Contribution ID: **94** Type: **Poster**

Context and Provenance for FAIR Health Data - The who, what, why, where, when and how

Monday 13 October 2025 18:00 (1h 30m)

The COVID-19 pandemic has altered how health data is regarded and was a distinct driver for change. The need for rapid analysis and assessment of health data at scale brought sharp focus to the challenges, highlighting the importance of Findable, Accessible, Interoperable and Reuseable (FAIR) data. The heterogeneous nature of health data, together with the wide array of systems and associated formats that record and represent data is a fundamental impediment to this. This is compounded by the representation of data, i.e., how it is coded and thus understood. This aspect alone varies widely, from localised efforts to encode data through to the use and application of international standardised controlled vocabularies (e.g. SNOMED).

One approach used to address these issues is the application of Common Data Models (CDM). A CDM that is gaining traction internationally is the Observation Health Data Sciences and Informatics (OHDSI) Outcomes Medical Observation Partnership (OMOP) CDM. Its aim is to harmonise the representation of heterogeneous health datasets to enable the FAIR assessment and analysis of converted datasets. The approach to creating OMOP CDM datasets utilises a process called Extract, Transform and Load (ETL). The Transform process converts source data representation to a target representation using OHDSI's own OMOP controlled vocabulary concepts by way of the OMOP CDM. Currently, how the Transform process is performed is ambiguous because the decisions necessary to select the target vocabulary terms to represent source data terms are not explicit. Thus supporting knowledge cannot be used to aid the Transform process. In effect, the whole approach to transforming data is variable, open to interpretation and thus subjective.

To be able to make better informed decisions when transforming data, a potential answer lies in the ability to represent and use context and provenance. In other disciplines, data provenance is used to record and represent the origin of the data, how it has been moved, processed and who has performed these. Yet, for OMOP datasets such approaches have yet to be made available. The more amorphous properties of context are somewhat problematic to be able to represent, but they are in part key to understanding why decisions are made and the factors that influenced them.

In November 2024, a workshop was run to focus upon 'How to be FAIR with data standards'. Its focus was the conversion of datasets to the OMOP CDM, such conversions can be lossy –in that data granularity may not be translated fully. 55 people from around the United Kingdom and Europe took part in the workshop. These constituted academics, clinicians and industrial representatives. Activities were designed to elicit perspectives and thoughts from attendees by posing questions intended to derive a response. Two themes were set out and groups were assigned accordingly. The main question posed was 'All OMOP datasets are unreliable'. Currently there is no way to prove whether or not they are reliable.

The workshop produced a rich set of perspectives focused upon current transformation challenges. Utilising these, a context and provenance data model has been developed. It represents the pertinent attributes that potentially affect decision making and the provenance of the data conversion. The data model is composed of several concepts, they are: OMOP Datasets, Quality, Standards, Sharing, Circumstances, Provenance, Version, Transformation, Decisions, Context, Datasets, People, Bias and Training.

The central concept of the model is OMOP Rulesets, i.e., the rules that have been created to transform source data to the OMOP CDM. The concepts of Quality and Standards support this, Quality representing the quality of the rulesets and Standards, the associated standards. Sharing represents FAIR data approaches to sharing rulesets.

The next key concept is Circumstances which relates to OMOP Rulesets. It models the who, why, what, where, when and how (bringing together the concepts of Provenance, Decisions and Context). It coalesces

the factors that influence data transformation. Provenance relates to the OMOP Ruleset concept, and Circumstances. It represents the factors of data origin, data transportation, changes made to the data, when and by whom. This is supported by Version to evidence changes made. The concept of Decisions models reasons why decisions were made and the transformation aim. Context concerns environmental aspects, where was the transformation performed, under what conditions and time pressures. Transformation represents the processes undertaken to convert the data to the OMOP CDM, this in turn relates to the concept of Datasets (the source data). Additionally, the concepts of People, Training and Bias represent the people that perform the decisions, their expertise and training, plus any potential bias.

The context and provenance data model seeks to make explicit the factors surrounding CDM data transformations. The conversion of data to the OMOP CDM is a time consuming and complex undertaking. Hence, there is a need to understand how better to transform data in an ever-efficient manner by providing methods and systems to support this. Additionally, the growth in federated data approaches underlines the need to create reliable OMOP datasets, to which the data model can provide provenance and context information for such transformations.

Further work will entail (i) running the workshop again at the Elixir All-hands meeting in June 2025 and to gain additional insight; (ii) development of a software tool to instantiate the data model to capture the attributes of data provenance and context for end users. The objectives of these are to further support OMOP CDM transformations and improve their reliability.

Primary author: URWIN, Esmond (University of Nottingham)

Co-authors: Mr RAE, Andy (University of Nottingham); Dr BECK, Tim (University of Nottingham); Dr FIGUEREDO, Grazziela (University of Nottingham); Prof. QUINLAN, Phil (University of Nottingham)

Presenter: URWIN, Esmond (University of Nottingham)

Session Classification: Poster Session

Track Classification: SciDataCon2025 Specific Themes: Open research through Interconnected, Interoperable, and Interdisciplinary Data