Connected Bradford - FDM Document

Last modified: 2022/11/29 - JDB

What is a FDM ?

How is it built ?

FAQ’s

# 1. What is a Flexible Data Model (FDM)?

A Flexible Data Model (FDM) is a standard way of structuring datasets. It has been devised by the Connected Bradford team as an adaptation of the ‘Common Data Model’ (CDM), which was created by the Observational Medical Outcomes Partnership (OMOP). The CDM transforms data from disparate databases into a common format (‘data model’), allowing for easy linkage and systematic analysis (you can read a brief overview of the CDM [here](https://www.ohdsi.org/data-standardization/the-common-data-model/)).

Connected Bradford aspires to adopt the CDM where possible, e.g. for patient-centred healthcare data. However, some datasets in Connected Bradford (e.g. education data, intensive care visits, etc.) have features that do not sit well within either the OMOP conceptual vocabulary or the structure of the CDM. The FDM therefore provides a data model for non-OMOP datasets so that they can be structured in an OMOP-friendly way, and are therefore easy to combine with OMOP CDM data during analysis.

## What tables are required?

For our current purposes, the minimum tables contained within the FDM are::

* **Person Table -** Shows the people contained in the current dataset. It is a subset of the master person table containing only people relevant to this FDM. It is created as part of the build of this FDM.
* **Observation Period Table -** This shows the earliest (visit\_occurrence\_start\_date) and latest (visit\_occurrence\_end\_date) related to each person in the current dataset.
* **Visit Occurrence Table -** Shows the events/encounters/visits for each person in the current dataset. Minimum required fields are visit\_occurrence\_id, person\_id, visit\_start\_date/visit\_start\_datetime.

# 2. General FDM Creation Process

The exact process of creating an FDM will differ for every dataset, but general steps are below. Ensure that your scripts contain descriptive comments that can be understood by others.

1. Look over the columns to see if there are any sensible ways that the data could be made longer (i.e. condensing multiple columns into one column with several rows).
2. If it’s manageable, cast key columns into relevant types, e.g. dates, integers, floats.
3. Discuss the viability of any further cleaning with Kuldeep and John.
4. Run Sam’s FDM Builder Script (linked below) to create the person table and the observation period table.
5. Create the data dictionary - start with the table that Sam’s script creates. Download this as a CSV, open in Excel / copy into Google Sheets. Describe the columns to the best of your ability.
6. Create the visit occurrence table - this will likely differ according to each dataset, but teammates can share the process they have taken and discuss options for your data.
7. Create a dashboard with the standard plots based on the person and visit occurrence tables (template linked below), then 2-3 plots unique to your dataset to show its potential. These plots don’t have to answer legitimate literature-informed research questions or show something profound, they are to illustrate a handful of variables that are available in your dataset.
8. Create a folder in the Google Drive named ‘CY\_FDM\_<YourDataset>’. Copy and paste your scripts into Google Docs and document your process (see current folders for examples).
9. Make sure your finished tables are within a dataset that John has access to.
10. Celebrate.

# 3. Frequently Asked Questions

* What is GCP? Google Cloud Platform is the environment in which we store and query the Connected Bradford data.
* What is BigQuery? BigQuery is the data warehouse within GCP. The language used is Google BigQuery Standard SQL.
* What is Vertex ? Vertex AI is a secure virtual development platform , also called a Jupyter Notebook, used to build your Python, R scripts etc to access the data.
* What are projects, datasets and tables in BigQuery? We all work within the same project, ‘yhcr-prd-phm-bia-core’. Datasets () are folder-like containers for tables (). Tables contain your data, made up of different columns (fields) and rows (records). Some datasets may comprise of several *source tables*. A *MYSPACE* dataset will be made for you, but you can create tables freely.
* What is a MYSPACE? A MYSPACE is an empty dataset created for each user. It is a space for you to freely create and delete tables, without affecting the wider datasets, it is your personal development space and can be used while you train and familiarise yourself with the environment. Not all users required a MYSPACE
* What does ‘source’ mean, e.g. ‘source tables’? Source tables are tables of data that are received from a particular data provider, relating to a particular dataset. There may be only one source table within a dataset (e.g. ICNARC [intensive care]), or there may be multiple tables within a dataset (e.g. education).
* What is the Master Person table ? - The Connected Bradford dataset contains a Master FDM (CY\_FDM\_MASTER) . All users have read access to this dataset. This does not contain any clinical data but does contain person, care\_site and observation\_period tables. The persons table contains a unique record for each individual that we have data on, from whichever dataset, each person being assigned a unique person\_id that is common throughout the data.. This means that in some cases we may not have anything more than they exist. In others we will know dates of birth, death, ethnicity etc.The Care site is based on the Primary Care data and will, where possible, refer to the patients partial postcode. The observation period is also based on the full data held. Is from first record to latest data held date (or date of death + 42 days).
* What are lookups ? - The dataset CY\_LOOKUPS contains a number of useful datasets that can potentially be useful for analysis of a specific FDM. All users have access to this dataset.Example tablase include tbl\_CTV3ToSnomed\_Map, this is a reference table showing the CTV3 read codes and the relevant SNOMED Code. The person to lsoa table shows each individual person\_id and the lsoa for this individual, thai may be useful for geographic analysis.
* [FDM Builder Script](https://console.cloud.google.com/vertex-ai/workbench/list/instances?project=yhcr-prd-phm-bia-core) - A Jupyter lab workspace is created for each person by Sam Relins. Contains ‘fdm\_builder\_basics\_tutorial.ipynb’ and ‘fdm\_builder\_extras\_tutorial.ipynb’. Adapt the basics script and ask Sam any questions. This doesn’t build a complete FDM but creates the person and observation period tables along with a rudimentary data dictionary. (Log in with your YHCR Account).
* [FDM Dashboard Template](https://datastudio.google.com/reporting/425e4ef7-6936-4bb4-a6bd-a0b230de08b7/page/8iurC) - Sam/Rob have created a basic template for visualising the standard parts of the datasets via a dashboard (i.e. based on the person and observation period). (Log in with your YHCR Account).
* [FDM Specification](https://docs.google.com/document/d/149pC6lvOyO-cYQN6fioU0Q-gSlfH4D0-JuBQkxDmLxY/edit?usp=sharing) **-** Details the specification for each FDM.
* [Connected Bradford Shared Google Drive](https://drive.google.com/drive/u/2/folders/0AMHinCA0NOcgUk9PVA) - A space for sharing the scripts used to create FDMs and accompanying documentation. (Log in with your NIHR Account).
* [White Rabbit](http://ohdsi.github.io/WhiteRabbit/WhiteRabbit.html) - a software tool that can be useful to triage a new dataset. It returns the following for each column: type (tells you the relevant type even if the data is all saved as strings), maximum length, number of rows, fraction empty (null), number of unique values, fraction unique. Though it says it can read straight from BigQuery, we are unable to follow the authentication process. Instead, it can be used on datasets in SQL Server. John has an executable file. (check)