Original Paper

David Amadi[1], Sylvia Kiwuwa-Muyingo[2], Tathagata Bhattacharjee[1], Amelia Taylor[3], Agnes Kiragga[2], Michael Ochola[2], Chifundo Kanjala, Arofan Gregor[4], Keith Tomlin[1], Jim Todd[1], Jay Greenfield[4]

Author Affiliation

[1] Department of Population Health, Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, UK

[2] African Population and Health Research Center, Nairobi, Kenya

[3] Malawi University of Business and Applied Sciences, Blantyre, Malawi

[4] Committee on Data (CODATA), Paris, France

Corresponding Author:

David Amadi, MSc,

London School of Hygiene and Tropical Medicine,

Department of Population Health,

Keppel Street London WC1E 7HT United Kingdom,

Phone: +44 20 7636 8636

Email: david.amadi@lshtm.ac.uk

Making metadata machine-readable as the first step to FAIR population health data

Abstract

Background

Metadata describes and provides context for other data and plays a pivotal role in enabling the FAIR (Findability, Accessibility, Interoperability, and Reusability) data principles. By providing comprehensive and machine-readable descriptions of digital resources, metadata empowers both machines and human users to seamlessly discover, access, integrate, and reuse data or content across diverse platforms and applications. However, the limited accessibility and machine-interpretability of existing metadata for population health data hinder effective data discovery and reuse.

Objective

To address these challenges, we propose a comprehensive framework utilizing standardized formats, vocabularies, and protocols to render population health data machine-readable, significantly enhancing its FAIRness and enabling seamless discovery, access, and integration across diverse platforms and research applications.

Methods:

The framework implements a three-stage approach:

1. DDI (Data Documentation Initiative) Integration: Leveraging the DDI Codebook metadata, detailed information for data and associated assets is documented, ensuring transparency and comprehensiveness.
2. OMOP CDM (Observational Medical Outcomes Partnership Common Data Model) Standardization: Data is harmonized and standardized into the OMOP CDM, facilitating unified analysis across heterogeneous datasets.
3. Schema.org and JSON-LD (JavaScript Object Notation for Linked Data) Integration: Machine-readable metadata is generated using Schema.org entities and embedded within the data using JSON-LD, boosting discoverability and comprehension for both machines and human users.

We demonstrated the implementation of these three stages using the infectious disease surveillance and response (IDSR) data from Malawi and Kenya.

Results

The implementation of our framework significantly enhanced the FAIRness of population health data, resulting in improved discoverability through seamless integration with platforms like Google Dataset Search. The adoption of standardized formats and protocols streamlined data accessibility and integration across various research environments, fostering collaboration and knowledge sharing. Additionally, the utilization of machine-interpretable metadata empowered researchers to efficiently reuse data for targeted analyses and insights, thereby maximizing the overall value of population health resources. The JSON-LD codes are accessible via GitHub repository, and the HTML code integrated with JSON-LD is available on the The Implementation Network for Sharing Population Information from Research Entities (INSPIRE) website.

Conclusion

The adoption of machine-readable metadata standards is essential for ensuring the FAIRness of population health data. By embracing these standards, organizations can enhance diverse resource visibility, accessibility, and utility, leading to a broader impact, particularly in low- and middle-income countries (LMICs). Machine-readable metadata can accelerate research, improve healthcare decision-making, and ultimately promote better health outcomes for populations worldwide.

Keywords: FAIR Data Principles; Metadata; Machine-Readable Metadata; DDI (Data Documentation Initiative); Standardization; JSON-LD; OMOP CDM; Data Science; Standards; Data Models;

Introduction

Population health data play a crucial role in understanding the dynamics of public health and informing evidence-based policies and interventions [1]. In low- and middle-income countries (LMICs), two common approaches for systematically and continuously monitoring health indicators in the population over time are the disease surveillance systems and the health and demographic surveillance systems (HDSS). Surveillance systems are designed to detect and respond to outbreaks of infectious diseases while HDSS are longitudinal systems that collect health and demographic data on a defined population [2], [3].

Despite the availability of valuable population health data from these systems, there remains a significant challenge in effectively sharing this information across research entities. Finding data and knowledge or information about population health requires accurate and comprehensive documentation of metadata, often referred to as "data about data", but many studies are unsure about how to do this [4]. Disease surveillance metadata encompasses crucial details, including the diseases under surveillance, the geographical areas covered, and the variables measured, each with defined formats and explanations. For example, variables may include disease incidence rates, demographic characteristics, and healthcare utilization. Additionally, a data dictionary provides comprehensive definitions and formats for each variable, aiding in the interpretation and utilization of surveillance data[5].

Metadata is a critical component of achieving the FAIR (Findable, Accessible, Interoperable, and Reusable) data principles[6]. The integration of these principles into data science aligns with the original vision of Wilkinson and colleagues, which goes beyond the traditional reuse of data but also other digital research components like inputs, outputs, algorithms, tools, and workflows that generate data [6], [7]. To realize this vision, it requires the use of not just one but several standards depending on the use case [6]. By employing metadata standards like Data Documentation Initiative (DDI), Dublin Core, and the Common Data Model (CDM) for representing exposures and outcomes, along with a domain-specific vocabulary for annotating these, we ensure the FAIRness of our data collection process [8], [9].

DDI Codebook and DDI Lifecycle serve as an international standard for systematically detailing data generated through surveys and observational methods, ensuring consistent documentation of content, structure, and provenance. This enhances the accessibility, discovery, and preservation of data and consistency in representation of data by users [10]. Notable portals like INDEPTH Data Repository (iSHARE), SAPRIN Data, and APHRC's microdata portal employ DDI for public data dissemination [11], [12], [13].

The Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) addresses the challenge of integrating and analyzing diverse healthcare data by providing a standardized information model. This model acts as a universal language, enabling seamless integration and consistent analysis of data from various home and clinic encounters [14], [15].

The core of the OMOP CDM's model lies in its well-defined structure comprising of 39 tables categorized into relevant healthcare domains [16]. These domains include standardized vocabularies, person-centric data (e.g., demographics, diagnoses), and standardized health system data (e.g., procedures, medications). This organization ensures consistency throughout the data and facilitates downstream analyses [15].

Furthermore, the standardized structure allows efficient data preparation through ETL processes for analysis with various tools. This facilitates uniform analysis techniques across studies. Additionally, standardized vocabulary enables domain-specific labeling of interventions and outcomes, crucial for machine learning and metadata documentation within the CDM framework [17].

OMOP CDM prioritizes ethical considerations by sharing de-identified and aggregated data, enabling network-wide analysis without sharing patient-level information. This promotes

transparency, reduces bias, and aligns with data protection regulations, while enabling autonomous data sharing and safeguarding individual privacy with data remaining secure at the source [18].

While data standardization and harmonization are essential for FAIR compliance, they are not sufficient on their own [6]. Another step involves rendering metadata machine-readable and more findable online, aligning with the objectives of Schema.org. Schema.org is a collaborative project by major search engines like Google to create a schema or shared vocabulary dedicated to developing and establishing metadata standards that enhance the discoverability and indexing of online content, including assets in population health and clinical data[19]. It has applicability across a vast range of domains, but its use in population health data remains underexplored. Nevertheless, harnessing schema.org offers the potential to significantly streamline data discovery efforts, as evidenced by its adoption by over 10 million websites [19].

For this reason, we are considering using JSON-LD (JavaScript Object Notation for Linked Data). JSON-LD is a lightweight Linked Data format capable of marking up internet content with metadata. It initially utilized schema.org entities like Action, BioChemEntity, CreativeWork, Event, MedicalEntity, Organization, Person, and Place, but JSON-LD's capabilities extend beyond these entities. It can create complex machine-readable documents that can seamlessly transition between different sets of metadata objects. In our use case and others, JSON-LD demonstrates the potential to FAIRify the arc of data science by enabling seamless interoperability and data sharing across diverse systems and platforms [20].

However, the current practice of FAIR implementation in population health faces substantial barriers. Obstacles include the limited availability of mature FAIR technology, the proliferation of diverse digital tools, and data often being locked within local formats or proprietary standards imposed by electronic health record (EHR) vendors. Moreover, these challenges are compounded by a prevailing siloed data mindset among the key stakeholders collecting population health data [21].

This paper proposes an approach to bridge the gap between FAIR principles and practice in population health data by incorporating machine-readable metadata for data and non-data digital assets like platforms into the population health dataset. We use the Infectious Disease Surveillance and Response (IDSR) data as a case study to demonstrate the processes. Despite its critical role in public health, the current structure of IDSR data presents a challenge in adhering to FAIR principles. Key information regarding data collection methodologies, access restrictions, and updates often lacks proper documentation or resides scattered across diverse formats. This fragmentation severely hinders the data's findability, accessibility, and interoperability, posing substantial obstacles to its effective utilization and compliance with FAIR standards. To address this, we leverage established standards like DDI and OMOP CDM and explore the potential of Schema.org entities alongside other standards using JSON-LD to extend and adapt these principles. This strategic combination of machine-readable metadata, standards like Schema.org, and other standards coupled with JSON-LD is a crucial step towards achieving *comprehensive* FAIR compliance in population health data.

The proposed approach integrates the collaborative frameworks of the GO-FAIR and WorldFAIR initiatives to promote interdisciplinary collaboration, culture change, and technology integration for the effective implementation of FAIR principles[22], [23]. This effort addresses the longstanding challenge of restricted data discoverability, which hinders the effective use, reuse, integration, and knowledge integration of data [24]. Data stewardship plans mandate that data, along with all assets generated from public funds, should be publicly accessible [25]. Good data stewardship practices, particularly those adhering to FAIR principles, significantly enhance data discoverability and facilitate its improved reuse. Universal access to health research data, regardless of location or resource limitations, is crucial for advancing research and supports the broader goal of promoting healthy lives

and well-being for all at all ages, as outlined in Sustainable Development Goal 3 (SDG 3) for improving global health and achieving sustainable development goals [26], [27].

Methods

We propose a step-by-step guide to making metadata FAIR using the IDSR database, which has been collected in population settings and HDSS sites in Africa, as a demonstration of our use case. Our methodology is built upon a flexible, multi-level, domain agnostic FAIRification framework, providing practical guidance to improve the FAIRness for both existing and future datasets. This framework encompasses three stages: first, we integrate DDI metadata, followed by the implementation of OMOP CDM, and finally, we leverage schema.org to refine metadata structure and accessibility [28]. We use the IDSR data as a case study, which is recommended by the World Health Organization (WHO) but may be implemented differently from one country to another.

To effectively implement the DDI framework for population health data, we first selected the National Data Archive (NADA) online catalog. This metadata repository adheres to the DDI 2 Codebook and Dublin Core Extensible Markup Language (XML) metadata standards [29]. It serves as a comprehensive platform for searching, comparing, applying for access to, and downloading metadata, datasets, questionnaires, and reports. NADA plays a pivotal role in ensuring the accessibility, discoverability, reusability, and collaboration of population health data. This is achieved by configuring the open-source web-based data cataloging application according to the guidelines provided in the NADA documentation [30].

Nesstar Publisher is used to create the DDI Codebook, a structured and descriptive document that captures essential metadata elements, which integrates the DDI framework into the IDSR dataset [31]. With the help of the International Household Survey Network (IHSN) metadata template and step-by-step guide, we describe the attributes of the IDSR data in the codebook, providing a rich and informative metadata record, including document description, study description, data file description, variable description, and additional materials [32], [33]. More advanced versions of DDI, such as Lifecycle, have the potential to extend this integration to longitudinal studies, such as HDSS and other cohort studies, accommodating distinct waves or rounds of data collection [34], [35].

The Implementation Network for Sharing Population Information from Research Entities (INSPIRE) developed ETL programs to upload the source data and metadata into the OMOP CDM, harmonizing and standardizing the data for unified data analysis in research using open-source OHDSI tools, which extend beyond its conventional use in clinical data[36], [37].

This builds on the methods begun by European Health Data & Evidence Network (EHDEN) [38] to grow the description of the IDSR data and metadata as a use case within the OMOP CDM. The process of FAIRification begins by identifying various digital resources present within the OHDSI artifacts, including *protocols*, *databases*, *study results*, *controlled vocabularies*, *software libraries*, and any other relevant digital assets. The metadata was described using schema.org, expressed in JSON-LD format.

The standard CDM has a structure shown in Fig 1. The yellow boxes are literals that describe elements of the schema.org MedicalObservationalStudy, including the title, identifier, database, study status, and type. The orange boxes describe a set of classes, or concepts that are contained in the dataset. These classes use standard vocabulary concepts that describe variables in the study, including the risk factors and exposures (schema:MedicalEntity) and the conditions reported (schema:MedicalCondition) in connection with them. The left of the figure shows the analyses carried out on the data. The lower part of the model outlines the initial step of generating the OMOP CDM instance through an action that uses the ETL process leveraging the Pentaho platform and SQL to integrate and process data from the IDSR source dataset or synthetic data.

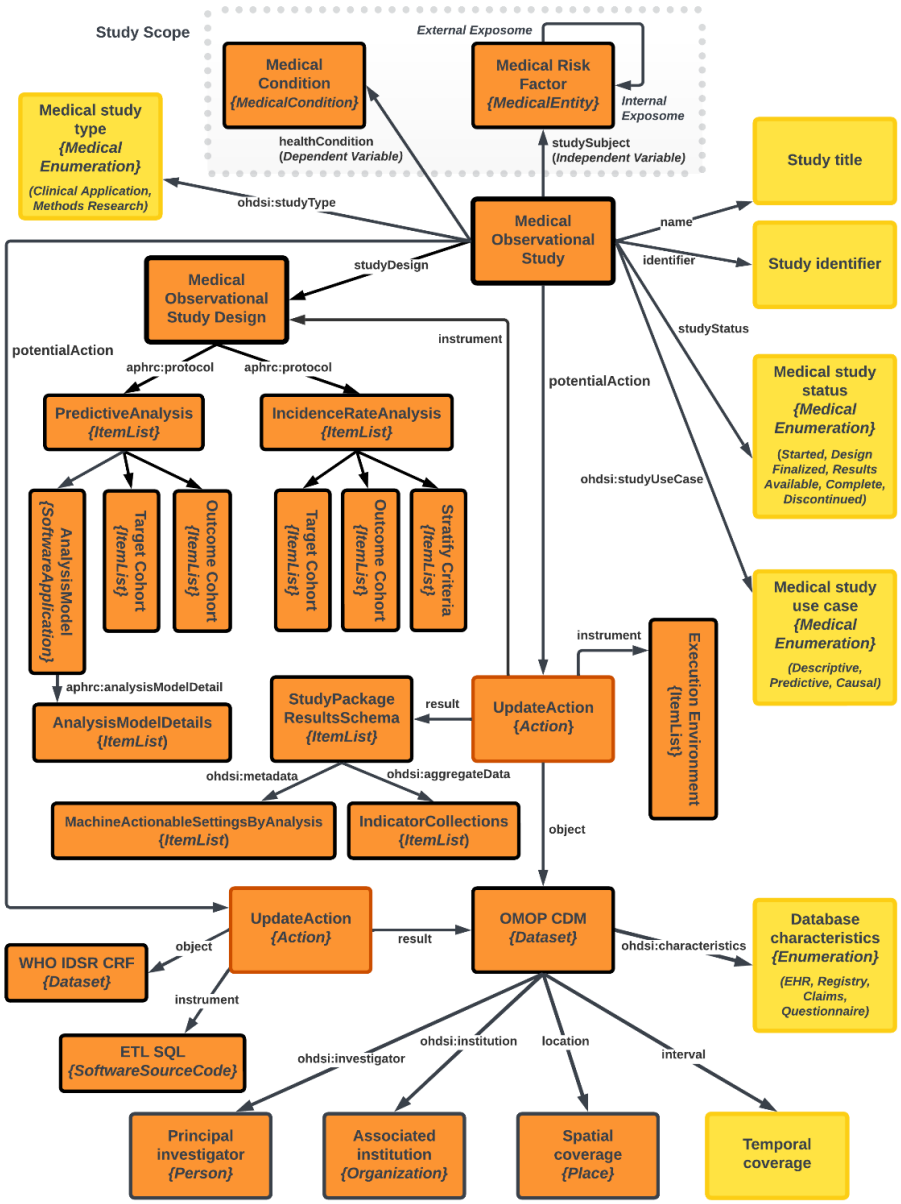


Figure 1: INSPIRE Model

The validity of the generated JSON-LD files can be verified using the schema.org validator, a user-friendly tool capable of validating JSON-LD code. Validation can be done by either submitting a URL that points to the JSON-LD file or by directly pasting the JSON-LD code into the tool. The JSON-LD is then embedded within the HTML code, enabling search engines such as Google [39] to effectively utilize this structured data for enhanced data discovery and comprehension purposes.

Results

DDI codebook

The DDI codebook presented through the NADA repository provides a comprehensive and structured documentation framework for the IDSR source data. This resource is essential for researchers seeking a comprehensive understanding of the dataset. For IDSR codebook refer to supplementary material (ddi_documentation.pdf) and is logically organized into four main sections:

Table 1: Catalog sections:

Section	Description
Study description	The Study section of the DDI codebook provides a comprehensive overview of the IDSR study, including its title, purpose, methodology, coverage, producers and sponsors, disclaimer, and copyright, and contact details.
Documentation	This section of the DDI codebook includes the WHO IDSR questionnaire for Malawi, Kenya, and Uganda, as well as other relevant documentation for the IDSR study, such as the study protocol.
Data description	Provides a detailed description of the IDSR data set, including the data files from Malawi, Kenya, and Uganda. The codebook provides detailed descriptions of all variables in the dataset, including their names, labels, definitions, and coding schemes.
Microdata	This section provides information about the IDSR source data, including the number of variables, their corresponding format, and a description of each variable. The raw microdata is not currently shared, but it has been ETled into the OMOP CDM and the results are accessible through ATLAS.

The detailed description of the IDSR study offers a deep understanding of the dataset, ensuring clarity for researchers. Accessible questionnaires further enrich the resource, providing invaluable insights into the data collection process.

The user interface is intuitively designed, offering seamless navigation and search capabilities. Researchers can effortlessly locate codebooks through keyword, title, or author searches. The availability of multiple download formats, including XML, PDF, and HTML, enhances accessibility.

A notable feature is the support for the repository extends its accessibility through multilingual metadata support, accommodating researchers from diverse linguistic backgrounds. This inclusivity further bolsters the codebook's utility and accessibility.

Additionally, the codebook maintains a detailed history of versions and updates made to the dataset. This feature ensures transparency and aids researchers in understanding potential impacts on their analyses.

JSON-LD

JSON-LD was used to create a structured representation of the IDSR data. This format leverages the schema.org vocabulary, a common language for describing things on the web. Specifically, we used properties from the MedicalObservationalStudy schema to capture essential information about the public health study, such as the study design (e.g., case series, cohort, observational, cross-sectional, longitudinal, or registry). Additionally, properties like schema.org/healthCondition, schema.org/studyLocation, schema.org/studySubject, and schema.org/guideline are used to describe patient population characteristics (e.g., inclusion criteria, age range), exposure or outcome measures of interest, and relevant health conditions or guidelines. Furthermore, this approach aligns with the efforts of OHDSI, which has adopted schema.org as a FAIRifying standard and extended its usage through OHDSI extension.

This structured representation benefits both human users and machine processing. For public health users familiar with schema.org, the data is easier to understand and interpret. Additionally, the use of a common vocabulary facilitates data exchange and integration with other systems that leverage schema.org. Figure 2 showcases a snippet of the JSON-LD code, illustrating how specific data elements are mapped to schema.org properties. Table 2 provides a more detailed tabular view of the IDSR JSON-LD class structure, including corresponding properties and their values.

.

Figure 2: Syntax Example: MedicalObservationalStudy

Table 2: JSON-LD class description

Class	Description
MedicalObservationalStudy	This represents the core IDSR study being described. It includes the Study Title, Identifier, Status, and Use Case.
MedicalRiskFactor	Describes both internal and external exposomes, i.e., the exposure of individuals in their environment.
DatabaseCharacteristics	This class describes the IDSR data set, which is implemented using the OMOP CDM v5.4.4. The IDSR data set is a federated system that includes data from Malawi, Kenya, and Uganda. The data set is stored in a relational database with the following tables: Person, Condition_Occurrence, Observation, Drug_Exposure, Procedure_Occurrence and Measurement
MedicalCondition	This class describes the medical conditions in the IDSR data, including their descriptions and OMOP CDM concept IDs.
MedicalObservationalStudyDesign	Provides detailed information on the study's design, including types of analysis (e.g., Predictive Analysis, Incident Rate Analysis).
UpdateAction	UpdateActions describe the workflow in the MedicalObservationStudy beginning with an action that takes the source data as input (object) and produces an OMOP CDM instance as output (result) using ETL SQL as an instrument. A second UpdateAction takes the OMOP CDM as an object and populates the OMOP CDM Results Schema as a result using both the MedicalObservationalStudyDesign and irs execution environment as instruments.

Our structured representation methodology makes the IDSR dataset easily discoverable through platforms like Google Dataset Search. The JSON-LD codes are available on GitHub Repository[40]. Additionally, the HTML code, embedded with JSON-LD, can be found on the INSPIRE website [41]. This open-source approach promotes transparency, reproducibility, and further development of our work.

Discussion

The integration of OMOP CDM, DDI, and Schema.org with JSON-LD provide access to the metadata within the IDSR framework. This led to substantial improvements in the FAIRness, standardization, interoperability, and analytical capabilities of the IDSR data, reinforcing the critical role of machine readability in this domain. Notably, the efficient sharing of vital metadata enables seamless data integration, collaborative research, and advanced analysis across diverse contexts, which is essential for improving public health outcomes. The model may also be used to promote the discoverability, accessibility, and reusability of observational research. One of the main benefits here is that IDSR data can become more visible and accessible on the Search Engine Results Pages (SERP), which can increase the click-through rate [40]. Moreover, sharing non-data research objects, such as analytical workflows and code, can significantly enrich the research ecosystem by enabling others to replicate, validate, and extend existing findings, fostering transparency and reproducibility.

However, the adoption of schema.org with JSON-LD in the context of population health data presents certain challenges. Specifically, the OHDSI vocabulary predominantly consists of medical terms, lacking comprehensive coverage of population health-related concepts such as HDSSs and IDSRs. Standard vocabularies like SNOMED-CT and LOINC often miss tests and questionnaires common in LMICs, particularly evident in capturing stages and public health-clinical interactions for diseases like AIDS. Terminologies like Columbia International eHealth Laboratory (CIEL), already integrated into OHDSI, offer significant potential as independent standard vocabularies to address these gaps [42].

To effectively address the specific gaps identified within the MedicalObservationalStudy model for African contexts, it is crucial to focus on enhancing vocabularies tailored to the region. This entails prioritizing risk factors and exposures currently inadequately captured by existing OHDSI standards. Particularly, OHDSI lacks vocabularies encompassing physical/chemical and social determinants of health, as well as mental health factors. Our participation in an OHDSI Working Group directly addresses these deficiencies by examining the relationships between diverse exposure histories (e.g., climate variations, pharmaceutical accessibility, healthcare availability for vulnerable populations like pregnant women) and the medical conditions presented by study subjects.

Importantly, technical expertise is necessary to optimize web pages for search engine optimization (SEO), ensuring effective implementation of schema.org with JSON-LD. This highlights the importance of community support in developing and refining the necessary resources and expertise to overcome these challenges and maximize the benefits of adopting schema.org with JSON-LD for population health data in Africa.

While the DDI codebook is a valuable tool for metadata documentation, it may not be as effective as the DDI lifecycle for promoting reuse under the FAIR principles, as noted by Chifundo Kanjala in "Open-access for existing LMIC demographic surveillance data using DDI" [35]. The DDI codebook is a static document that describes the data in a study at a single point in time, while the dynamic DDI Lifecycle can be used to describe the data throughout its lifecycle, from collection to dissemination. Moreover, the DDI Lifecycle is more machine-actionable, automating tasks such as data validation and interoperability. As a result, the adoption of the DDI lifecycle presents a promising avenue for future research, to further enhance the accessibility and reusability of population health data in LMIC contexts.

The study's findings underscore the substantial benefits of adopting OMOP CDM, DDI, and Schema.org with JSON-LD and suggest that these benefits outweigh the accompanying challenges. However, it is imperative to proactively address these challenges to ensure the successful implementation and adoption of these technologies. By actively tackling the challenges and offering robust support to users, the IDSR or HDSS community can significantly enhance the FAIRness and accessibility of their data and digital assets, enabling a broader spectrum of users to leverage its potential for research, decision-making, and public health interventions.

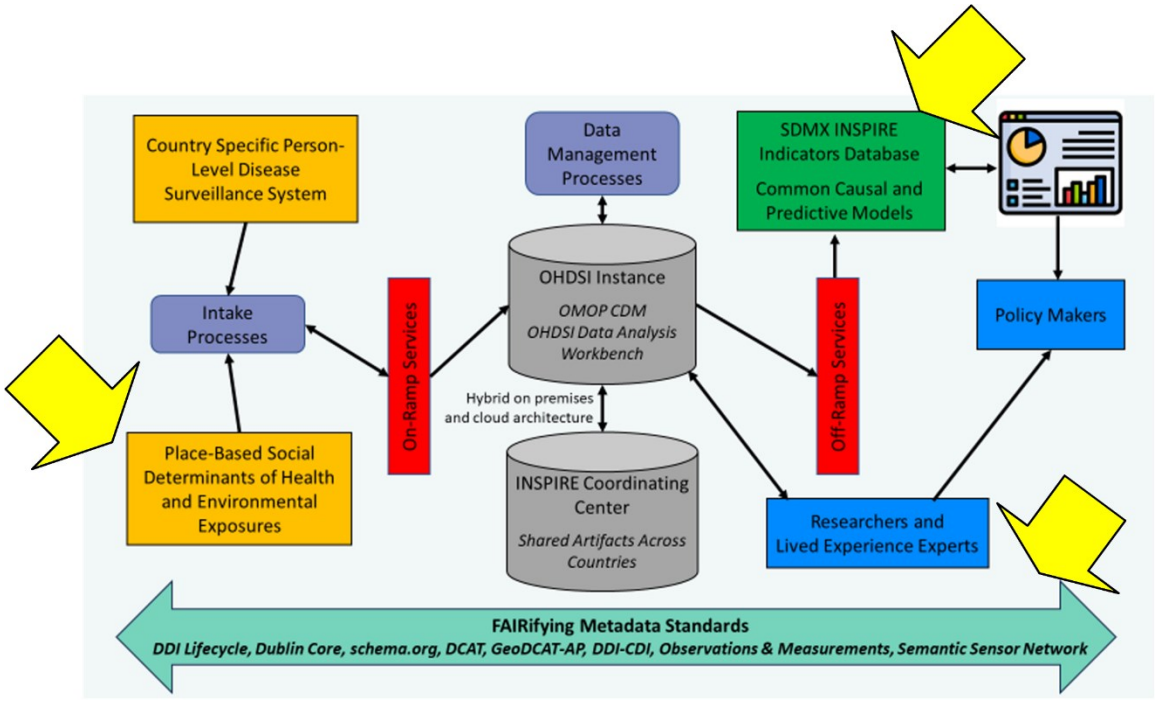
Future Direction

In future strategies, the FAIRification process will emphasize the adoption of schema.org and JSON-LD MedicalObservationStudy as a standard within the Cross-Domain Interoperability Framework (CDIF). This framework is capable of describing a workflow that includes the WHO IDSR platform, the OMOP CDM, and the OHDSI data analysis workbench. Whether each locality has its own platform, the data stays at home and only methods and aggregate results shared, or the data is pooled and moves through the entire workflow at once. Indeed, because the MedicalObservationalStudy can describe the entire arc of clinical and population health research, it more or less guarantees the reproducibility of studies.

In the future, our aim is to expand the capabilities of MedicalObservationalStudy in several ways. Firstly, we intend for it to describe a workflow that includes climate events and social determinants of health, crucial for understanding outcomes [43].]. Furthermore, we seek to enhance the MedicalObservationalStudy to describe a workflow that terminates in a data cube in which each cell disaggregates an aggregate result by many dimensions [44] . This would be a workflow that INSPIRE augments with a Statistical Data and Metadata eXchange (SDMX) instance to present the OMOP CDM Results Schema as an indicator repository [45]. This addition to the workflow mirrors the United Nations (UN) SDG platform and ensures broader utilization.

While SDMX is prominent, the Data Documentation Initiative-Comprehensive Data Integration (DDI-CDI) offers arguably a more capable one that statistical organizations may adopt in the future [44]. Also, Fast Healthcare Interoperability Resources (FHIR) integration with OHDSI which is ongoing may provide other paths that a future workflow takes. [46]. Through the navigation of these evolving platforms and their standards, our work with the MedicalObservationalStudy aims to achieve compatibility, relevance, and interoperability within the public health community. The arrows in the figure below show where machine-readable and machine-actionable metadata are still needed:

Fig 3: Next Steps



Finally, as new, and emerging data formats emerge, schema.org's flexibility and extensibility will play a crucial role in accommodating these formats, ensuring continued compatibility and interoperability within the evolving landscape of population health data.

Conclusions

The utilization of machine-readable metadata plays a vital role in ensuring FAIRification of population health data. By embracing universal standards, such as schema.org, organizations can not only enhance their search engine optimization but also make their data more discoverable on the web. This will maximize the impact and utility of population health data, particularly in LMICs. This paper highlights the importance of promoting and adopting machine-readable metadata standards in LMICs to advance the FAIRification and accessibility of population health data.

Ethical Approval

This research centers on implementing the FAIR (Findable, Accessible, Interoperable, and Reusable) principles. There is no involvement of human subjects in our research, ethical approval from a local ethics committee is not applicable.

Data Availability

No new data were generated or analysed in support of this research.

Authors Contribution

DA designed the study and authored the original manuscript. JG oversaw the study's implementation. JT, JG, SK-M, AT and TB played key roles in conceptualization and methodology. JT, KT, JG, TB, AK, AG, MO, CK and SK-M reviewed and edited the manuscript. All authors contributed to the article and approved the submitted version.

Funding

INSPIRE has received funding from Wellcome (grants 224834_Z_21_Z and 226589_Z_22_Z) and IDRC Canada (grant 109622-001) which has contributed to this work.

Conflict of Interest

We declare that we have no competing interests concerning this study.

Acknowledgment

The authors would like to extend their gratitude to the MUBAS team in Malawi, the APHRC data science Nairobi team, the CODATA team, and the INSPIRE project colleagues for their invaluable contributions to the project.

References

[1] R. C. Brownson, J. F. Chriqui, and K. A. Stamatakis, "Understanding Evidence-Based Public Health Policy," *Am J Public Health*, vol. 99, no. 9, pp. 1576–1583, Sep. 2009, doi: 10.2105/AJPH.2008.156224.

[2] I. S. Fall *et al.*, "Integrated Disease Surveillance and Response (IDSR) strategy: current status, challenges and perspectives for the future in Africa," *BMJ Glob Health*, vol. 4, no. 4, p. e001427, Jul. 2019, doi: 10.1136/bmjgh-2019-001427.

[3] O. Sankoh and P. Byass, "The INDEPTH Network: filling vital gaps in global epidemiology," *Int J Epidemiol*, vol. 41, no. 3, pp. 579–588, Jun. 2012, doi: 10.1093/ije/dys081.

[4] W. G. van Panhuis *et al.*, "A systematic review of barriers to data sharing in public health," *BMC Public Health*, vol. 14, no. 1, p. 1144, Nov. 2014, doi: 10.1186/1471-2458-14-1144.

[5] J. B. Pettengill, J. Beal, M. Balkey, M. Allard, H. Rand, and R. Timme, "Interpretative Labor and the Bane of Nonstandardized Metadata in Public Health Surveillance and Food Safety," *Clinical Infectious Diseases*, vol. 73, no. 8, pp. 1537–1539, Oct. 2021, doi: 10.1093/cid/ciab615.

[6] M. D. Wilkinson *et al.*, "The FAIR Guiding Principles for scientific data management and stewardship," *Sci Data*, vol. 3, no. 1, p. 160018, Mar. 2016, doi: 10.1038/sdata.2016.18.
[7] "Data science," *Wikipedia*. Sep. 13, 2023. Accessed: Sep. 18, 2023. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Data_science&oldid=1175162291
[8] J. Riley, "UNDERSTANDING METADATA".
[9] A. Devaraju and R. Huber, "An automated solution for measuring the progress toward FAIR research data," *Patterns (N Y)*, vol. 2, no. 11, p. 100370, Oct. 2021, doi: 10.1016/j.patter.2021.100370.
[10] A. E. Green and C. Humphrey, "Building the DDI," *IQ*, vol. 37, no. 1–4, p. 36, May 2014, doi: 10.29173/iq500.
[11] "iSHARE Repository." Accessed: Aug. 30, 2023. [Online]. Available: https://www.indepth-ishare.org/index.php/home
[12] "Central Data Catalog." Accessed: Aug. 30, 2023. [Online]. Available: https://saprindata.samrc.ac.za/index.php/catalog
[13] "Microdata Portal," APHRC. Accessed: Aug. 30, 2023. [Online]. Available: https://aphrc.org/microdata-portal/
[14] O. H. D. S. and Informatics, *Chapter 4 The Common Data Model | The Book of OHDSI*. Accessed: Apr. 24, 2024. [Online]. Available: https://ohdsi.github.io/TheBookOfOhdsi/
[15] J. M. Overhage, P. B. Ryan, C. G. Reich, A. G. Hartzema, and P. E. Stang, "Validation of a common data model for active safety surveillance research," *Journal of the American Medical Informatics Association*, vol. 19, no. 1, pp. 54–60, Jan. 2012, doi: 10.1136/amiajnl-2011-000376.
[16] M. Wornow, "Walkthrough of the OMOP CDM (Part 1)." Accessed: Apr. 24, 2024. [Online]. Available: https://michaelwornow.net/2022/12/30/omop-cdm-part-1
[17] O. H. D. S. and Informatics, *Chapter 5 Standardized Vocabularies | The Book of OHDSI*. Accessed: Apr. 01, 2024. [Online]. Available: https://ohdsi.github.io/TheBookOfOhdsi/
[18] C. M. Hallinan *et al.*, "Seamless EMR data access: Integrated governance, digital health and the OMOP-CDM," *BMJ Health Care Inform*, vol. 31, no. 1, p. e100953, Feb. 2024, doi: 10.1136/bmjhci-2023-100953.
[19] "Schema.org - Schemas - Schema.org." Accessed: Feb. 24, 2023. [Online]. Available: https://schema.org/docs/schemas.html
[20] "JSON-LD - JSON for Linking Data." Accessed: Sep. 18, 2023. [Online]. Available: https://json-ld.org/
[21] C. U. Guerrero, M. V. Romero, M. Dolman, and M. Dumontier, "FAIR Begins at home: Implementing FAIR via the Community Data Driven Insights." arXiv, Mar. 13, 2023. Accessed: Aug. 18, 2023. [Online]. Available: http://arxiv.org/abs/2303.07429
[22] "FAIR Principles," GO FAIR. Accessed: May 23, 2023. [Online]. Available: https://www.go-fair.org/fair-principles/
[23] "WorldFAIR," CODATA, The Committee on Data for Science and Technology. Accessed: Sep. 03, 2023. [Online]. Available: https://codata.org/initiatives/decadal-programme2/worldfair/
[24] M. Boeckhout, G. A. Zielhuis, and A. L. Bredenoord, "The FAIR guiding principles for data stewardship: fair enough?," *Eur J Hum Genet*, vol. 26, no. 7, pp. 931–936, Jul. 2018, doi: 10.1038/s41431-018-0160-0.
[25] G. Peng, "The State of Assessing Data Stewardship Maturity – An Overview," vol. 17, no. 0, Art. no. 0, Mar. 2018, doi: 10.5334/dsj-2018-007.
[26] S. J. Nass, L. A. Levit, L. O. Gostin, and I. of M. (US) C. on H. R. and the P. of H. I. T. H. P. Rule, "The Value, Importance, and Oversight of Health Research," in *Beyond the HIPAA Privacy Rule: Enhancing Privacy, Improving Health Through Research*, National Academies Press (US), 2009. Accessed: Nov. 24, 2023. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK9571/
[27] G. Royston, N. Pakenham-Walsh, and C. Zielinski, "Universal access to essential health information: accelerating progress towards universal health coverage and other SDG health targets," *BMJ Glob Health*, vol. 5, no. 5, p. e002475, May 2020, doi: 10.1136/bmjgh-2020-002475.
[28] D. Welter *et al.*, "FAIR in action - a flexible framework to guide FAIRification," *Sci Data*, vol. 10, no. 1, p. 291, May 2023, doi: 10.1038/s41597-023-02167-2.

[29] "NADA | Microdata Cataloging Tool." Accessed: Jun. 23, 2023. [Online]. Available: https://nada.ihsn.org/
[30] "Overview | NADA Documentation." Accessed: Sep. 18, 2023. [Online]. Available: https://ihsn.github.io/nada-documentation/intro/#why-nada
[31] "DDI Metadata Editor (Nesstar Publisher 4.0.10) | IHSN." Accessed: Sep. 18, 2023. [Online]. Available: http://www.ihsn.org/software/ddi-metadata-editor
[32] "Quick Reference Guide for Data Archivists — Guide for Data Archivists documentation." Accessed: Sep. 27, 2023. [Online]. Available: https://guide-for-data-archivists.readthedocs.io/en/latest/
[33] "DDI-Codebook 2.5 | Data Documentation Initiative." Accessed: Aug. 24, 2023. [Online]. Available: https://ddialliance.org/Specification/DDI-Codebook/2.5/
[34] I. Barkow *et al.*, "Generic Longitudinal Business Process Model," 2013, doi: 10.3886/DDILONGITUDINAL05.
[35] C. Kanjala *et al.*, "Open-access for existing LMIC demographic surveillance data using DDI," *IQ*, vol. 40, no. 2, p. 18, Feb. 2017, doi: 10.29173/iq783.
[36] "Frontiers | INSPIRE datahub: a pan-African integrated suite of services for harmonising longitudinal population health data using OHDSI tools." Accessed: May 06, 2024. [Online]. Available: https://www.frontiersin.org/journals/digital-health/articles/10.3389/fdgth.2024.1329630/full
[37] S. Kiwuwa-Muyingo, J. Todd, T. Bhattacharjee, A. Taylor, and J. Greenfield, "Enabling data sharing and utilization for African population health data using OHDSI tools with an OMOP-common data model," *Frontiers in Public Health*, vol. 11, 2023, Accessed: Jun. 28, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpubh.2023.1116682
[38] EHDEN, "FAIRification of observational studies and databases: the EHDEN and OHDSI use case." Accessed: Feb. 24, 2023. [Online]. Available: https://fairplus.github.io/the-fair-cookbook/content/recipes/applied-examples/ehden-ohdsi.html
[39] Google, "Dataset Search." Accessed: Feb. 26, 2023. [Online]. Available: https://datasetsearch.research.google.com/
[40] jaygee-on-github, "INSPIRE-Domain-Model-for-Network-Cohort-Studies." Jul. 08, 2023. Accessed: Jul. 21, 2023. [Online]. Available: https://github.com/jaygee-on-github/INSPIRE-Domain-Model-for-Network-Cohort-Studies
[41] "Inspire Data Network." Accessed: May 06, 2024. [Online]. Available: https://inspiredata.network/ext/fairomop
[42] "OCL." Accessed: May 06, 2024. [Online]. Available: https://app.openconceptlab.org/#/orgs/CIEL/
[43] "OHDSI GIS Workgroup." Observational Health Data Sciences and Informatics, Mar. 17, 2023. Accessed: Jul. 21, 2023. [Online]. Available: https://github.com/OHDSI/GIS
[44] "DDI-Cross Domain Integration (DDI-CDI)," CODATA, The Committee on Data for Science and Technology. Accessed: May 12, 2023. [Online]. Available: https://codata.org/initiatives/decadal-programme2/ddi-cross-domain-integration/
[45] M. Kellerman and E. David, "Implementation of the SDMX International Standard in Statistical Data Specification," 2021.
[46] "HL7 International and OHDSI Announce Collaboration to Provide Single Common Data Model for Sharing Information in Clinical Care and Observational Research – OHDSI." Accessed: Apr. 25, 2024. [Online]. Available: https://www.ohdsi.org/ohdsi-hl7-collaboration/

Abbreviations

FAIR	Findable, Accessible, Interoperable, Reusable
DDI	Data Documentation Initiative
ETL	Extraction, Transform, Load

INSPIRE	Implementation Network for Sharing Population Information with Research Entities
HDSS	Health and Demographic Surveillance System
OMOP	Observational Medical Outcomes Partnership
OHDSI	Observational Health Data
CDM	Common Data Model
SDMX	Statistical Data and Metadata Exchange
IDSR	Infectious Disease Surveillance and Response
JSON-LD	Javascript Object Notation – Linked Data
EHDEN	European Health Data Evidence Network
SDG	Sustainable Development Goals
WHO	World Health Organisation
NADA	National Data Archive
LMIC	Low and Middle-Income Countries
APHRC	African Population Health Research Centre
LSHTM	London School of Hygiene and Tropical Medicine
CODATA	Committee on Data (International Science Council)