example querying <https://datasetsearch.research.google.com/>, the data should pop up when JSON-LD is embedded.

## Making a sitemap file and letting crawlers know

Setting up a robots.txt file at the root of a website will give search engine crawlers a guide on which web pages can be crawled, and which pages can’t be crawled. Crawling all pages is possible by adding this line to the robots.txt file:

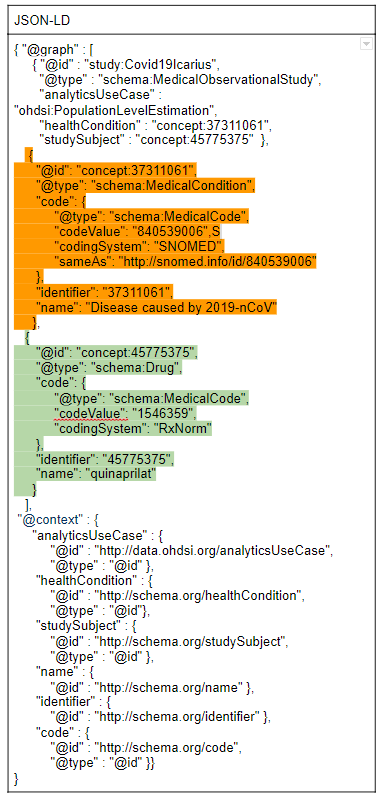
User-agent: \*

The Hugo framework automatically generates a sitemap when building a website (in this example: <https://covid19.ohdsi.app/sitemap.xml>)

Figure 5 shows part of the model instantiated with metadata for a particular COVID-19 study in JSON-LD format. (This matches the metadata shown in figure 3). Figure 4 shows the relationship between the Turtle, YAML and JSON-LD files.

The generated JSON-LD files can be validated using [Google’s Rich Result Test](https://search.google.com/test/rich-results). This tool can easily validate JSON-LD code, by either submitting a URL directing to the JSON-LD, or by pasting JSON-LD code.

An important note to make here is that to make the data findable for Google Dataset search the model and final JSON-LD markup needed to mention the “schema:Dataset” class. In a first iteration, we had used our new “ohdsi:database” class and the website was not found (even if the ohdsi:database is-a schema:Dataset statement from the model is repeated in the JSON-LD, so the Google Dataset search doesn’t seem to use inference to determine this).



**Figure 5.** Example JSON-LD representation of metadata instantiation of the OHDSI Covid19Icarius study.

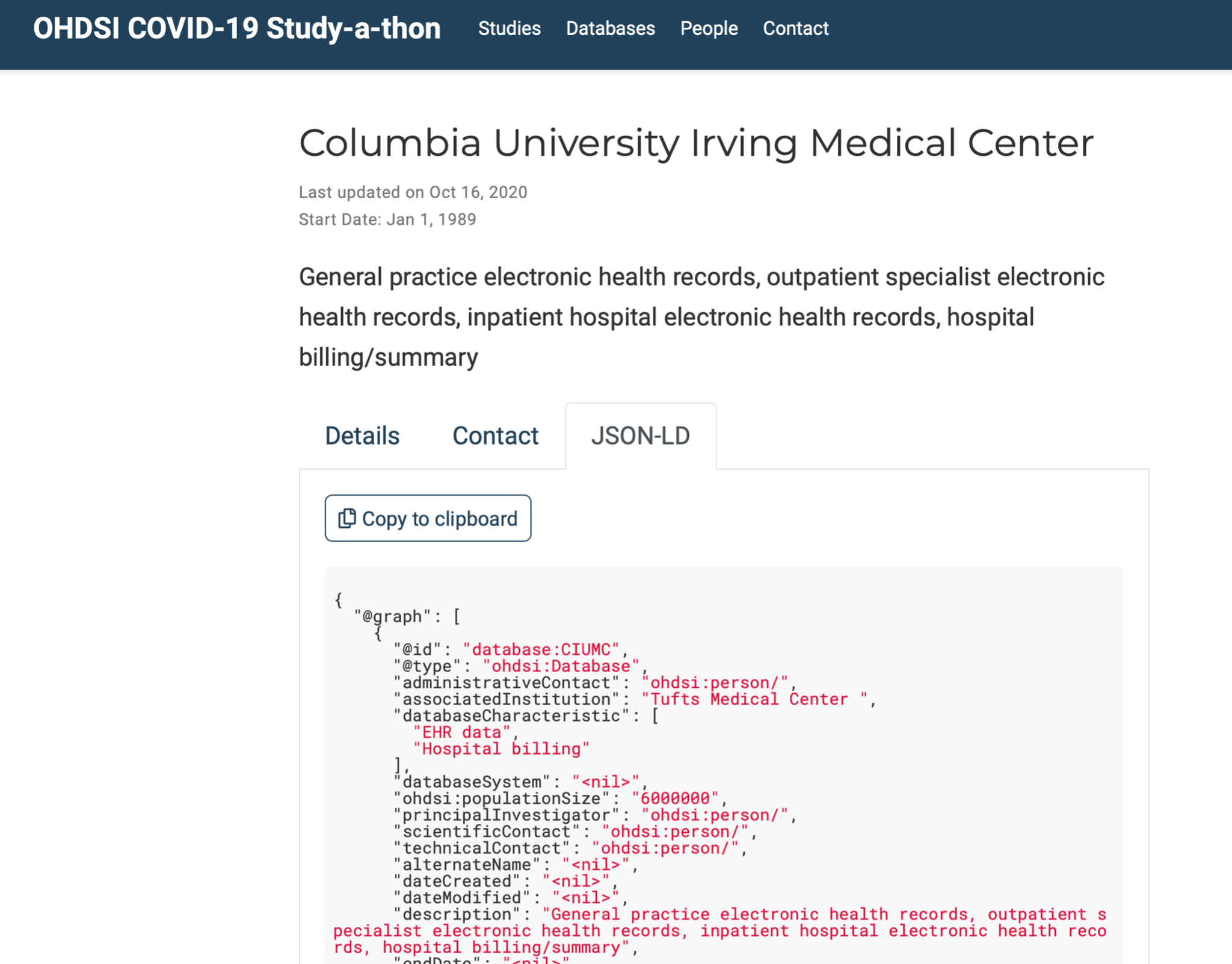
## Static website generator HUGO

The static site generator [HUGO](https://gohugo.io/) was used to create the [covid19.ohdsi.app](https://covid19.ohdsi.app/) website. To populate the website with metadata, so-called [archetypes](https://gohugo.io/content-management/archetypes/) were created. These archetypes are content templates that use the YAML format, as the JSON-LD format is not the easiest format to enter data, to manually populate the static website pages with metadata. Hence, these archetype templates are used for metadata instantization of the website (see chapter [3](#_h4wm4s0j5v2) and [4](#_5yp2oql1egdz) of this recipe). Part of the content is populated automatically, by scraping information from the [OHDSI forum](https://forums.ohdsi.org/) and [README files](https://github.com/ohdsi-studies/StudyRepoTemplate) in the study-specific GitHub repositories. The metadata in YAML is converted to JSON-LD in the back end of the website. In the front end of the website, this meta data is shown in a human-readable way.

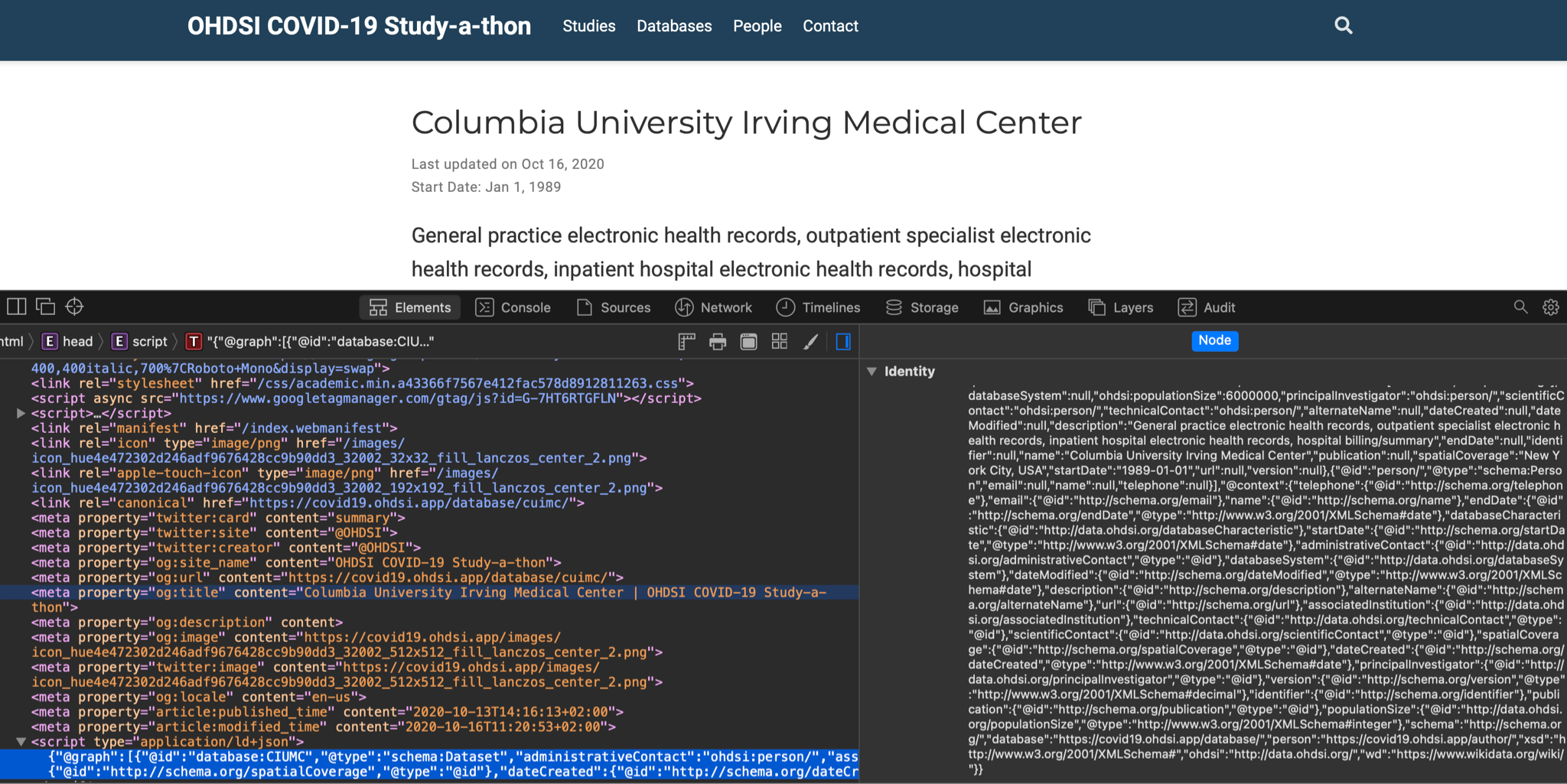
The source code of the website can be found [on Github](https://github.com/thehyve/ohdsi-covid19-site).

## Final result

The final result of applying all these steps looks like this, as HTML page in the browser with JSON-LD metadata ‘under the hood:



**A.**

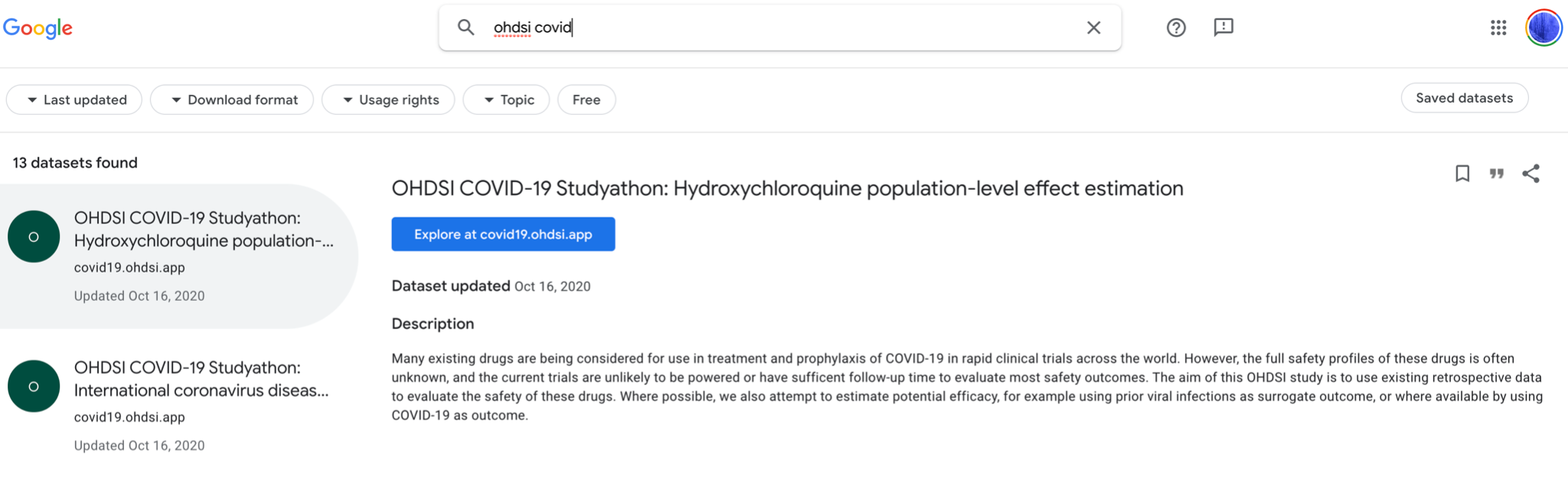


**B.**

**Figure 6A.** The top panel of the figure shows the ‘database’ view of the Hugo covid19 app, which offers the ability for users to download the JSON-LD for the records.

**Figure 6B.** The lower panel of the figure on the other hand shows how the HTML head element is augmented with a script element which allows embedding the same JSON-LD information payload and make it available to search engines’ crawlers (more on this in the following section).

Finally, with all this in place, one may now use the dedicated Google Dataset Search site to discover new covid related datasets, and we will find the dataset from this website in the results:



**Figure 7**. The OHDSI covid19 datasets as viewed from Google Dataset Search

# Conclusion

The key reusable element in this recipe is the use of schema.org JSON-LD statements in any website to promote findability of metadata in search engines. Using the Hugo site generator has the advantage that it is very versatile and includes a lot of scientific/blogging authoring tools out of the box. However, the particular implementation of using YAML as input format (which is ‘canonical Hugo’) and converting this during the static website generation process to JSON-LD is quite laborious and also potentially introduces issues during the conversion. Therefore, it is probably easier and more scalable to let the scientists write their metadata directly in JSON-LD, helped by some good examples and templates, and the Hugo software could even be extended to facilitate that as well.

Another next step that could be taken, besides publishing the JSON-LD metadata directly on the web, is to wrap it into a container such as [RO-CRATE](https://www.researchobject.org/ro-crate/) and publish these directly into a scientific artefacts repository. This would also help to sustain the availability of the data, because for all it’s great properties, static websites are only live as long as the webhost stays up, and the average website disappears after a few years. Hosting the website on [IPFS](https://ipfs.io/) and using a pinning service could be yet another remedy to fix this problem, but this would be out of scope of this recipe.

# Table of Data Standards

|  |
| --- |
| **Data Formats** |
| [Turtle](https://www.w3.org/TR/turtle/) |
| [YAML](https://yaml.org/spec/1.2/spec.html) |
| [JSON-LD](https://www.w3.org/TR/json-ld11/) |

# Authors

|  |  |  |
| --- | --- | --- |
| **Name** | **Affiliation** | **ORCID** |
| Kees van Bochove | The Hyve | [0000-0002-8589-0609](https://orcid.org/0000-0002-8589-0609) |
| Emma Vos | The Hyve | [0000-0002-7024-8883](https://orcid.org/0000-0002-7024-8883) |
| Jolanda Strubel | The Hyve | [0000-0001-6675-4639](https://orcid.org/0000-0001-6675-4639) |

# Licence

This page is released under the Creative Commons 4.0 BY license.

# Recommended reading

1. EHDEN Deliverable [D4.5: Interoperability roadmap](https://zenodo.org/record/4474373)
2. How to choose the correct data model for your (observational/clinical/study?) data:
   1. FHIR
   2. OMOP CDM
   3. Open EHR
   4. i2b2 tranSMART
   5. CDISC/SDTM
   6. …
3. Findability of data by search engines: follow-up of [Search Engine Optimization](https://fairplus.github.io/cookbook-dev/recipes/findability/seo.html) recipe, describing web crawling, robots.txt, sitemaps and possible roadblocks.
4. A metadata profile for observational studies (like the [A Transcriptiomics MetaData Profile](https://fairplus.github.io/cookbook-dev/recipes/interoperability/transcriptomics-metadata.html) recipe).