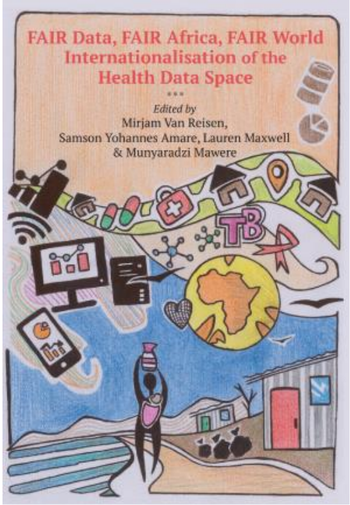
Creation of a FAIR Data Point for a Clinical Trial: the schistosome controlled human infection dataset

Aliya Aktau, Mirjam van Reisen

Chapter in:
Fair Data Fair Africa Fair World:
Internationalisation of the Health Data Space


FAIR Data, FAIR Africa, FAIR World
Internationalisation of the
Health Data Space
Edited by
Mirjam Van Reisen,
Samson Yohannes Amare, Lauren Maxwell
& Munyaradzi Mawere
TB

Cite as: Aktau, A. & Van Reisen, M. (2025). Creation of a FAIR Data Point for a Clinical Trial: the CoHSI2 dataset. https://doi.org/10.5281/zenodo.15382957. In Van Reisen, M., Amare, S. Y., Maxwell, L. & Mawere, M. (Eds.), FAIR data, FAIR Africa, FAIR world: Internationalisation of the Health Data Space. (pp. 397–428). Bamenda: Langaa. URL: https://www.researchgate.net/publication/391750151_FAIR_Data_FAIR_Africa_FAIR_World_The_Internationalisation_of_the_Health_Data_Space

The About the Authors note can be found here: https://raee.eu/wp-content/uploads/2025/05/About-the-Authors-and-Editors.pdf
The list of figures and tables can be found here: https://raee.eu/wp-content/uploads/2025/05/List-of-Figures-and-Tables.pdf

Contents

Foreword……………………………………………………..................xi

Mouhammad Mpezamihigo

Acronyms………………………………………………………..………....xv

Chapter 1: FAIR Data, FAIR Africa, FAIR World: The Internationalisation of the Health Data Space………………………...1

Mirjam van Reisen, Samson Yohannes Amare, Lauren Maxwell & Munyaradzi Mawere

Chapter 2: Bridging Borders with FAIR Data: Transforming Digital Ecosystems for Maternal Health and Public Health Surveillance in Africa…………………………………………………………..…25

Samson Yohannes Amare

Chapter 3: Introducing Data Sovereignty Over Patient Data: Patient Data Ownership in Residence of Health Facilities in Kenya………65

Reginald Nalugala, Putu Hadi Purnama Jati, Samson Yohannes Amare, Maxwell Omare, Jacinta Wairimu, Charles Kahiro, William Nandwa, Seth Okeyo, Dennis Kinoti, Aliya Aktau, Albert Mulingwa & Mirjam van Reisen

Chapter 4: GO TRAIN: A Protocol for Metadata Creation for the FAIRification of Patient Data Health Records……………………...97

Aliya Aktau, Samson Yohannes Amare, Mirjam van Reisen, Getu Tadele Taye, Tesfit Gebreslassie Gebremeskel, Putu Hadi Purnama Jati & Ruduan Plug

Chapter 5: De Novo FAIRification: A Literature Review……..…..143

Zhengyu Lin

Chapter 6: Federating Tools for FAIR Patient Data: Strengthening Maternal Health and Infectious Disease Surveillance from Clinics to Global Systems…………………………………………………..……..157

Samson Yohannes Amare, Getu Tadele Taye, Ruduan Plug, Araya Abrha Medhanyie & Mirjam van Reisen

Chapter 7: Adoption of FAIR-OLR Architectures to Support Insights from Patient Health Data Records in Africa…………………….…195

Putu Hadi Purnama Jati, Samson Yohannes Amare, Abdullahi Abubakar Kawu, William Nandwa, Getu Tadele Taye & Mirjam van Reisen

Chapter 8: A Critical Incident Assessment of a FAIR Implementation Study on Patient Health Data in Africa……………………………..…241

Abdullahi Abubakar Kawu, Putu Hadi Purnama Jati, Lars Schrijver, Mirjam van Reisen, Getu Tadele Taye, Dympna O'Sullivan & Lucy Hederman

Chapter 9: The Potential of Adoption of FAIR Guidelines in Digital Healthcare in Kazakhstan……………………………………………..…277

Aliya Aktau

Chapter 10: Harmonising Antenatal Care Records Across Four African Countries: A Comparative Analysis and Development of a Standardised Data Model……………………………………………...301

Samson Yohannes Amare, Liya Mamo Weldu & Tesfit Gebremeskel

Chapter 11: Complexities of Recording and Reporting Perinatal Data: A Case Study from Namibia……………………………………….…325

Beatrix Callard

Chapter 12: Enhancing Data Privacy and Availability in Indonesia's SATUSEHAT Platform through FAIR-OLR Principles…………….…………………………………………………361

Putu Hadi Purnama Jati, Markus Sintong Tambah Lasroha & Mirjam van Reisen

Chapter 13: Creation of a FAIR Data Point for a Clinical Trial: The CoHSI2 Dataset..……………..…………………………………..…397

Aliya Aktau & Mirjam van Reisen

Chapter 14: Implementation of De Novo FAIRification in Relational Legacy Systems: The Case of the Electronic Medical Record System for Maternal Health in Afya.ke..………………… ………………...429

Zhengyu Lin

Chapter 15: Narratives in Public Agenda-Setting for FAIR Data and Health Data Management in Africa: Enhancing Maternal Health and Infectious Disease Outcomes..……………………………………….467

Putu Hadi Purnama Jati & Mirjam van Reisen

Chapter 16: Testing the Cross-Border Africa Health Data Space: Monitoring, Tracking and Prevention of Mother to Child Syphilis Transmission………………………………………………………...517

Samson Yohannes Amare, Liya Mamo Weldu, Araya Abrha Medhanyie & Mirjam van Reisen

Chapter 17: GO FAIR: Ontology Development of Health Semantics with Cultural Specificity: Traditional Health Practices using Tsebel in Conflict Zones……………………………………………………..553

Bereket Godifay Kahsay, Mirjam van Reisen & Zhengyu Lin

Chapter 18: Satu Data Indonesia and FAIR Data: Advancing Coherent Data Management in Government Administration…….585

Intan K. Utami & Mirjam van Reisen

Chapter 19: Integrating the Personal Health Train Methodology into Healthcare Systems for Enhanced Elderly Care in the Netherlands………………………………………………………….625

Ria Landa-Figueroa & Lars Schrijver

Chapter 20: FAIR Data Implementation for Analysis of Research Data in Human Trafficking and Migration…….…………………665

Kai Smits, Mehul Upase, Nimish Pandey, Senjuti Bala & Mirjam van Reisen

Chapter 21: A Higher Education Curriculum for Cultural Competence, Representation and Social Responsibility in AI and FAIR Data Practices……………………………………………...685

Sakinat Folorunso, Francisca Oladipo, Mirjam van Reisen & Ibrahim Abdullahi

About the Editors……………………………………………...721

Chapter 13

Creation of a FAIR Data Point for a Clinical Trial: the schistosome controlled human infection dataset

Aliya Aktau, Mirjam van Reisen[i]

Abstract

This study explores the FAIRification process of a dataset from a a controlled human infection study with *Schistosomiasis mansoni*, using a FAIR by increment approach. Existing datasets were progressively enhanced to meet FAIR (Findable, Accessible – under well-defined conditions, Interoperable, and Reusable) principles, culminating in the creation of a fully developed FAIR Data Point. The study presented ten concrete steps to deploy a fully-fledged FAIR Data Point that was deployed on the Internet and publicly available on the Internet. The deployment required expertise in semantic web data mapping and IT deployment, requiring a specialised background for the preparation and deployment. This process ensures data verification, enhances the reliability of publications, and maintains the value of the data for future reuse in other studies. The integration of new tools has facilitated the retrospective FAIRification of legacy data. The study recommends incorporating FAIR processes into data collection methodologies, standardising practices, and embedding FAIRification from the beginning of the study. Despite progress, challenges remain in the deployment of FAIR Data Points, particularly in testing interoperability and refined access control mechanisms. Granular capabilities for data security and privacy protection, especially for sensitive data, are still under development. The study concludes that while further advancements are necessary, FAIR Data Points are essential for enhancing academic transparency and accountability, and promoting their adoption will significantly benefit the scientific community.

Keywords: FAIR Data Point, scientific data management, open science, clinical studies, CoHSI2

Introduction

The FAIR Principles, proposing data is curated as Findable, Accessible (under well-defined conditions), Interoperable, and Reusable, emphasise the necessity for computers to autonomously access and process published data without requiring human intervention (Bonino et al., 2016). Wilkinson et al (2016) identified 15 facets under the four principles that operationalise the framework of making data FAIR.

The Findability principle of FAIR data focuses on ensuring that datasets are discoverable through machine-readable and indexable metadata. This metadata enables users to determine whether a data provider holds relevant records. Additionally, beyond basic discoverability, metadata must include trustworthiness indicators, licensing conditions, data representation formats, and semantic details to support access and reuse decisions. A FAIR Data Point (FDP) was envisaged as a software layer over data resources. The FDP adheres to the FAIR principles by organising metadata into four complementary layers: Layer 1: FDP Metadata; Layer 2: Data Catalogue Metadata; Layer 3: Dataset Metadata; Layer 4: Data Record Metadata (Bonino et al., 2016). This structured approach ensures effective data discovery, evaluation, and accessibility in accordance with FAIR principles.

This study investigates the establishment of a full and detailed FAIR Data Point for a clinical study. The dataset comprised of a controlled human *Schistosoma mansoni* infection study led by Koopmans et al (2023) in which a male-only controlled human *Schistosoma mansoni* infection was performed in Schistosoma-naïve volunteers (NL72661.058.20). The study using female schistosomes built on an earlier clinical study carried out with male schistosomes (Koopman et al., 2023).

Schistosomiasis is a global parasitic disease with no available vaccine. Traditional Phase 2 and 3 field trials for vaccine candidates in Schistosoma-endemic regions require large populations and extended durations to assess efficacy, making them resource-intensive. This study seeks to establish a female-only controlled human Schistosoma mansoni infection model to provide early-stage proof-of-concept data for vaccine candidates. Additionally, this model will facilitate research on schistosome immune responses, particularly relevant for vaccines targeting female schistosome-specific antigens (NL72661.058.20 / CoHSI2).

Two approaches to FAIRification can be distinguished. Jacobsen et al. (2019) propose a generic workflow, while Groenen et al. (2021) propose a workflow for De Novo FAIRification. This is defined as the process of creating and managing data in a way that inherently complies with the FAIR (Findable, Accessible, Interoperable, and Reusable) principles from the point of collection or generation. Unlike retrospective FAIRification, which applies FAIR principles to pre-existing datasets, De Novo FAIRification ensures that data is structured, annotated, and stored in a FAIR-compliant manner from the outset.

FAIRification by Increment refers to the gradual transformation of existing datasets to align with the FAIR (Findable, Accessible, Interoperable, and Reusable) principles. Unlike De Novo FAIRification, which ensures data is FAIR from the outset, FAIRification by Increment means retrofitting FAIR principles onto pre-existing data through an iterative process. This approach is particularly useful for legacy datasets that were not originally structured for machine readability, interoperability, or reusability.

FAIRification is a progressive process. To determine the right level of 'data being FAIR' the objective of the FAIRification needs to be assessed, and evaluated against the understanding of which of the 15 facets of FAIR are the most essential and which of those are more optional.

Even though the intention of the investigators of the clinical trial was from the outset, to ensure source data of the study were available in a format that was fully in accordance with the FAIR-principles, a complete workflow was not available at the start of the study. Therefore, the data was FAIRified retrospectively, following a FAIR by Increment approach. This study investigates the FAIRification of (meta)data with an incremental approach.

Relevance of the study

Wilkinson (2016) set out the value of scientific data as the main source of academic production, which has been undervalued. Open science focuses on the transformation of academia to account for its responsibility to ensure that publications and data are open for other scholars, and the public, to build and benefit from the work undertaken. This is relevant from the perspective of efficiency, and accountability to public resources and adheres to the core of the academic philosophy, to build on work that was previously undertaken including by other scholars.

While the FAIR-idea had rapid uptake and was associated with a problem that was intuitively understood, at the time when the EU stipulated the use of FAIR data to subsidised research projects in 2020, it was unclear how the FAIRification process could be adopted (Stocker et al., 2022). Even though Jacobsen et al (2019) had proposed a workflow, the implementation of this workflow depended on FAIR Supporting Resources, which were generally immature or unavailable.

A particular sticking point is the Findability and Accessibility requirement, which in an academic context, depends on federated architectures. Such architectures were not readily available and were often not well balanced with other engineering requirements, such as data safety. If data and metadata were assigned, the question of how to practically deploy the data in a secure way, and allow data-visiting, in a cost-effective way, was unclear.

Moreover, an individual academic has little to gain by FAIRification of data, unless this is followed by others. Hence the uptake of one depends on the adoption by others. The data published in the FAIR Data Point described in this article is the result of a study by Koopman et al. (2023), early adopters of the FAIRification obligation in science.

A final sticking point was the difficulty for investigators to understand what would be gained by the FAIRification process. This study has the aim of taking a FAIRification process from beginning to end, to demonstrate all the steps involved, and to show the outcome for this dataset in enhancing its potential to serve current and future academic inquiries. At the very least, the publication of the data on the FDP will enhance the accountability of publications and findings of the study and will allow other researchers to investigate and understand the source data.

FAIR design considerations

There are three steps involved to complete the FAIRification process:

- Determine the FAIRification approach
- Implement the FAIR workflow that relates to the selected approach
- Install all attributes on an FDP

These three steps are set out in detail below.

FAIRification approach: De Novo FAIRification or FAIR by increment

The FAIRification process involves transforming data to align with the FAIR principles, Findable, Accessible, Interoperable, and Reusable. Two primary approaches to this process are De Novo FAIRification and FAIRification by Increment, each differing in their methodology and application.

De Novo FAIRification is an approach that involves designing and implementing data systems that are inherently FAIR from their inception. In the study by Groenen et al. (2021), a registry for vascular anomalies was developed with FAIR principles integrated from the beginning. The process encompassed five phases:

- Pre-FAIRification: defining objectives, assembling a multidisciplinary team, and planning resources.
- Facilitating FAIRification: developing semantic models and establishing technical infrastructure to support FAIR data.
- Data collection: implementing data collection protocols that ensure data conforms to FAIR standards upon entry.
- Generating FAIR data in Real-Time: automatically converting collected data into machine-readable formats compliant with FAIR principles.
- Using FAIR data: Utilising the FAIR data for research and clinical purposes, ensuring interoperability and reusability across platforms.

This strategy – in which the FAIRification is integrated into the data production, ensures that data is FAIR from the point of creation, facilitating seamless integration and utilisation (Kersloot, 2021; Lin, 2025a; 2025b). In studies where De Novo FAIRification is not a possibility because data is already existing, a workflow for FAIRification by Increment is proposed. This involves the retrospective transformation of existing datasets to comply with FAIR principles (Amare, 2025).

Here the workflow proposed by Jacobsen et al. (2019) can be followed:

- Pre-FAIRification: assessing current data and metadata, setting FAIRification goals, and identifying necessary resources.
- FAIRification: defining semantic models, enhancing data and metadata to be linkable and interoperable, and applying necessary transformations.
- Post-FAIRification: hosting the FAIRified data inaccessible repositories and evaluating the FAIRness to ensure compliance.

This retrospective approach allows for the enhancement of legacy datasets to meet FAIR standards, improving their accessibility and usability over time.

FAIRification by Increment is relevant for legacy data, and in instances when source data is obtained prior to the FAIRification process. The consideration for which workflow is best used, depends largely on the timing when the FAIRification can take place. If source data already exists, then source data is progressively enhanced. The existing datasets are improved to increase FAIR compliance, and to adhere progressively to the 15 FAIR facts defined by Wilkinson et al (2016).

During the FAIRification, the FAIR principles are implemented in stages, starting with metadata enrichment, followed by improvements in data accessibility, interoperability, and reusability. In FAIRification by increment, as is the case by De Novo FAIRification, controlled vocabularies, ontologies, and persistent identifiers are used to enhance findability and machine-actionability. This task is referred to as metadata standardisation. To increase syntactic interoperability data formats are gradually mapped and transformed into standardised, machine-readable formats such as Resource Description Framework (RDF), JavaScript Object Notation for Linked Data (JSON-LD), or eXtensible Markup Language (XML). This step is also critical to a De Novo FAIRification setup. In both approaches FAIR Supporting Resources (FSRs) are used to incorporate existing semantic frameworks, knowledge graphs, and linked data repositories to enhance data integration. In FAIRification by increment, some different tools are used to obtain the result of deploying source data as FAIR. The FAIRified end product, usually an FDP, incorporates the existing semantic frameworks, knowledge graphs, and linked data repositories to enhance data integration and provides information on how the data can be accessed and reused and under which conditions. Table 1 shows a comparison between FAIRification by Increment and the De Novo FAIRification.

Table 1. Comparison of FAIRification by Increment and De Novo FAIRification

Aspect	**FAIRification by Increment**	**De Novo FAIRification**
Timing	Applied after data creation	Applied during data creation
Approach	Gradual enhancement of existing data	Designed to be FAIR from the start
Complexity	May require data restructuring and cleanup	Requires initial planning and infrastructure
Metadata Integration	Added retroactively	Integrated during data generation

The differences between the two approaches can be understood by the following three elements:

- Timing: De Novo FAIRification integrates FAIR principles during the initial design and data collection phases, whereas FAIRification by Increment applies these principles to existing datasets after their creation.
- Implementation complexity: De Novo FAIRification requires comprehensive planning and infrastructure development upfront, while FAIRification by Increment may involve

complex transformations and mappings to retrofit FAIR principles onto legacy data.

- Data quality: De Novo FAIRification ensures high-quality, FAIR-compliant data from the outset, whereas FAIRification by Increment aims to enhance the FAIRness of existing data, which may have varying levels of initial quality.

De Novo FAIRification is advantageous, due to the data quality. However, if the source data is already collected, FAIR by increment offers an alternative approach. Any FAIRification approach is designed to be adaptable, allowing iterations and refinements to achieve optimal FAIR compliance (Jacobsen et al., 2019).

Generic FAIRification workflow

To create a FAIR Data Point, data should be curated according to FAIR principles. The FAIRification workflow proposed by Jacobsen et al. (2019) outlines a structured approach to transform datasets to align with the FAIR principles. This process involves three main phases, each comprising specific steps (Jacobsen et al., 2019):

1. Pre-FAIRification Phase:

1.1. Identification of FAIRification Objectives. In this phase the goals for making the data FAIR is determined, considering the specific needs and potential benefits for the target community.

1.2. Analysis of the Source Data. In this phase the current state of the data is inspected, including its structure, quality, and existing metadata, to understand the extent of work required for FAIRification.

1.3. Analysis of Metadata. In this phase, the existing metadata is examined to determine its adequacy in supporting data discovery and reuse, identifying gaps that need addressing.

2. FAIRification Phase:

2.1. Definition of the Semantic Model. In this phase, a semantic model for data and metadata is identified.

- Selection of semantic model for the metadata. This involves the creation of a semantic model for metadata. The realisation of this step will ensure the description of the dataset and the data, to facilitate findability and accessibility.

- Selection of semantic model for source data and creation of a data model. This involves the development of a semantic framework that accurately represents the meaning and context of the source data. The completion of this step will facilitate data interoperability and reuse.

2.2. Creating the format for Data and Metadata to be Linkable:

- Linking metadata connects them to appropriate standards and vocabularies, which facilitates interoperability and reuse.
- Linking source data with relevant external resources, facilitates understanding of the provenance of the data, the context in which the data was generated, and usability.

3. Post-FAIRification Phase:

3.1. Deployment and hosting of FAIR data and metadata. This phase involves the repositing of the FAIRified data and metadata in data stores or platforms that support sustained accessibility and compliance with FAIR principles.

3.2. Assessment of the FAIR data and metadata. This task focuses on the evaluation of the FAIRness of the data and metadata. Tools using established metrics can be used to ensure the data FAIRification meets the desired standards and objectives (Jacobsen et al., 2019).

The realisation of the workflow requires a multidisciplinary team, guided by FAIR data stewards, to effectively execute the FAIRification process.

Installation of an FDP

After the completion of the FAIRification, an FDP is installed. This is a software model that is structured into four layers, each serving a distinct purpose in ensuring data discoverability, interoperability, and machine-actionability. The FDP layers are the following:

(i) The FDP Metadata Layer provides technical details and provenance information, including a formal description of required, recommended, and optional metadata elements. It follows the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) and Dublin Core Terms (DCT).

(ii) The Data Catalogue Metadata Layer describes the datasets available in the FDP using the W3C Data Catalogue Vocabulary

(DCAT), which defines a catalogue as a curated collection of dataset metadata.

(iii) The Dataset Metadata Layer provides details on individual datasets, including available formats and access points (e.g., APIs, XML, RDF, tab-delimited files). It also follows DCAT standards, distinguishing between datasets and their various distributions (formats or endpoints).

(iv) The Data Record Metadata Layer describes the specific data items within a dataset, including data types and semantic annotations. It ensures rich metadata that supports automatic linking and integration, using shared vocabularies and ontologies (e.g., Systematized Nomenclature of Medicine Clinical Terms [SNOMED CT]) for genetic and disease data classification).

By structuring metadata in these layers, FDPs enhance data interoperability, facilitate machine-readability, and support automated data integration within the FAIR ecosystem. Through metadata, the FDP provides the properties offering insight into the datasets it exposes, as well as the license and access conditions, and it can also provide access to the data itself if permission is provided (Bonino et al., 2016) or alternatively allow pre-defined SPARQL queries to run an operation over a pre-agreed range of metadata in one triple store or in #=n federated stores (Taye et al., 2024; Amare et al., 2025).

Based on DCAT, Bonino et al (2016) define a dataset as a collection of data that is published or curated by a single agent and made available for access or download in one or more formats. It serves as the core unit of data organisation within a data catalogue. A distribution refers to a specific available form of a dataset. A single dataset may have multiple distributions, which can differ in terms of file formats (e.g., XML, RDF, CSV) or access endpoints (e.g., Application Programming Interface (API), Query Language and Resource Description Framework (SPARQL) endpoint). Distributions provide multiple ways of accessing the same dataset based on user needs and technical requirements.

While a dataset could refer to a national health survey dataset containing patient demographics and health conditions, the distributions would refer to a CSV file for tabular analysis, RDF format for linked data applications with API access for programmatic queries. This distinction enables interoperability by allowing datasets to be accessed in different technical environments while maintaining semantic consistency.

The fourth layer is important in that it gives information about the data items contained in the dataset. This information allows the users to assess what is the actual content of the dataset, by describing the data types represented in the data (Bonino et al., 2016).

Methodology

Location of the study

This study is carried out in The Netherlands at Leiden University Medical Centre.

Timeline of the study

The study started in 2021, with an inventory of datapoints for the clinical study. The data was analysed in 2022, and in 2023, a common data model was completed. In 2024, an FDP was set up and completed. Following the completion of the FDP, the research team was requested to critically look at all the items and provide feedback. The comments received were used to make improvements and corrections to the FDP, which was finally made public in January 2025.

Data collection

There are four types of source data for the COHSI2 study:

- The data collected for the study is reported in electronic Case Report Forms (eCRF). The eCRF forms are considered source data.
- The notes of the researchers are also considered source data.
- The medical file of a participant who shows a reaction (adverse event) to the intervention. This will be source data. This is particularly the case if the participant will require medical consultation or hospitalisation.

- The diaries were produced by the study volunteers. These will be held in the investigator's site file (NL72661.058.20).

Solicited adverse effects were: itching, fever (by examination), rash, urticaria, headache, fatigue, malaise, coughing, myalgia, arthralgia, night sweats, back pain, anorexia, nausea, vomiting, abdominal pain, and diarrhoea (NL72661.058.20).

Privacy preservation approach

The data of the participants in the clinical data is sensitive and contains personal information. The source data is anonymised. Given the low number of participants, there is a risk of reidentification, which needs to be taken into account when permission on data reuse is requested, meaning that data use agreements need to clarify measures being taken to ensure that the privacy of the participants is protected, including limiting general access to the dataset but performing selected queries for defined purposes.

FAIRification approach

Because data was already collected, the workflow follow a FAIR by design approach, following Jacobsen (2019).

Findings

The results of the study are documented in three sections. The first section describes the steps of the data and metadata FAIRification. This is described as a process carried out in six distinct steps. The second section describes the features of the data dashboard. The third section focuses on the deployment of the data and metadata in the triple store and the features of interoperability and reuse this offers. The next section documents the realisation of the FDP. The final section describes the deployment and hosting of the FDP.

Steps in the creation of the CoHSI2 data and metadata FAIRification

The creation of the data FAIRification involves the following practical steps.

Step 1: Prepare Excel data

Before converting the dataset, ensure the following:

- The dataset has a clear header row with meaningful column names.
- Avoid merged cells, as they can cause issues during conversion.
- Ensure consistency in formats (e.g., dates should follow the YYYY-MM-DD format).

Step 2: Understand data attributes to find ontologies

To find correct ontologies that are corresponding to the data attributes, the following two steps were undertaken

- Review of the column names in the Excel file
- Identification of the meaning of each attribute
- Identify the attributes relevant to matching ontologies.

This is followed by the identification of the potential sources for relevant ontologies.

- Determine the domain of the dataset (e.g., healthcare, finance, environment), for this project we use Healthcare/Biomedical ontologies
- Ontology repositories to find relevant ontologies:
- BioPortal (https://bioportal.bioontology.org/)
- Ontobee (http://www.ontobee.org/)
- OLS (Ontology Lookup Service) by EBI (https://www.ebi.ac.uk/ols/).

Step 3. Match attributes to ontology terms

At this stage, each attribute must be searched within the selected ontology repository. The definitions should be compared to verify their relevance. The ontology term and its corresponding URI must be documented for each matching attribute.

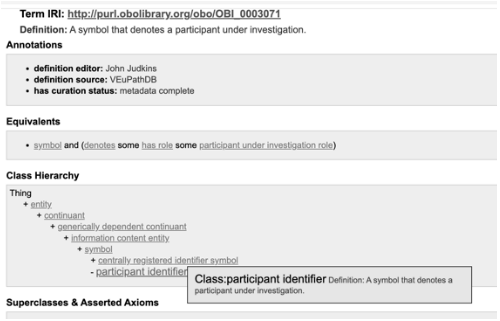
Step 4. Data mappings

In the following step, the data mappings are prepared on the different data elements of the study for each data type. The data mapping links classes to ontologies and defines the relationship, so that knowledge can be obtained from the data linked through these metadata that are linked to each source data instance. In semantic data, the relationship between an ontology and a class is central to how knowledge is represented and structured. An ontology is a formal representation of knowledge within a domain, defining the types of entities that exist in that domain, their relationships, and the rules governing their interactions. It serves as a blueprint for understanding concepts and the relationships between them. In the context of semantic data, an ontology helps represent data in a machine-readable and interoperable way. The knowledge is built up as the metadata form triples (subject – predicate – object) through which expanding layers of relationships become discoverable.

A class in an ontology represents a category or a set of entities that share common characteristics or properties. Classes are essentially the types or concepts in the ontology. For example, in a medical ontology, 'Patient' could be a class, representing all patients with common attributes such as age, gender, and medical history.

The following tables (Table 1, Table 2, Table 3, Table 4, and Table 5) present ontology terms categorised by data type, which have been selected for this study. The ontologies are clustered per class. They are based on the Open Biological and Biomedical Ontologies (OBO) Library, which is a collection of standardised ontologies primarily used in the biological and biomedical sciences. These ontologies are used to represent knowledge about various biological entities, in a machine-readable format. The OBO Library includes many widely used ontologies as Gene Ontology (GO), Disease Ontology (DO), and Sequence Ontology (SO), which help to provide consistent annotations and facilitate data integration across different biological domains. OntoPortal, which includes Ontobee, is a platform that enables the discovery, exploration, and use of ontologies in a centralised and accessible way. OntoPortal provides an interface to view and interact with the ontologies that are part of the OBO Library, thereby enabling better accessibility, integration, and utilisation of the ontologies for research and data analysis.

An example is the class 'participant' which is here defined as a participant in the study. The participant relates to an identity symbol. The ontology has Internationalised Resource Identifier (IRI) which is a unique identifier on the worldwide web.



Term IRI: http://purl.obolibrary.org/obo/OBI_0003071

Definition: A symbol that denotes a participant under investigation.

Annotations

- **definition editor:** John Judkins
- **definition source:** VEuPathDB
- **has curation status:** metadata complete

Equivalents

- symbol **and** (denotes **some** has role **some** participant under investigation role)

Class Hierarchy

```
Thing
  + entity
    + continuant
      + generically dependent continuant
        + information content entity
          + symbol
            + centrally registered identifier symbol
            - participant identifier
```

Class:participant identifier Definition: A symbol that denotes a participant under investigation.

Superclasses & Asserted Axioms

Figure 1. Class Participant and IRI

The following data map was created for the CoHSI2 study: (i) metadata of the study (ii) screening terms (iii) adverse events according to ICD10 (iv) test results per weekly visits (v) treatments. This is also referred to as the RDF skeleton. ICD10 is the International Statistical Classification of Diseases and Related Health Problems 10th Revision.

Table 2. Ontology set 1: metadata of the study

Study information	**Description**	**Ontology term URI**
(variable) Name		
Participant ID	A symbol that denotes a participant under investigation	http://purl.obolibrary.org/obo/OBI_0003071
Cohort	Dosing group and dose amount in CoHSI2 Study	http://purl.obolibrary.org/obo/NCIT_C61512
CoHSI_date	Date of CoHSI	http://purl.obolibrary.org/obo/NCI

Study information	Description	Ontology term URI
(variable) Name		
		T_C69208
CoHSI_cerc	Number of female Schistosoma mansoni cercariae for challenge	http://purl.obolibrary.org/obo/NCIT_C124387

Table 3. Ontology set 2 – screening

Screening	Description	Ontology term URI
(variable) Name		
Participant ID	A symbol that denotes a participant under investigation	http://purl.obolibrary.org/obo/OBI_0003071
Cohort	A research study that compares a particular outcome in groups of individuals who are alike in many ways but differ by a certain	http://purl.obolibrary.org/obo/NCIT_C15208

Screening	**Description**	**Ontology term URI**
(variable) Name		
	characteristic	
Screening Age	Age in years (at week 00)	http://purl.obolibrary.org/obo/NCIT_C25150
Screening Gender	Characteristics of people that are socially constructed, including norms, behaviours, and roles based on sex. As a social construct, gender varies from society to society and can change over time. (Adapted from WHO.)	http://purl.obolibrary.org/obo/NCIT_C17357
Screening Body Mass Index (BMI)	Subject's body mass index (in kg/m2)	http://purl.obolibrary.org/obo/NCIT_C168828

Table 4. Ontology set 3 – adverse events according to ICD10

Adverse Event	**Description**	**Ontology term URI**
(variable) Name		
Participant ID	A symbol that denotes a participant under investigation	http://purl.obolibrary.org/obo/OBI_0003071
Cohort	A research study that compares a particular outcome in groups of individuals who are alike in many ways but differ by a certain characteristic	http://purl.obolibrary.org/obo/NCIT_C15208
Participant ID	A symbol that denotes a participant under investigation	http://purl.obolibrary.org/obo/OBI_0003071
Cohort	A research study that compares a particular outcome in groups of individuals who are alike in many	http://purl.obolibrary.org/obo/NCIT_C15208

Adverse Event	Description	Ontology term URI
(variable) Name		
	ways but differ by a certain characteristic	
Adverse Event description	The verbatim description of the adverse event	http://purl.obolibrary.org/obo/OAE_0000001
ICD-10 code (version:2010)	The tenth version of the International Classification of Diseases (ICD), published by the World Health Organization in 1992	http://purl.obolibrary.org/obo/NCIT_C192551
Adverse Event onset date	The calendar date on which an adverse event starts	http://purl.obolibrary.org/obo/NCIT_C78536
Adverse Event onset time	The time at which an adverse event starts	http://purl.obolibrary.org/obo/NCIT_C78539
Adverse Event end date	The calendar date on which an adverse event	http://purl.obolibrary.org/obo/NCIT_C78537

Adverse Event	Description	Ontology term URI
(variable) Name		
	ends	
Adverse Event end time	The stop time of the adverse event	http://purl.obolibrary.org/obo/NCIT_C83046
Adverse Event severity	A numeric value corresponding to the degree of severity of an adverse event	http://purl.obolibrary.org/obo/NCIT_C166200
Adverse Event treatment	Treatment administered to patients experiencing adverse events	http://purl.obolibrary.org/obo/NCIT_C45501
Adverse Event remarks	A written explanation, observation or criticism added to an Adverse Event	http://purl.obolibrary.org/obo/NCIT_C25393

Table 5. Ontology set 4 - test results per weekly visit

Week X	Description	Ontology term URI
(variable)		

Name		
Participant ID	A symbol that denotes a participant under investigation	http://purl.obolibrary.org/obo/OBI_0003071
Cohort	A research study that compares a particular outcome in groups of individuals who are alike in many ways but differ by a certain characteristic	http://purl.obolibrary.org/obo/NCIT_C15208
Visit	Week number of the scheduled visit	http://purl.obolibrary.org/obo/NCIT_C83101
Date	The date on which a visit occurred	http://purl.obolibrary.org/obo/NCIT_C83031
Visit other	The week number of the unscheduled visit	http://purl.obolibrary.org/obo/NCIT_C170510
Visit reason	Reason for the unscheduled visit	
Erythrocyte Sedimentation Rate (ESR)	A quantitative measurement of the distance that	http://purl.obolibrary.org/obo/NCIT_C74611

(mm)	red blood cells travel in one hour in a sample of unclotted blood.	
Haemoglobin (mmol/L)	A quantitative measurement of the amount of haemoglobin present in a blood sample	http://purl.obolibrary.org/obo/NCIT_C64848
Haematocrit (L/L)	A measure of the volume of red blood cells expressed as a percentage of the total blood volume. (Normal in males is 43-49%, in females 37-43%???)	http://purl.obolibrary.org/obo/NCIT_C64796
Erythrocyte Mean Corpuscular Volume (MCV) (fL)	The mean cell volume is the average volume of a red blood cell. This is a calculated value derived from the haematocrit and the red cell count.	http://purl.obolibrary.org/obo/NCIT_C64799
Thrombocytes (x10^9/L)	The determination of the number of	http://purl.obolibrary.org/obo/NCIT_C51951

	platelets in a biospecimen.	
Leukocytes (x10^9/L)	A test to determine the number of leukocytes in a biospecimen (The amount of a leukocyte)	http://purl.obolibrary.org/obo/OBA_VT0000217
Neutrophils (x10^9/L)	A test to determine the number of neutrophils in a blood sample	http://purl.obolibrary.org/obo/NCIT_C51950
Eosinophils (x10^9/L)	The determination of the number of eosinophils in a blood sample	http://purl.obolibrary.org/obo/NCIT_C64550
Basophils (x10^9/L)	The determination of the absolute number of basophils in a blood sample	http://purl.obolibrary.org/obo/NCIT_C64470
Lymphocytes (x10^9/L)	The determination of the number of lymphocytes in a blood sample	http://purl.obolibrary.org/obo/NCIT_C51949

Monocytes (x10^9/L)	The determination of the classical monocytes in a biospecimen.	http://purl.obolibrary.org/obo/NCIT_C181281
Sodium (mmol/L)	A quantitative measurement of the amount of sodium present in a sample of serum	http://purl.obolibrary.org/obo/NCIT_C61029
Potassium (mmol/L)	A quantitative measurement of the amount of potassium present in a sample of serum	http://purl.obolibrary.org/obo/NCIT_C61030
Creatinine (umol/L)	A quantitative measurement of the amount of creatinine present in a sample of serum	http://purl.obolibrary.org/obo/NCIT_C61023
Blood urea nitrogen (BUN) (mmol/L)	A quantitative measurement of the amount of urea nitrogen present in a serum sample	http://purl.obolibrary.org/obo/NCIT_C61019
Aspartate aminotransferas	A quantitative measurement of	http://purl.obolibrary.org/obo/NCI

e (ASAT) (U/L)	aspartate aminotransferase present in a sample of serum	T_C61018
Alanine transaminase (ALAT) (U/L)	A quantitative measurement of the amount of alanine aminotransferase present in a sample of serum	http://purl.obolibrary.org/obo/NCIT_C61017
Alkaline phosphatase (U/L)	A quantitative measurement of alkaline phosphatase present in a sample of serum	http://purl.obolibrary.org/obo/NCIT_C61016
Gamma-glutamyl transferase (gGT) (U/L)	A quantitative measurement of the amount of gamma glutamyl transpeptidase present in a sample of serum	http://purl.obolibrary.org/obo/NCIT_C61025
Bilirubin total (umol/L)	The measurement of the total amount of bilirubin present in a blood sample *(JJ added in a blood sample!)*	http://purl.obolibrary.org/obo/NCIT_C38037

Serum glucose (mmol/L)	A quantitative measurement of the amount of glucose present in a sample of serum.	http://purl.obolibrary.org/obo/NCIT_C61027

Table 6: Ontology set 5 - treatments

Treatment	Description	Ontology term URI
(variable) Name		
Participant ID	A symbol that denotes a participant under investigation	OBI:0003071
Cohort	Dosing group and dose amount in CoHSI2 Study	http://purl.obolibrary.org/obo/NCIT_C61512
T_Visit (Treatment visit)	A visit by a patient or study participant to a medical professional.	http://purl.obolibrary.org/obo/NCIT_C39564
T_Visit_Other (Treatment visit other)	Treatment visit purpose if doesn't accord to the schedule	http://purl.obolibrary.org/obo/NCIT_C142240

Treatment	**Description**	**Ontology term URI**
(variable) Name		
T_Date(Treatment date)	The time of a treatment.	http://purl.obolibrary.org/obo/AGRO_00010133
T_Medication (Treatment with medication)	Treatment of disease through the use of drugs.	http://purl.obolibrary.org/obo/NCIT_C15986
T_AL (Remarks on AL; Treatment with arthemether/lumefantrine, remarks)	A widely used artemisinin-based combination.	http://purl.obolibrary.org/obo/IDOMAL_0000148
T_PZQ (Dosing schedule praziquantel)	A plan specification that specifies 1) the quantity of which some material entity will be allocated to the eventual realization of some action specification, and 2) the temporal regions in which each action specification is to be realized.	http://purl.obolibrary.org/obo/APOLLO_SV_00000500

Treatment	Description	Ontology term URI
(variable) Name		
T_Remarks (Treatment remarks)	A written explanation, observation or criticism added to textual material.	http://purl.obolibrary.org/obo/NCIT_C25393

The data maps were used to create metadata that have semantic meaning, and are machine-readable identified through the IRI or a Uniform Resource Identifier (URI) of the ontology, each of which are relevant for making data machine-readable. They play a key role in ensuring that data can be unambiguously identified, accessed, and integrated across different systems and platforms in the semantic web and linked data web.

Step 5. Validate and refine the data map

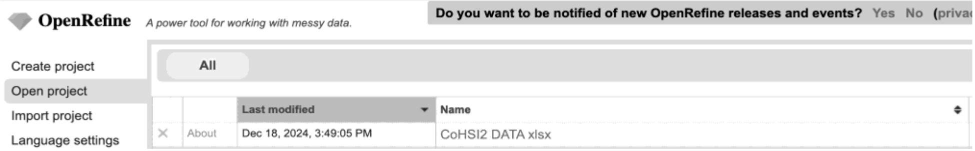
At this point in the workflow, it is important to consult and check for missing or ambiguous mappings and consult domain experts if necessary. Mappings should be refined based on data usage needs to ensure accuracy and relevance.

Step 6: Convert Excel data to FAIR formats

To make data interoperable and machine-readable, it is necessary to express the data in JSON-LD or RDF. In order to do this transformation, the data from the excel sheet can be converted. A no-code tool like OpenRefine can be used, which supports exports to RDF, JSON-LD, and other formats. OpenRefine is a tool for clearing up messy data. OpenRefine can handle all sort of data; import formats include CSV, Excel, Tab-Separated Values (TSV), JSON, Google Spreadsheets, RDF, text file with custom separators. To build an RDF infrastructure, it is recommended that OpenRefine is installed with RDF extension.

These are the steps taken with a conversion into RDF with OpenRefine:

- Load your Excel file into OpenRefine.
- Clean and standardise column names and values if there any



OpenRefine *A power tool for working with messy data.*

Do you want to be notified of new OpenRefine releases and events? Yes No (priva

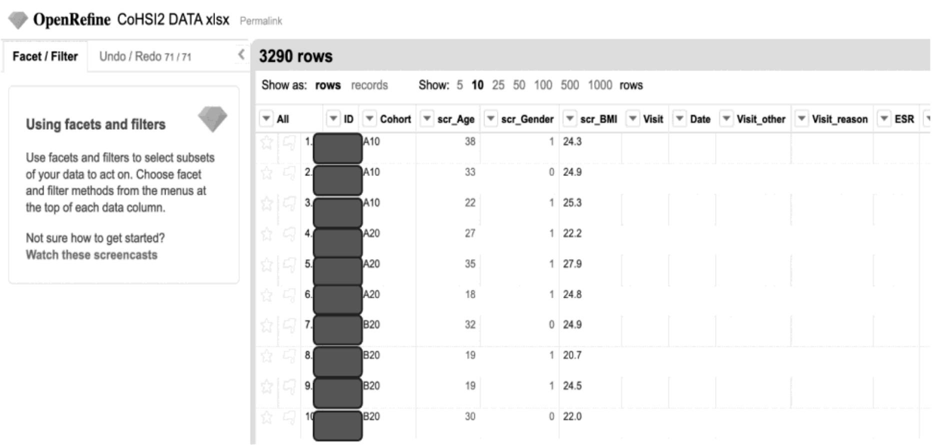
Create project
Open project
Import project
Language settings

All

		Last modified	Name
×	About	Dec 18, 2024, 3:49:05 PM	CoHSI2 DATA xlsx

Figure 2. Loading the Excel file in OpenRefine

- Create RDF structure from the dataset



OpenRefine CoHSI2 DATA xlsx Permalink

Facet / Filter   Undo / Redo 71 / 71

Using facets and filters

Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.

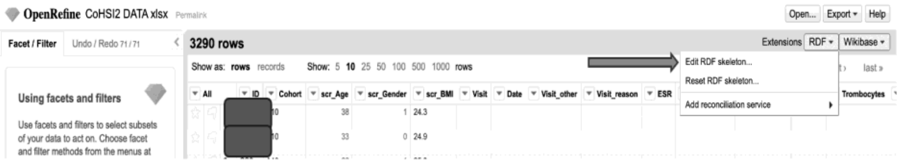
Not sure how to get started?
Watch these screencasts

3290 rows

Show as: **rows** records   Show: 5 **10** 25 50 100 500 1000 rows

All	ID	Cohort	scr_Age	scr_Gender	scr_BMI	Visit	Date	Visit_other	Visit_reason	ESR
1.	[redacted]	A10	38	1	24.3					
2.	[redacted]	A10	33	0	24.9					
3.	[redacted]	A10	22	1	25.3					
4.	[redacted]	A20	27	1	22.2					
5.	[redacted]	A20	35	1	27.9					
6.	[redacted]	A20	18	1	24.8					
7.	[redacted]	B20	32	0	24.9					
8.	[redacted]	B20	19	1	20.7					
9.	[redacted]	B20	19	1	24.5					
10	[redacted]	B20	30	0	22.0					

Figure 3. Cleaning and standardisation of column names and values



OpenRefine CoHSI2 DATA xlsx Permalink

Open... Export ▾ Help

Facet / Filter   Undo / Redo 71 / 71

Using facets and filters

Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at

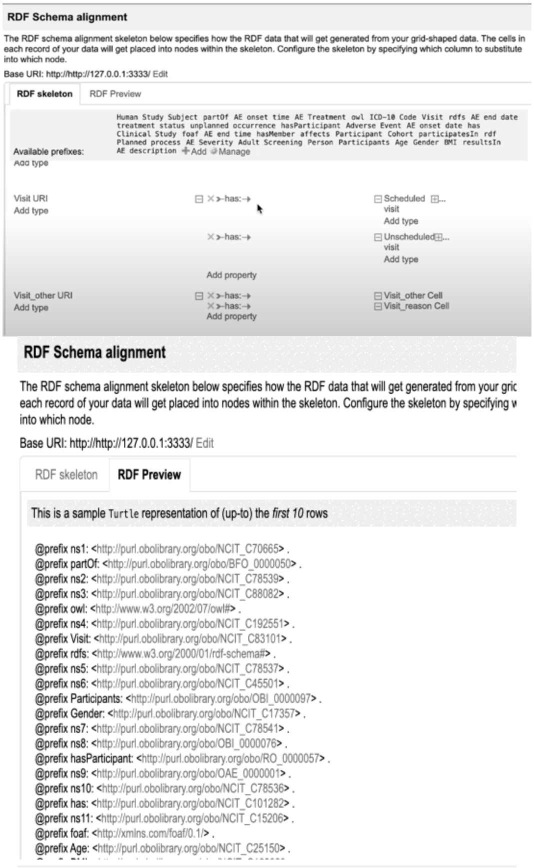
3290 rows   Extensions RDF ▾ Wikibase ▾

Show as: **rows** records   Show: 5 **10** 25 50 100 500 1000 rows

All	ID	Cohort	scr_Age	scr_Gender	scr_BMI	Visit	Date	Visit_other	Visit_reason	ESR	Trombocytes
	[redacted]	10	38	1	24.3						
	[redacted]	10	33	0	24.9						

Figure 4. Convert the input data in RDF

The RDF schema can now be generated in OpenRefine. The OpenRefine allows the RDF preview to open in a window. The RDF schema is exported into the OPenRefine to annotate the source data.



RDF Schema alignment

The RDF schema alignment skeleton below specifies how the RDF data that will get generated from your grid-shaped data. The cells in each record of your data will get placed into nodes within the skeleton. Configure the skeleton by specifying which column to substitute into which node.

Base URI: http://http://127.0.0.1:3333/ Edit

RDF skeleton | RDF Preview

Available prefixes: `Human Study Subject partOf AE onset time AE Treatment owl ICD-10 Code Visit rdfs AE end date treatment status unplanned occurrence hasParticipant Adverse Event AE onset date has Clinical Study foaf AE end time hasMember affects Participant Cohort participatesIn rdf Planned process AE Severity Adult Screening Person Participants Age Gender BMI resultsIn AE description` Add Manage

Add type

Visit URI
Add type
⊟ × has: → ⊟ Scheduled visit ⊞... Add type
× has: → ⊟ Unscheduled visit ⊞... Add type
Add property

Visit_other URI
Add type
⊟ × has: → ⊟ Visit_other Cell
× has: → ⊟ Visit_reason Cell
Add property

RDF Schema alignment

The RDF schema alignment skeleton below specifies how the RDF data that will get generated from your gric each record of your data will get placed into nodes within the skeleton. Configure the skeleton by specifying w into which node.

Base URI: http://http://127.0.0.1:3333/ Edit

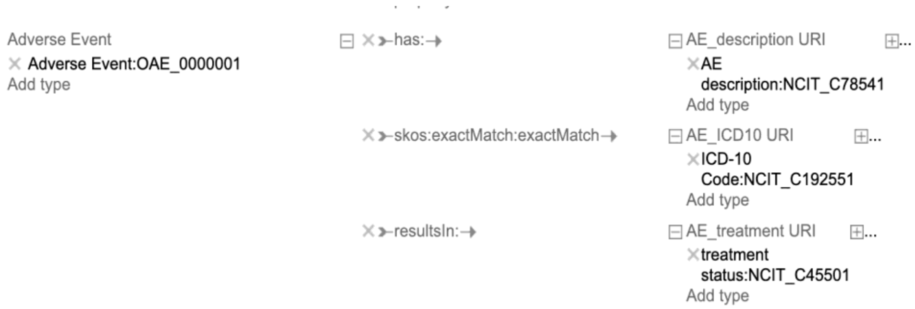
RDF skeleton | RDF Preview

This is a sample `Turtle` representation of (up-to) the *first 10* rows

```
@prefix ns1: <http://purl.obolibrary.org/obo/NCIT_C70665> .
@prefix partOf: <http://purl.obolibrary.org/obo/BFO_0000050> .
@prefix ns2: <http://purl.obolibrary.org/obo/NCIT_C78539> .
@prefix ns3: <http://purl.obolibrary.org/obo/NCIT_C88082> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix ns4: <http://purl.obolibrary.org/obo/NCIT_C192551> .
@prefix Visit: <http://purl.obolibrary.org/obo/NCIT_C83101> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix ns5: <http://purl.obolibrary.org/obo/NCIT_C78537> .
@prefix ns6: <http://purl.obolibrary.org/obo/NCIT_C45501> .
@prefix Participants: <http://purl.obolibrary.org/obo/OBI_0000097> .
@prefix Gender: <http://purl.obolibrary.org/obo/NCIT_C17357> .
@prefix ns7: <http://purl.obolibrary.org/obo/NCIT_C78541> .
@prefix ns8: <http://purl.obolibrary.org/obo/OBI_0000076> .
@prefix hasParticipant: <http://purl.obolibrary.org/obo/RO_0000057> .
@prefix ns9: <http://purl.obolibrary.org/obo/OAE_0000001> .
@prefix ns10: <http://purl.obolibrary.org/obo/NCIT_C78536> .
@prefix has: <http://purl.obolibrary.org/obo/NCIT_C101282> .
@prefix ns11: <http://purl.obolibrary.org/obo/NCIT_C15206> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix Age: <http://purl.obolibrary.org/obo/NCIT_C25150> .
```

Figure 5. RDF Schema

The OpenRefine will create a triple (subject-predicate-object) between the ontologies integrated into the data model. The RDF schema is related to a base URI.



Adverse Event
× Adverse Event:OAE_0000001
Add type

⊟ × has: → ⊟ AE_description URI ⊞...
× AE description:NCIT_C78541
Add type

× skos:exactMatch:exactMatch → ⊟ AE_ICD10 URI ⊞...
× ICD-10 Code:NCIT_C192551
Add type

× resultsIn: → ⊟ AE_treatment URI ⊞...
× treatment status:NCIT_C45501
Add type

Figure 6. Example 'Adverse Event' of predicate selection to complete triples generated in OpenRefine

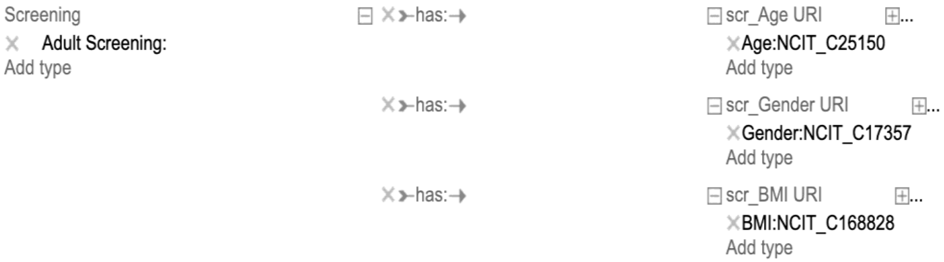


Figure 7. Example 'Screening' of predicate selection to complete triples generated in OpenRefine

The model is then exported into *.ttl format for the creation of the dashboard.

Steps in the creation of dashboard

Step 7: Upload in triple store

In the next step, the file needs to be uploaded to a triple store. We have chosen that the file is subsequently uploaded as a Turtle (.ttl) file into the Linked Data Hub. Terse RDF Triple Language (Turtle) is a format used in data science and linked data applications to represent structured information in a machine-readable way. It is a serialisation format for RDF, which is used to describe relationships between data entities in the Semantic Web and Linked Data.

This process involves navigating to the 'Datasets' or 'Graphs' section and selecting the option to upload RDF data. Once the upload is complete, the data can be visualised using the built-in graph views within the Linked Data Hub. An illustrative example demonstrating the interconnections between various metadata elements within the graph is presented in Figure 8.

The integration of source data through metadata provides a crucial analytical advantage. By expressing the data as RDF triples, it becomes possible to conduct more sophisticated queries that reveal intricate relationships. These queries can be performed in multiple ways: by identifying specific nodes and their connections to associated features; by recognising recurring patterns within the dataset; or by detecting unique feature combinations across the data.

The Linked Data Hub gives the possibility of showing the associations of the data in graphs or as a response to a query. These questions can be formulated on the API where of the Linked Data Hub. Other triple stores, such as AllegroGraph, have similar APIs that allow the data to be analysed and can also be used.

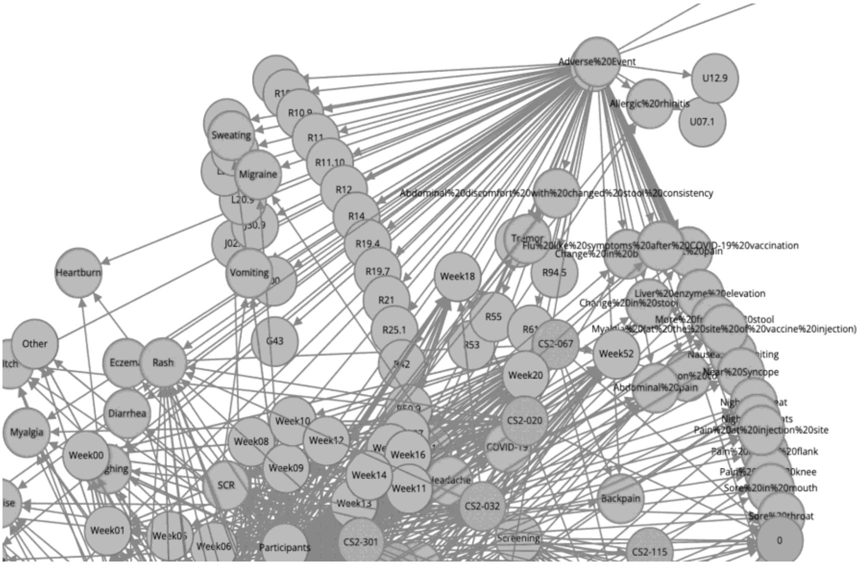


Figure 8. Interconnections between various metadata elements

The particular distribution to one of the adverse events is visualised in this graph with the node occurring on the right. This picture is blurred to avoid reidentification of the study participants.

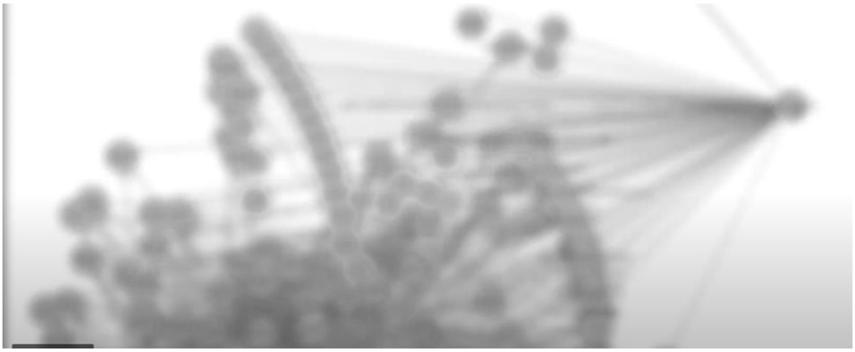


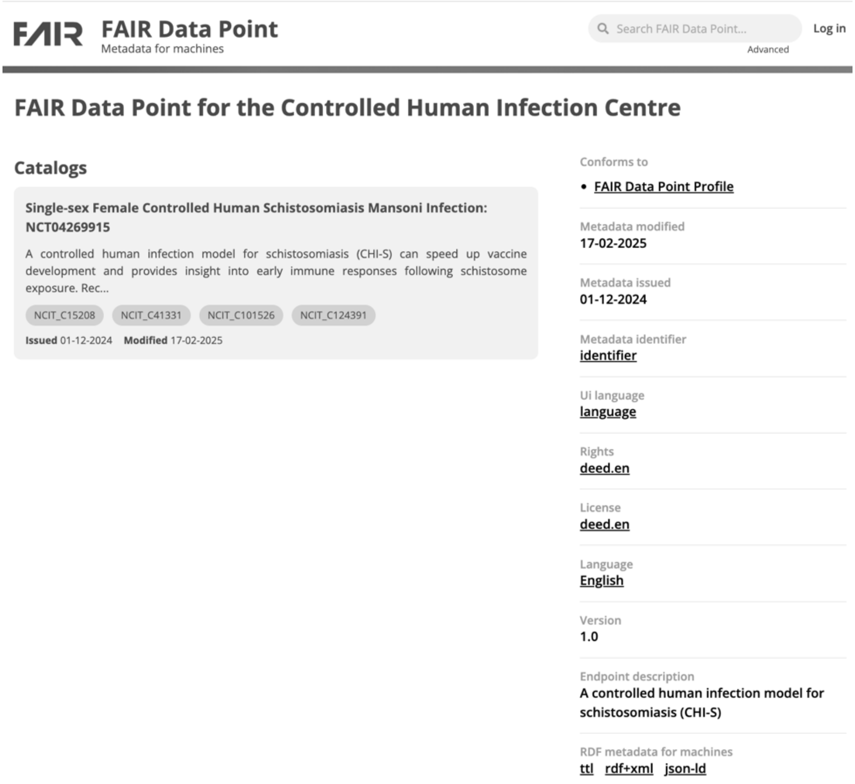
Figure 9. Node of adverse events (blurred to prevent reidentification)

Step 8: Setting up the FAIR Data Point

To set up the FAIR Data Point the workflow set out on https://github.com/FAIRDataTeam/FAIRDataPoint/ is followed.

The FAIR Data Point consists of several layers.

In the first layer, the catalogues within the FAIR Data Point are listed, along with their various characteristics, including the legal framework and licensing arrangements. Within the catalogue block, the purpose of the dataset is described, as well as the principal ontologies that define the catalogue. Additionally, this layer specifies the format of the data in Turtle (ttl), RDF, XML, and JSON-LD formats. New catalogues can be created by the FAIR Data Point controller, who can edit and add catalogues after logging in to the FAIR Data Point backend.



FAIR Data Point for the Controlled Human Infection Centre

Catalogs

Single-sex Female Controlled Human Schistosomiasis Mansoni Infection: NCT04269915

A controlled human infection model for schistosomiasis (CHI-S) can speed up vaccine development and provides insight into early immune responses following schistosome exposure. Rec...

NCIT_C15208 NCIT_C41331 NCIT_C101526 NCIT_C124391

Issued 01-12-2024 **Modified** 17-02-2025

Conforms to

- FAIR Data Point Profile

Metadata modified
17-02-2025

Metadata issued
01-12-2024

Metadata identifier
identifier

Ui language
language

Rights
deed.en

License
deed.en

Language
English

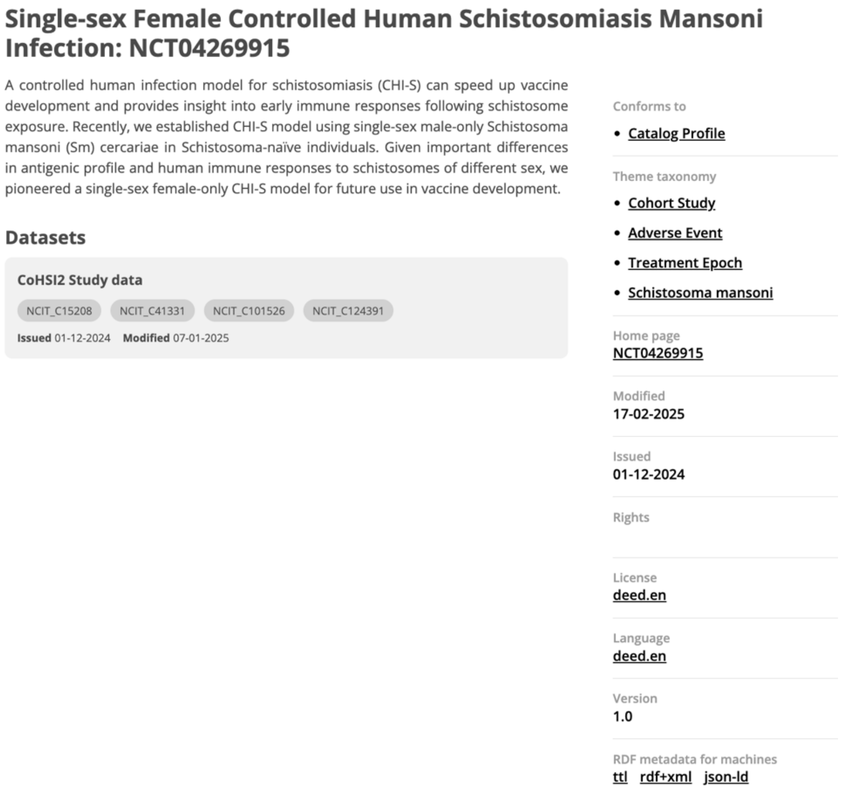
Version
1.0

Endpoint description
A controlled human infection model for schistosomiasis (CHI-S)

RDF metadata for machines
ttl rdf+xml json-ld

Figure 10. FAIR Data Point layer 1

The *Catalogue* is selected to navigate to the second layer of the FAIR Data Point. This layer provides specific information about the datasets associated with the catalogue, including a reference to the related study. The dataset is linked to a clinical trial study listed in the National Library of Medicine: https://clinicaltrials.gov/study/NCT04269915. New datasets, pertaining to the catalogue, can be added, by the FAIR Data Point controller.



Single-sex Female Controlled Human Schistosomiasis Mansoni Infection: NCT04269915

A controlled human infection model for schistosomiasis (CHI-S) can speed up vaccine development and provides insight into early immune responses following schistosome exposure. Recently, we established CHI-S model using single-sex male-only Schistosoma mansoni (Sm) cercariae in Schistosoma-naïve individuals. Given important differences in antigenic profile and human immune responses to schistosomes of different sex, we pioneered a single-sex female-only CHI-S model for future use in vaccine development.

Datasets

CoHSI2 Study data

NCIT_C15208 NCIT_C41331 NCIT_C101526 NCIT_C124391

Issued 01-12-2024 **Modified** 07-01-2025

Conforms to
- Catalog Profile

Theme taxonomy
- Cohort Study
- Adverse Event
- Treatment Epoch
- Schistosoma mansoni

Home page
NCT04269915

Modified
17-02-2025

Issued
01-12-2024

Rights

License
deed.en

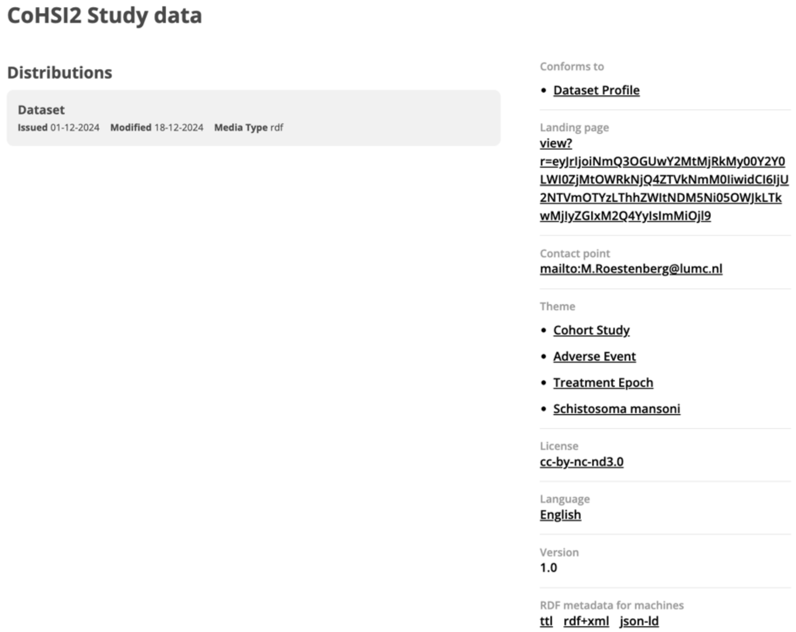
Language
deed.en

Version
1.0

RDF metadata for machines
ttl rdf+xml json-ld

Figure 11. FAIR Data Point layer 2

Navigating to the dataset opens the third layer of the FAIR Data Point. This brings us to the distribution of the dataset. The third layer provides detailed information about this specific dataset. Navigating to the landing page on the right side, opens the dashboard that provides the summary as well as details about the data's distribution. This will allow a third party interested in the data to inspect the data content, without having access to the metadata of the source data, or the source data itself. The page also gives the license pertaining to this dataset. Besides, it should provide information on the period that the dataset is available. The contact details of the researcher are also provided so that further information on the data can be obtained. The third layer is shown in Figure 12.



CoHSI2 Study data

Distributions

Dataset
Issued 01-12-2024 Modified 18-12-2024 Media Type rdf

Conforms to

- Dataset Profile

Landing page
view?
r=eyJrIjoiNmQ3OGUwY2MtMjRkMy00Y2Y0
LWI0ZjMtOWRkNjQ4ZTVkNmM0IiwidCI6IjU
2NTVmOTYzLThhZWItNDM5Ni05OWJkLTk
wMjIyZGIxM2Q4YyIsImMiOjl9

Contact point
mailto:M.Roestenberg@lumc.nl

Theme

- Cohort Study
- Adverse Event
- Treatment Epoch
- Schistosoma mansoni

License
cc-by-nc-nd3.0

Language
English

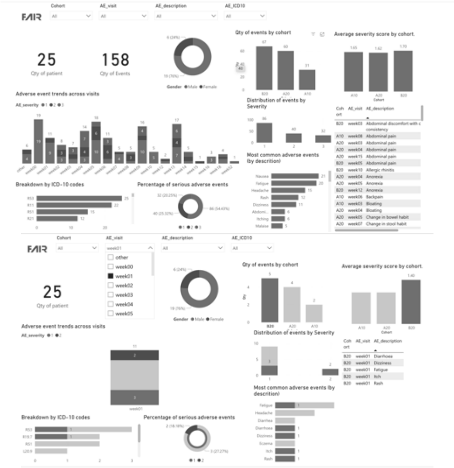
Version
1.0

RDF metadata for machines
ttl rdf+xml json-ld

Figure 12. FAIR Data Point layer 3

Step 9: The data dashboard

The data dashboard is essential for enabling viewers to understand the type of data contained in the dataset. Developed in Power BI, the dashboard provides insights into the clinical study. Its interactive features allow users to explore the distribution of the data in detail.



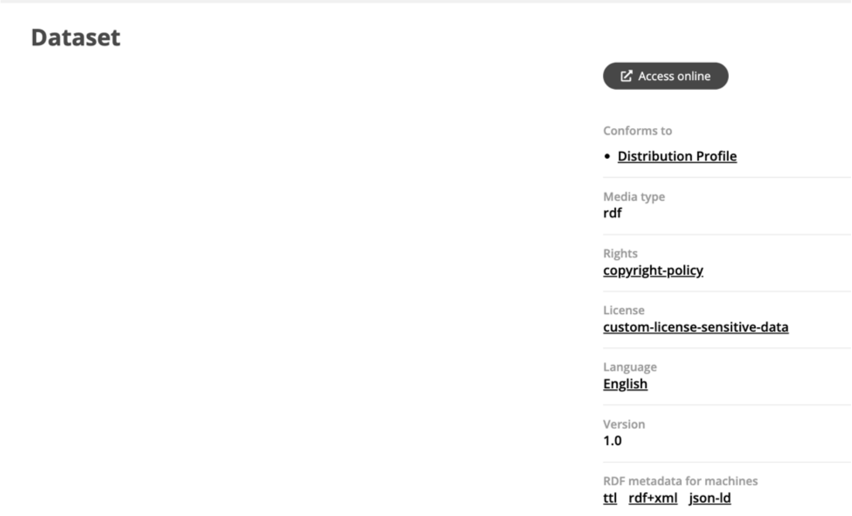
Cohort	AE_visit	AE_description
B20	week03	Abdominal discomfort with c consistency
A10	week08	Abdominal pain
A20	week03	Abdominal pain
A20	week04	Abdominal pain
A20	week15	Abdominal pain
B20	week05	Abdominal pain
B20	week10	Allergic rhinitis
A20	week04	Anorexia
A20	week05	Anorexia
B20	week12	Anorexia
A10	week06	Backpain
A10	week03	Bloating
A20	week04	Bloating
A20	week05	Change in bowel habit
A20	week07	Change in stool habit

Cohort	AE_visit	AE_description
B20	week01	Diarrhoea
B20	week01	Dizziness
B20	week01	Fatigue
B20	week01	Itch
B20	week01	Rash

Figure 13. Screenshots from Dynamic Dashboard in the FAIR Data Point

Step 10: Access to the triple store

To access the metadata in the triple store, navigation to the fourth layer is required by selecting the Dataset. This layer provides information on accessing the metadata stored in the triple store. By selecting Access Online, users can request a certificate for access. Additionally, the relevant rights and licenses governing data use are provided. The certificate grants access to the data in RDF format, enabling querying and analysis.



Dataset

Conforms to
- Distribution Profile

Media type
rdf

Rights
copyright-policy

License
custom-license-sensitive-data

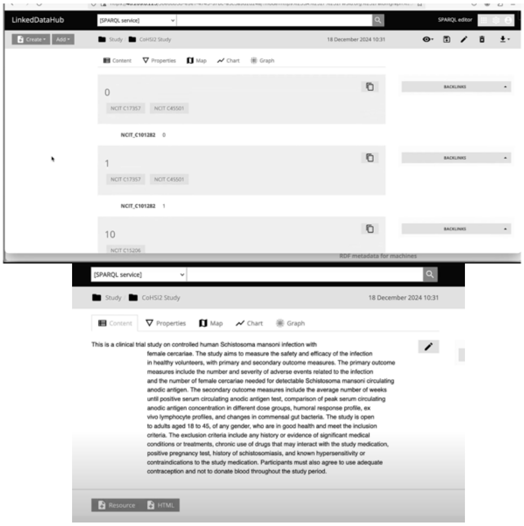
Language
English

Version
1.0

RDF metadata for machines
ttl rdf+xml json-ld

Figure 14. Access online through a certificate

Once the certificate is installed on the computer, the user gets access to the dataset in the Linked Data Hub. The API then provides a SPARQL query function, through which the data can be investigated.



This is a clinical trial study on controlled human Schistosoma mansoni infection with female cercariae. The study aims to measure the safety and efficacy of the infection in healthy volunteers, with primary and secondary outcome measures. The primary outcome measures include the number and severity of adverse events related to the infection and the number of female cercariae needed for detectable Schistosoma mansoni circulating anodic antigen. The secondary outcome measures include the average number of weeks until positive serum circulating anodic antigen test, comparison of peak serum circulating anodic antigen concentration in different dose groups, humoral response profile, ex vivo lymphocyte profiles, and changes in commensal gut bacteria. The study is open to adults aged 18 to 45, of any gender, who are in good health and meet the inclusion criteria. The exclusion criteria include any history or evidence of significant medical conditions or treatments, chronic use of drugs that may interact with the study medication, positive pregnancy test, history of schistosomiasis, and known hypersensitivity or contraindications to the study medication. Participants must also agree to use adequate contraception and not to donate blood throughout the study period.

Figure 15. API of the Linked Data Hub

The access to the Triple store can be mediated by approved queries, with permission provided prior to the running of the queries, and such a query could be run over multiple Linked Data stores as a federated query with prior approval.

The FAIR Data Point metadata-structure

The FAIR Data Point structures different types of metadata about the research, the dataset, and the source data. Metadata plays a crucial role in structuring, describing, and managing data across different layers of research information systems. The metadata of research, the data catalogue, and the source data each serve distinct but interconnected functions.

The metadata of research provides essential context and provenance for a study, ensuring reproducibility, credibility, and proper attribution. It serves a descriptive function by including information such as the research title, authors, affiliations, abstract, keywords, and publication details. Additionally, it fulfils a provenance function by documenting the origin of the research, including its methodology, data sources, ethical considerations, and funding. Metadata also supports discovery by enabling indexing in repositories and databases, making the research findable through search engines and linked data applications. Furthermore, it plays an interoperability role by ensuring compatibility with metadata standards such as Dublin Core, DataCite, or schema.org, which facilitate integration with other research outputs.

The metadata of the data catalogue functions as a structured inventory of datasets, organising and categorising them to enhance discoverability and governance. It plays an organisational role by grouping datasets based on themes, disciplines, projects, or institutions. The metadata also ensures standardisation by defining common attributes, such as dataset descriptions, versioning, licensing, and access conditions. Additionally, it enforces access control by specifying who can use the data, under what conditions, and through which mechanisms, such as open access, embargo, or restricted use. Another crucial function of the metadata in a data catalogue is interlinking, as it connects datasets to related studies, ontologies, and repositories through persistent identifiers such as Digital Object Identifier (DOIs), IRIs, or URIs. When data is linked using consistent identifiers, machines can process and combine information from different sources, enhancing data integration.

The metadata of the source data ensures data integrity, traceability, and machine-actionability at the dataset level. It serves a structural function by defining the format, schema, and data model, such as ttl, RDF, XML, and JSON-LD. Moreover, it has a semantic function by using ontologies and vocabularies to provide meaning to data fields, ensuring consistent interpretation across systems. The metadata also supports quality control by including information on data accuracy, completeness, version history, and provenance tracking. Additionally, it serves a technical function by describing how the data is stored, processed, and accessed, including details on APIs or Query Language and Resource Description Framework (SPARQL) endpoints.

In the FAIR Data Point, these three layers of metadata are interconnected in a way that enhances research and data management. Research metadata provides high-level context and publication details, while data catalogue metadata organises and makes datasets discoverable. Meanwhile, source data metadata ensures that the data remains usable and machine-readable. Together, these layers support the FAIR principles, ensuring that data is Findable, Accessible, Interoperable, and Reusable, thereby enhancing data-driven research and collaboration.

Discussion

In this study the FAIRification process of a full dataset was documented. The CoHSI2 source data was already generated in Castor and well structured. Castor allows the output in different formats. In this workflow the data was included in an excel sheet and FAIRification started from there. The FAIRification steps included:

- Adding metadata descriptors of the overall study using Dublin Core or DCAT.
- Assigning persistent identifiers (PIDs) to datasets.
- Mapping terminology semantically describing data to standardised ontologies (e.g., OBO) and machine-readable identifiers.
- Transforming datasets into structured formats (e.g., RDF, JSON-LD).
- Clean and enrich the source data with metadata organised in the data mappings.
- Creating a data dashboard for public access to give insights in the source data available
- Repositing the data and metadata in a triple store.
- Hosting data catalogues and datasets metadata in an FDP software with API access.
- Providing access to the metadata of the source data through an access mechanism (certificate)
- Provide querying access to the metadata of the source data through a triple store API.

The workflow of this study followed a FAIRification by Increment approach. This workflow preserves and enhances the value of legacy datasets by enabling the gradual adoption of FAIR principles without disrupting ongoing research activities. The result of the FAIRification process supports data reuse and facilitates the integration of FAIR workflows across disciplines. Over time, this approach encourages the gradual implementation of FAIR principles, potentially leading to a FAIR-by-design approach in future research. Steps in data preparation become clearer earlier in the process, potentially allowing for better mapping and planning of data collection.

The source data included sensitive data and personal data, with the possibility of deidentification due to the limited number of participants in the study. Therefore, particular care was given to make sure that the visualisation of the data and the graph do not show any information through which a natural person can be identified. The conditions for access to the triple store need to be managed by strict data use agreements that ensure the privacy preservation of the participants in the study.

This study took an implementation approach in real life, with real source data developed in a real clinical study. This allowed contextual factors that influenced the workflow to emerge explicitly. Delays were caused by the lack of clarity of what characteristics of workplaces were needed for a secure and workable environment for FAIRification. The conditions and costs of such workplaces and of tools were another challenge. The intensity of different steps, not yet fully supported by FSR, also caused delays. While a full FDP was achieved, the difficulties involved resulted in hesitation among researchers on the practicality of FAIRification and the costs versus benefits of it.

Due to the difficulties involved in having a public academically supported workspace, this study made use of commercial FSRs, such as the PowerBI dashboard, the Amazon Web Services (AWS) droplet, and Azur for storage, which all depend on digital backbones outside Europe. While these are good tools with high functionalities and easy-to-use structures, a future installation could focus on creating greater independence of these tools based and managed in the US. This would decrease undesirable dependencies, reduce costs, and increase efficiencies.

The study shows that the workflows are implementable, and FAIRification is at a point where it can be routinely embedded in data-handling workflows. More supporting resources are available, as is training and practical support to go about the different steps. It is also expected that more data handling in research will plan the FAIRification from the outset, allowing a FAIRification by Design or De Novo FAIRification which will be advantageous for data quality. De Novo FAIRification will eliminate the need for later remediation, ensuring long-term data quality, interoperability, and reusability. This approach is particularly beneficial in highly regulated domains such as biomedical research, clinical data management, and digital health ecosystems. This will also make data handling processes more efficient, and it will decrease the dependency on highly specialised FAIR data stewards.

While this study was performed by a FAIR data expert, dependency on such experts for popularising FAIRification is not always feasible, due to the scarcity of these experts and the costs-related aspects. While the documentation of it shows reproducibility and easy-to-handle steps for data stewards, it is expected that FAIR experts will be needed to support the data handling processes. Some of the more complex steps in the FAIRification process will continue to require the input of FAIR specialists, especially the understanding of available ontology resources, and the understanding of alternative workflows and steps that may resolve challenges or respond better to specific situations.

More research is needed on the potential for the reuse of data in clinical studies. This could not be investigated based on just one dataset. Also, the interoperability of data insights across different disciplines and fields, all with their own semantics, and different research philosophies, remains a question open for investigation in the future.

Conclusion

This study investigates a FAIRification process of data from a controlled human infection study with *Schistosomiasis mansoni* following a FAIR by increment approach. Existing source datasets were assessed, and these were progressively improved to increase FAIR compliance, resulting in a fully fledged FAIR Data Point, where data can be inspected and access to the metadata and source data can be obtained.

The FAIRification of the data is a critical contribution to ensure data can be verified, to enhance the reliability of publications and to ensure that the value of the data is maintained for future reference, including its reuse for other studies. The installation of the FAIR Data Point is facilitated by the emergence of new tools, which facilitate the retrospective FAIRification of legacy data. It is recommended to integrate FAIR processes in the methodology of data collection, standardising collection, recording methods, and creating FAIRification as an integral part of the data handling from the start of the study.

The installation of FAIR Data Points still has challenges, which has caused a delay in their deployment. As a result, tests of interoperability across FAIR Data Points require more work. Also refining the access, control, and permission mechanisms are still lacking granular capabilities. This is important among others to ensure the security of the data as well as the privacy preservation of sensitive data that might be identified in a dataset. More development is needed to strengthen the concept of the FAIR Data Points, to promote uptake and adoption in the scientific community. The overall conclusion of this study is that FAIR Data Points form a critical contribution towards greater academic transparency and accountability.

Acknowledgments

The authors thank Prof. Dr. Roestenberg and her team for spearheading the efforts to establish a FAIR Data Point for the clinical data.

The authors thank Olzhas Aldabergenov for the support provided during the implementation of this study.

Authors' Contributions

Aliya Aktau conducted this study as part of her PhD research. She wrote the first version of this chapter and edited all the drafts. **Mirjam van Reisen** reviewed all the versions and provided suggestions and comments and contributed to the text.

Ethical Considerations

Tilburg University, Research Ethics and Data Management Committee of Tilburg School of Humanities and Digital Sciences REDC#2020/013, June 1, 2020-May 31, 2024, on Social Dynamics of Digital Innovation in remote non-western communities. Uganda National Council for Science and Technology, Reference IS18ES, July 23, 2019-July 23, 2023.

References

Amare, S. Y. (2025). Bridging borders with FAIR data: Strengthening maternal health and public health surveillance in Africa. In M. Van Reisen, S. Y. Amare, & M. Mawere (Eds.), *FAIR data, FAIR Africa, FAIR world: Internationalisation of the Health Data Space*. Bamenda: Langaa

Amare, S. Y., Taye, G. T., Plug, R. B. F., Medhanyie, A. A., & Van Reisen, M. (2025). Federating tools for FAIR patient data: Strengthening maternal health and infectious disease surveillance from clinics to global systems. In Van Reisen, M., Amare, S. Y., Maxwell, L. & Mawere, M. (Eds.), *FAIR data, FAIR Africa, FAIR world: Internationalisation of the Health Data Space.* Bamenda: Langaa

American University of Nigeria. (n.d.). *FAIR data points.* Retrieved from https://aun.mu.edu.et/ahds/fair-data-points/

BioPortal. (n.d.). *COHSI2STUDY ontology.* Retrieved from https://bioportal.bioontology.org/ontologies/COHSI2STUDY?p=classes

Bonino da Silva Santos, L. O., Wilkinson, M. D., Kuzniar, A., Kaliyaperumal, R., Thompson, M., Dumontier, M., & Burger, K. (2016). FAIR data points supporting big data interoperability. In M. Zelm, G. Doumeingts, & J. P. Mendonça (Eds.), *Enterprise interoperability in the digitized and networked factory of the future*. ISTE Press.

ClinicalTrials.gov. (n.d.). *NCT04269915 study.* Retrieved from https://clinicaltrials.gov/study/NCT04269915

GitHub. (n.d.). *FAIRDataPoint.* Retrieved from https://github.com/FAIRDataTeam/FAIRDataPoint/

Groenen, K. H. J., Jacobsen, A., Kersloot, M. G., dos Santos Vieira, B., van Enckevort, E., Kaliyaperumal, R., Arts, D. L., 't Hoen, P. A. C., Cornet, R., Roos, M., & Kool, L. S. (2021). The De Novo FAIRification process of a registry for vascular anomalies. *Orphanet Journal of Rare Diseases, 16*(1), 376. https://doi.org/10.1186/s13023-021-02004-y

Jacobsen, A., Kaliyaperumal, R., Bonino da Silva Santos, L. O., Mons, B., Schultes, E., Thompson, M., & Roos, M. (2019). A generic workflow for the data FAIRification process. *Data Intelligence, 1*(1–2), 56–65. https://doi.org/10.1162/dint_a_00020

Koopman, J. P. R., Houlder, E. L., Janse, J. J., Casacuberta-Partal, M., Lamers, O. A. C., Sijtsma, J. C., de Dood, C., Hilt, S. T., Ozir-Fazalalikhan, A., Kuiper, V. P., Roozen, G. V. T., de Bes-Roeleveld, L. M., Kruize, Y. C. M., Wammes, L. J., Smits, H. H., van Lieshout, L., van Dam, G. J., van Amerongen-Westra, I. M., Meij, P., Corstjens, P. L. A. M., … Roestenberg, M. (2023). Safety and infectivity of female cercariae in Schistosoma-naïve, healthy participants: A controlled human Schistosoma mansoni infection study. *EBioMedicine, 97*, 104832. https://doi.org/10.1016/j.ebiom.2023.104832

Kersloot, M. G., Jacobsen, A., Groenen, K. H. J., dos Santos Vieira, B., Kaliyaperumal, R., Abu-Hanna, A., … Arts, D. L. (2021). De-novo FAIRification via an Electronic Data Capture system by automated transformation of filled electronic Case Report Forms into machine-readable data. *Journal of Biomedical Informatics*, *122*. https://doi.org/10.1016/j.jbi.2021.103897

Lin, Z. (2025a). *A scalable approach for De Novo FAIRification in legacy systems: Enabling real-time RDF transformation, semantic integration, and automated data upload* [Doctoral dissertation, Leiden University]. Leiden University Repository.

Lin, Z. (2025b) Implementation of De Novo FAIRification in Relational Legacy Systems: the Case of The Electronic Medical Record system for maternal health in Afya.ke. In Van Reisen, M., Amare, S. Y., Maxwell, L. & Mawere, M. (Eds.), *FAIR data, FAIR Africa, FAIR world: Internationalisation of the Health Data Space.* Bamenda: Langaa

Mroestenberg. (n.d.). *Home.* Retrieved from http://mroestenberg.com/

Mroestenberg. (n.d.). *Profile.* Retrieved from http://mroestenberg.com/profile/77aaad6a-0136-4c6e-88b9-07ffccd0ee4c

NL72661.058.20 / CoHSI2. (2021, January 22). *Establishing a female-only controlled human Schistosoma mansoni infection model: A safety and dose finding study (CoHSI2) Version 2.0*. Retrieved from https://www.thelancet.com/cms/10.1016/j.ebiom.2023.104832/attachment/fcb0f951-42e7-41ee-ad4a-6af5ad032689/mmc2.pdf

Power BI. (2025). *Data visualization.* Retrieved from https://app.powerbi.com/view?r=eyJrIjoiNmQ3OGUwY2MtMjRkMy00Y2Y0LWI0ZjMtOWRkNjQ4ZTVkNmM0IiwidCI6IjU2NTVmOTYzLThhZWItNDM5Ni05OWJkLTkwMjIyZGIxM2Q4YyIsImMiOjl9

Stocker, M., Stokmans, M., Van Reisen, M.: Agenda setting on FAIR Guidelines in the European Union and the role of expert committees. Data Intelligence 4(4) (2022). doi: 10.1162/dint_a_00168

Taye, G. T., Amare, S. Y., Gebreslassie, T. G., Jati, P. H. P., Kawu, A. A., Kievit, R., Plug, R., Smits, K., Stocker, J., & Van Reisen, M. (2024). Training Manual: A practical guide to FAIR Data Stewardship. Brussels: EEPA. VODAN. https://aun.mu.edu.et/ahds/training/

VODAN Africa. (2021, March 22). *Completion of a FAIR Data Point December 2024 - Aliya Aktau, LUMC, Leiden University.* YouTube Retrieved from https://www.youtube.com/watch?v=15PmyFO8iFs

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., Bonino da Silva Santos, L., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data, 3*, 160018. https://doi.org/10.1038/sdata.2016.18

[i] This chapter is updated on 7 May 2025 to reflect changes following comments received by Roestenberg. The changes focus particularly on the title of the study, identified in the FAIR Data Point which was changed from COHSI2 to 'schistosome controlled human infection' study.