EHDEN

EUROPEAN HEALTH DATA & EVIDENCE NETWORK

FAIRification of OHDSI assets

Kees van Bochove, Maxim Moinat, Emma Vos, Ilaria Maresi, Tess Korthout

Who we are

Kees van Bochove, Founder
kees@thehyve.nl

Maxim Moinat, Data Engineer
maxim@thehyve.nl

the hyve

Ilaria Maresi, FAIR Data Engineer
ilaria@thehyve.nl

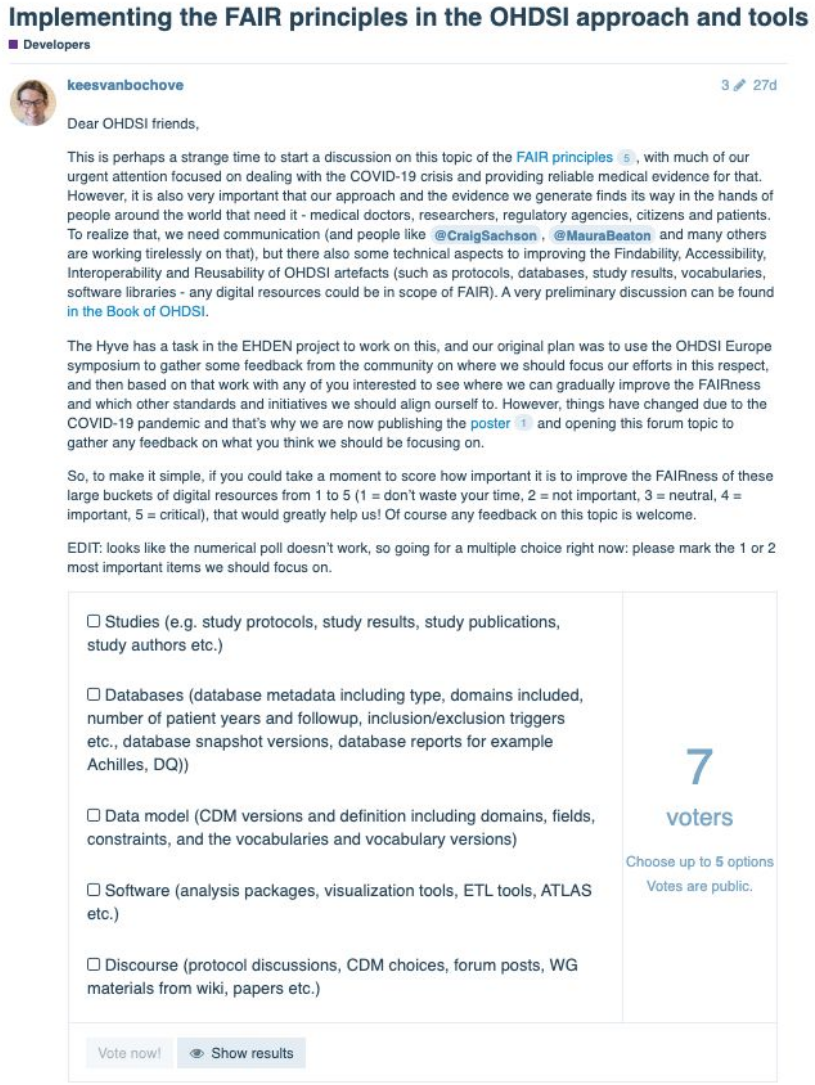
Emma Vos, FAIR Data Engineer
emma@thehyve.nl

Tess Korthout, Data Engineer
tess@thehyve.nl

AGENDA

- The FAIR principles: quick recap
- State of FAIRness of the OHDSI landscape
- How to implement FAIR
- A use case: the COVID-19 study-a-thon

https://forums.ohdsi.org/t/implementing-the-fair-principles-in-the-ohdsi-approach-and-tools/10387



Implementing the FAIR principles in the OHDSI approach and tools

Developers

keesvanbochove 3 27d

Dear OHDSI friends,

This is perhaps a strange time to start a discussion on this topic of the FAIR principles 5, with much of our urgent attention focused on dealing with the COVID-19 crisis and providing reliable medical evidence for that. However, it is also very important that our approach and the evidence we generate finds its way in the hands of people around the world that need it - medical doctors, researchers, regulatory agencies, citizens and patients. To realize that, we need communication (and people like @CraigSachson , @MauraBeaton and many others are working tirelessly on that), but there also some technical aspects to improving the Findability, Accessibility, Interoperability and Reusability of OHDSI artefacts (such as protocols, databases, study results, vocabularies, software libraries - any digital resources could be in scope of FAIR). A very preliminary discussion can be found in the Book of OHDSI.

The Hyve has a task in the EHDEN project to work on this, and our original plan was to use the OHDSI Europe symposium to gather some feedback from the community on where we should focus our efforts in this respect, and then based on that work with any of you interested to see where we can gradually improve the FAIRness and which other standards and initiatives we should align ourself to. However, things have changed due to the COVID-19 pandemic and that's why we are now publishing the poster 1 and opening this forum topic to gather any feedback on what you think we should be focusing on.

So, to make it simple, if you could take a moment to score how important it is to improve the FAIRness of these large buckets of digital resources from 1 to 5 (1 = don't waste your time, 2 = not important, 3 = neutral, 4 = important, 5 = critical), that would greatly help us! Of course any feedback on this topic is welcome.

EDIT: looks like the numerical poll doesn't work, so going for a multiple choice right now: please mark the 1 or 2 most important items we should focus on.

- ☐ Studies (e.g. study protocols, study results, study publications, study authors etc.)
- ☐ Databases (database metadata including type, domains included, number of patient years and followup, inclusion/exclusion triggers etc., database snapshot versions, database reports for example Achilles, DQ))
- ☐ Data model (CDM versions and definition including domains, fields, constraints, and the vocabularies and vocabulary versions)
- ☐ Software (analysis packages, visualization tools, ETL tools, ATLAS etc.)
- ☐ Discourse (protocol discussions, CDM choices, forum posts, WG materials from wiki, papers etc.)

7 voters

Choose up to 5 options

Votes are public.

Vote now! Show results

EHDEN

EUROPEAN HEALTH DATA & EVIDENCE NETWORK

Open Science & the FAIR principles

A short refresher of making data Findable, Accessible, Interoperable and Reusable

imi efpia

The Roots of FAIR

- Public-private partnership to advance:
  - Open Science
  - Sustainability & reuse of data
- Workshop in Leiden in 2014
  - Towards a Modular Blueprint ‘Floor-plan’ of a safe and fair Data Stewardship, Trading and Routing environment, provisionally called the Data FAIRPORT

https://www.lorentzcenter.nl/lc/web/2014/602/info.php3?wsid=602

Open Science

netherlands eScience center

elixir NETHERLANDS

DTL

DUTCH TECHCENTRE FOR LIFE SCIENCES

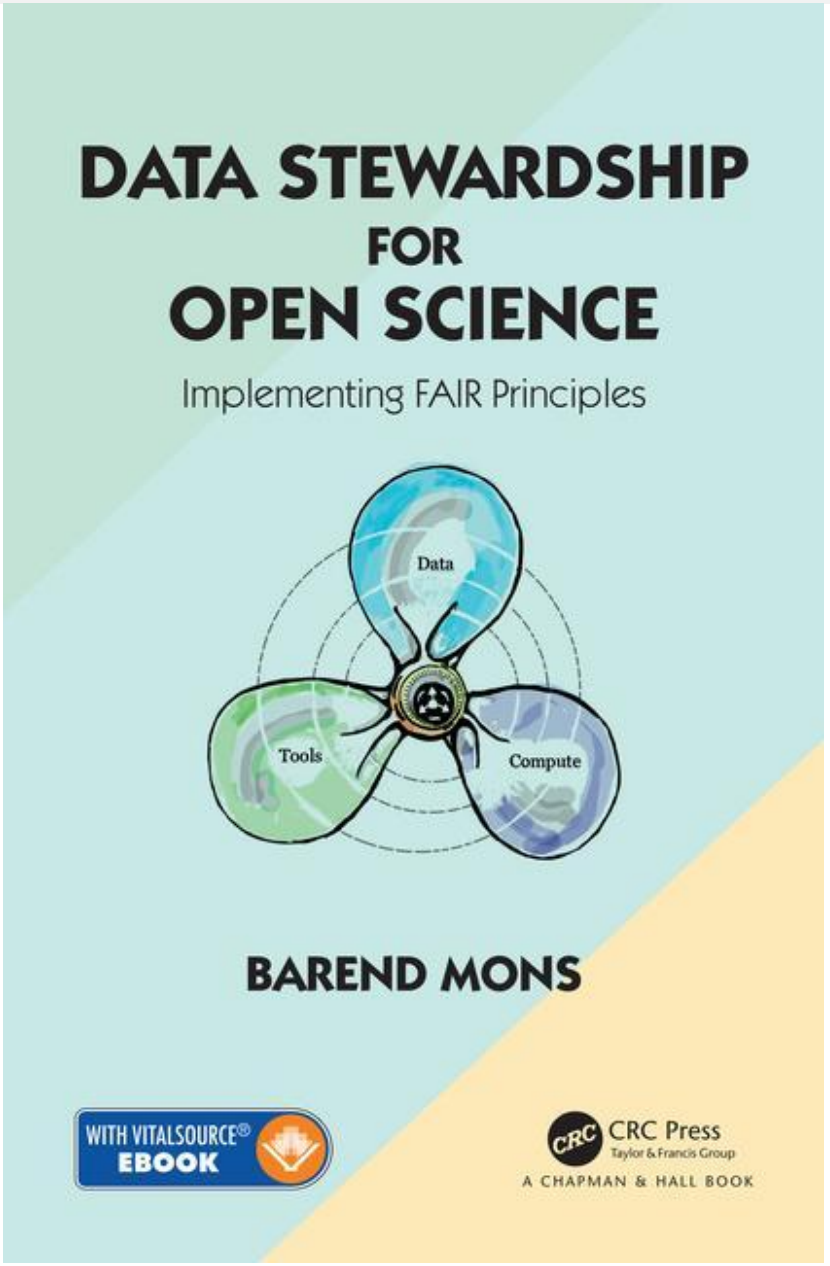
RDA RESEARCH DATA ALLIANCE

GO FAIR

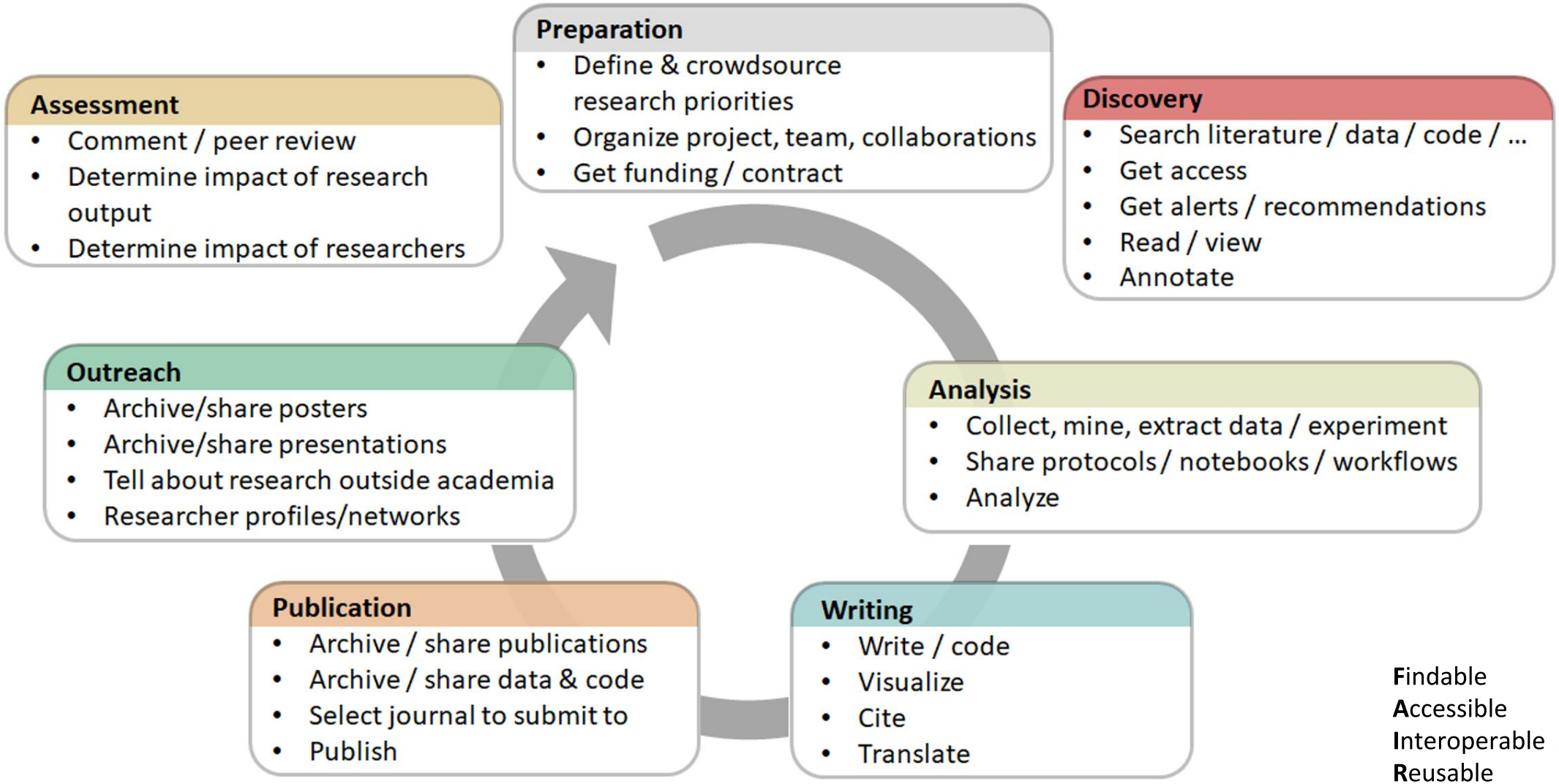
webinar Origins and perspectives for FAIR Data with DTL the hyve

https://youtu.be/5YgAH3f9LiU

imi efpia



Open Science In Practice

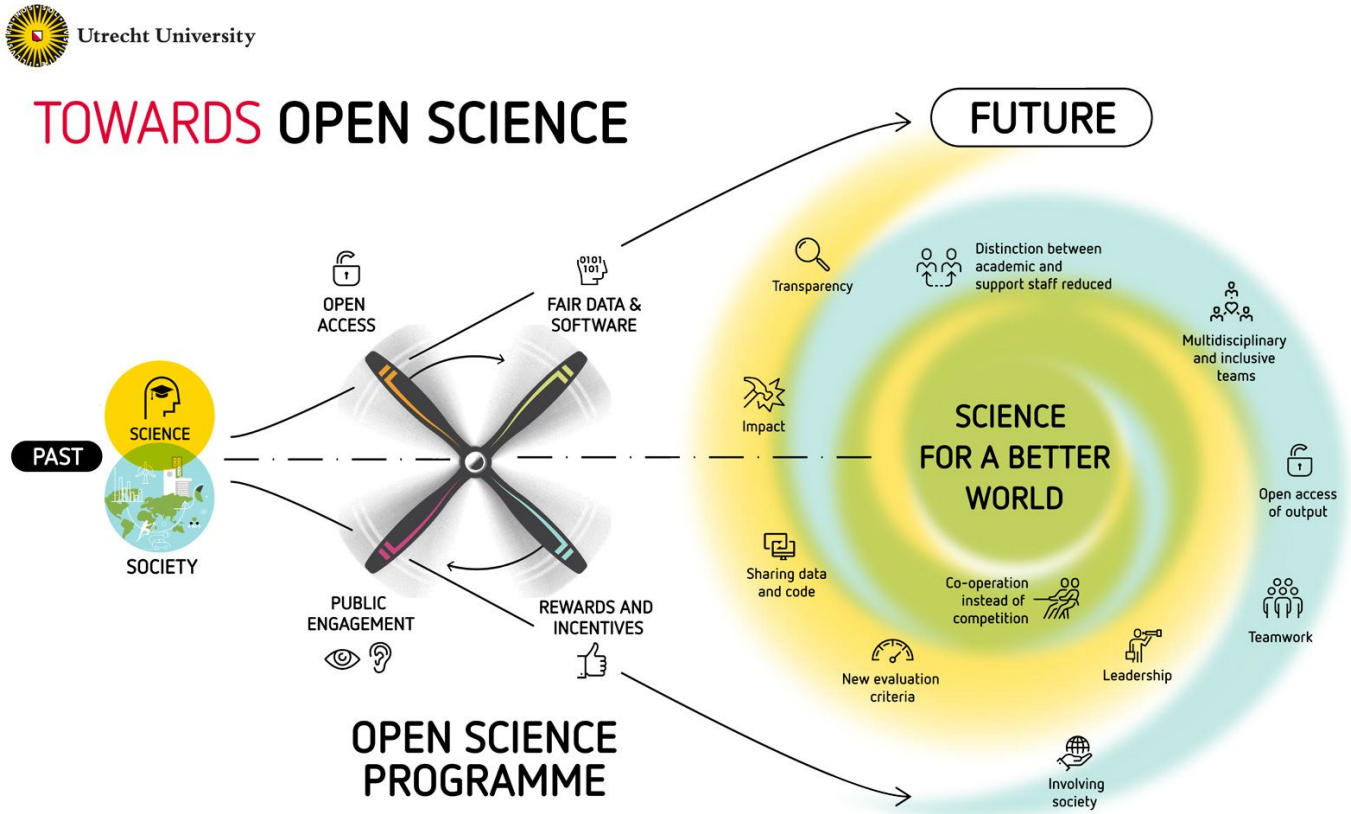


Findable
Accessible
Interoperable
Reusable

https://doi.org/10.5281/zenodo.2587951

Open Science and OHDSI

https://ohdsi.github.io/TheBookOfOhdsi/OpenScience.html



https://www.uu.nl/en/research/open-science

OHDSI
Observational Health Data Sciences and Informatics

The Book of OHDSI

vational research is a field which will benefit greatly from We actively seek and encourage fresh methodological work.

Reproducibility: Accurate, reproducible, and well-calibrated evidence is necessary for health improvement.

Community: Everyone is welcome to actively participate in OHDSI, whether you are a patient, a health professional, a researcher, or someone who simply believes in our cause.

Collaboration: We work collectively to prioritize and address the real world needs of our community's participants.

Openness: We strive to make all our community's proceeds open and publicly accessible, including the methods, tools and the evidence that we generate.

Beneficence: We seek to protect the rights of individuals and organizations within our community at all times.

https://www.ohdsi.org/who-we-are/mission-vision-values/

EHDEN

EUROPEAN HEALTH DATA & EVIDENCE NETWORK

State of FAIRness of the OHDSI landscape

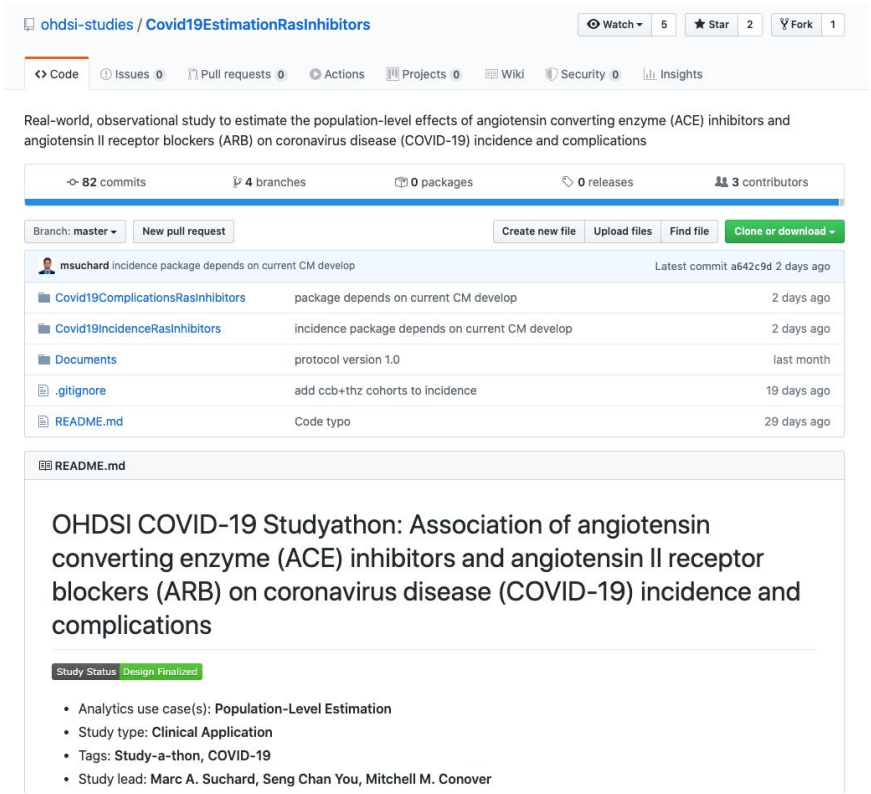
imi efpia

OHDSI Digital Resources

- Data model
  - OMOP CDM versions and definitions
  - Vocabularies
- Data network
  - Data source metadata
- Software
  - Analysis packages, visualization tools, ETL tools, ATLAS
- Studies
  - Study protocols, results, publications, authors
- Other materials
  - Protocol discussions, CDM choices, forum posts, working group documents, papers

Current state: Data Network and Study pages

Database	Data Type	Contact email	Country	# of Patients (000s)	CDM Status
AltaMed Health Services	EHR		USA	638	CDMv4 complete
ARS	Claims		Italy	4,000	CDMv4 in progress
AU-ePBRN (Australian Electronic practice based research network)	EHR (Primary care data linked with hospital admissions)	Jitendra (z3339253@unsw.edu.au); Teng (siaw@unsw.edu.au)	Australia	1,100	CDMv5.3.1 Complete
AUSOM	EHR	Dahye Shin at dasoon0031@naver.com	Korea	2,2,270	CDMv5 complete
BTRIS	EHR/CTDMS	Vojtech.Huser at nih dot gov	USA	500	CDMv4 complete (without era tables) (pilot implementati , ETL posted
Clinical Practice Research Datalink (CPRD)	EHR		UK	11,560	CDMv4 converted to CDMv5 complete, CDMv4 ETL posted
CMS	Claims		USA	2.000	CDMv4 draft



The FAIR Principles

Findable:

F1. metadata are assigned globally unique and **persistent identifier**;

F2. data are described with rich **metadata**;

F3. metadata clearly and explicitly include the identifier of the data it describes;

F4. metadata **registered or indexed** in a searchable resource;

Accessible:

A1. metadata retrievable by their identifier using a **standardized** communications **protocol**;

A1.1 protocol is **open**, free, and universally implementable;

A1.2. protocol allows for an **authentication and authorization** procedure, where necessary;

A2. metadata accessible, even when data are no longer available;

Interoperable:

I1. metadata use a **formal**, accessible, shared, and broadly applicable language for **knowledge representation**.

I2. metadata use **vocabularies** that follow FAIR principles;

I3. metadata include qualified **references** to other (meta)data;

Reusable:

R1. metadata are richly described with a plurality of accurate and relevant **attributes**;

R1.1. metadata are released with a clear and accessible data usage **license**;

R1.2. metadata are associated with detailed **provenance**;

R1.3. metadata meet domain-relevant **community standards**;

http://www.nature.com/articles/sdata201618

OHDSI Assets are Interoperable and Reusable

Interoperable:

I1. (meta)data use a **formal**, accessible, shared, and broadly applicable language for **knowledge representation**.

I2. (meta)data use **vocabularies** that follow FAIR principles;

I3. (meta)data include qualified **references** to other (meta)data;

OMOP CDM

OMOP Standard Vocabularies

Relationships between source and standard vocabularies

Reusable:

R1. meta(data) are richly described with a plurality of accurate and relevant **attributes**;

R1.1. (meta)data are released with a clear and accessible data usage **license**;

R1.2. (meta)data are associated with detailed **provenance**;

R1.3. (meta)data meet domain-relevant **community standards**;

OHDSI software licenced under Apache 2.0, data not consistently licensed

OMOP CDM, Vocabulary and Methods Library for best-practice observational research

OHDSI Assets can be more Reusable

Interoperable:

I1. (meta)data use a **formal**, accessible, shared, and broadly applicable language for **knowledge representation**.

I2. (meta)data use **vocabularies** that follow FAIR principles;

I3. (meta)data include qualified **references** to other (meta)data;

Metadata is not in an interoperable format

Reusable:

R1. meta(data) are richly described with a plurality of accurate and relevant **attributes**;

R1.1. (meta)data are released with a clear and accessible data usage **license**;

R1.2. (meta)data are associated with detailed **provenance**;

R1.3. (meta)data meet domain-relevant **community standards**;

OHDSI Assets are Accessible

Findable:

F1. (meta)data are assigned a globally unique and **persistent identifier**;

F2. data are described with rich **metadata**;

F3. metadata clearly and explicitly include the identifier of the data it describes;

F4. (meta)data are **registered or indexed** in a searchable resource;

Accessible:

A1. (meta)data are retrievable by their identifier using a **standardized** communications **protocol**;

A1.1 the protocol is **open**, free, and universally implementable;

A1.2. the protocol allows for an **authentication and authorization** procedure, where necessary;

A2. metadata are accessible, even when the data are no longer available;

SqlRender to translate OHDSI SQL to other SQL dialects

Achilles,
(EHDEN data catalogue)

OHDSI Assets are not (machine) Findable

Findable:

F1. (meta)data are assigned a globally unique and **persistent identifier**;

F2. data are described with rich **metadata**;

F3. metadata clearly and explicitly include the identifier of the data it describes;

F4. (meta)data are **registered or indexed** in a searchable resource;

Accessible:

A1. (meta)data are retrievable by their identifier using a **standardized** communications **protocol**;

A1.1 the protocol is **open**, free, and universally implementable;

A1.2. the protocol allows for an **authentication and authorization** procedure, where necessary;

A2. metadata are accessible, even when the data are no longer available;

EHDEN

EUROPEAN HEALTH DATA & EVIDENCE NETWORK

How to implement FAIR

From FAIR assessment to linked data

imi efpia

Workflow & Proof of Principle

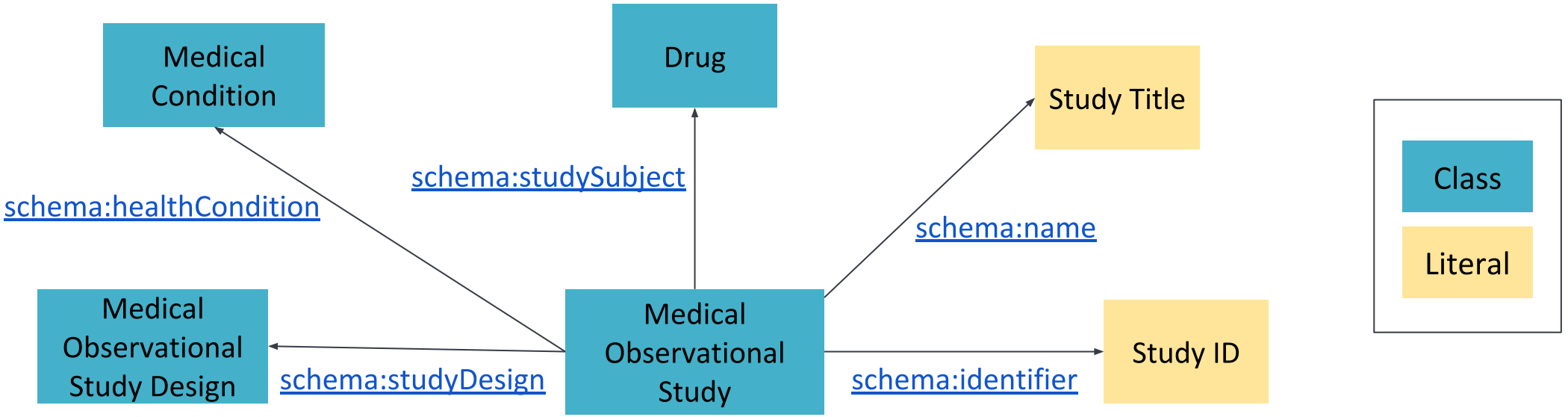
- Identified digital resources in OHDSI
- Selected study-a-thons and study databases as resources to begin FAIRifying
- Assessed FAIRness of study-a-thon and study databases
- Chose one aspect to improve → *Findability of studies*
- Improving Findability via **study-a-thon website** annotated with **rich & findable metadata**

Rich metadata

Metadata should be generous and extensive, to make data **Findable** and **Reusable**.

➔ Map study metadata elements to **Schema.org** concepts

- Schema.org is a vocabulary that can be used to structure metadata on the Internet.



Use Schema.org concepts for metadata elements

Map study metadata elements to **Schema.org** concepts:

Class	Metadata element	Property	Range
schema.org/MedicalObservationalStudy	Study id	schema.org/identifier	Data type: https://schema.org/Text Data type: https://schema.org/URL
	Study title	schema.org/name	Data type: https://schema.org/Text
	Study description	schema.org/description	Data type: https://schema.org/Text
	Study protocol	schema.org/studyDesign	Class: schema.org/MedicalObservationStudyDesign
	Medical Condition studied	schema.org/healthCondition	Class: schema.org/MedicalCondition
	Drug studied	schema.org/studySubject	Class: schema.org/Drug

Representing metadata as linked data

Findable metadata

Capture metadata as **machine-readable**, **structured data**: **JSON-LD**

➔ Encoding Schema.org in JSON-LD allows metadata to be searchable using Google and other search engines

Linked data is more **Findable** and **Interoperable** = more **Reusable**

```
{
"@context" : "https://schema.org",
"@type" : "MedicalObservationalStudy",
"name" : "Covid19EstimationHydroxychloroquine",
"healthCondition" :
        {"@type" : "MedicalCondition",
        "name" : "COVID-19"},
"studySubject":
        {"@type" : "Drug",
        "name" : "hydroxychloroquine"}
}
```

EHDEN

EUROPEAN HEALTH DATA & EVIDENCE NETWORK

A use case: the COVID-19 study-a-thon

Creating FAIR assets

imi efpia

Demo

- https://covid19.ohdsi.app/

  - Human-readable page
    - https://covid19.ohdsi.app/study/ace-arb/
  - Machine-readable metadata
    - view-source:https://covid19.ohdsi.app/study/ace-arb/
  - Generated from one rich YAML document

Resources

- https://covid19.ohdsi.app/
- https://github.com/thehyve/ohdsi-covid19-site

www.ehden.eu

@IMI_EHDEN

IMI_EHDEN

github.com/EHDEN

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.