CO-CONNECT

Metadata Profile Registration Process for HDR

Innovation Gateway

October 2021

This is a guide for the preparation of dataset metadata for the CO-CONNECT project to enable registration on the HDRUK Gateway, referred to in Step 2 CO-CONNECT Implementation Document. It presents a step-by-step guide to each process and explains the order in which these processes happen.

Figure 1 is a flowchart showing the steps needed to compose the metadata profile on the HDR Innovation Gateway. Each step is described and sets out responsibilities for both the Data Partner and the CO-CONNECT Team.

For security reasons, the CO-CONNECT Team will not have access to any of the Data Partner's data at any time, instead, each Data Partner will share a standardised set of metadata files with the CO-CONNECT Team who will use these files to define bespoke transformation and mapping rules.
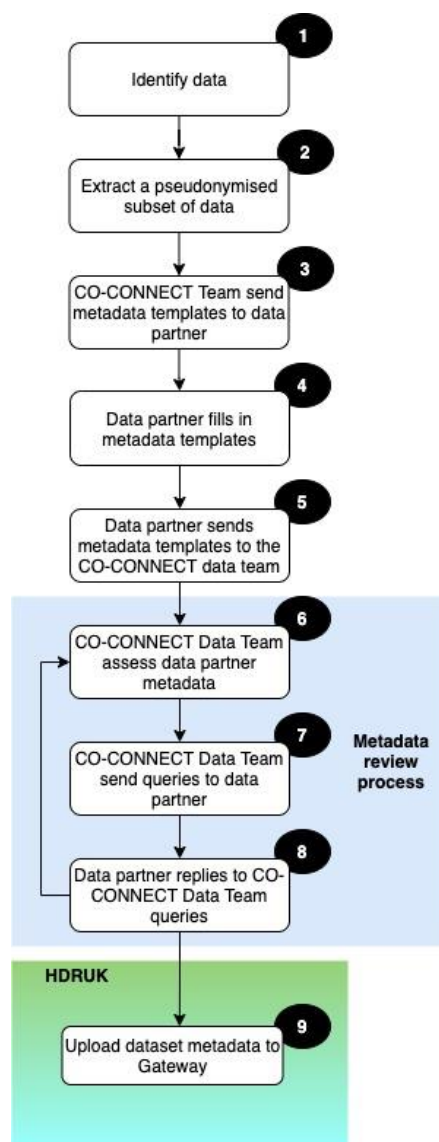

Figure 1: Metadata development steps.

The metadata profile registration process consists of nine distinct steps, these are described as follows:

Step 1: Identify data

The Data Partner will select the most relevant set or subset of data to be included in the CO-CONNECT repository. The CO-CONNECT Team can help advise each Data Partner which datasets are most relevant to include, but the final decision regarding data for upload remains with the Data Partner. To identify the same patients across other overlapping CO-CONNECT datasets, the Data Partner will map source patient identifiers to NHS/CHI Number (if not already using NHS/CHI Number). This will allow the CO-CONNECT Team to understand the overlap of data between different studies, but data will only be linked with explicit permission of all data controllers of each data set.

Step 2: Extract pseudonymised subset of data

Once the data set(s) have been identified, the Data Partner will then anonymise the NHS/CHI Number using a one-way hashing algorithm, with a pre-defined Salt, which will be securely shared with all Data Partners (please refer to the Data Anonymisation document here). From this pseudonymised data, the Data Partner will extract a representative subset of that data.

Step 3: CO-CONNECT Team send metadata templates to Data Partner

The CO-CONNECT Team will send the Data Partner two metadata templates to complete, these are:

1. Summary Metadata Template

The Summary Metadata Template is used to describe the cohort, providing a summary of the data (including title, abstract, key contacts, etc.), it's coverage (geographical, age range, etc.), its provenance, any accessibility requirements, linkages and metadata overview. This information is needed to understand what a dataset is and what it contains at a high level.

- HDRUK Summary Metadata Template (*click the link*)

2. Structural Metadata Template

The Structural Metadata Template addressed the metadata structure of the dataset. This template gathers table names and descriptions, column names and descriptions, their associated data types and whether the data is sensitive. It is essential that the Data Partner descriptions support the CO-CONNECT mapping process.

- HDRUK Structural Metadata Template (*click the link*)

Both templates create a descriptive entry for the dataset within the HDR Gateway.

Step 4: Data Partner fills in metadata templates

On receipt, the Data Partner should complete the Structural and Summary Metadata Templates. Figure 2 and 3 below, provide examples of how these forms should be completed. The CO-CONNECT Team can support Data Partner with this process.

Figure 2: Example of Summary Metadata Template contents

Table Name	Table Description	Column Name	Column Description	Data Type	Sensitive
			(data dictionary description)	*(INT/REAL/VARCHAR/etc.)*	*FALSE/TRUE*
COVID Table 01	COVID Study X	Patient ID	A person's identifier	VARCHAR	TRUE
COVID Table 01	COVID Study X	Age	Age of patient	INT	TRUE
COVID Table 01	COVID Study X	Visit Number	Number of patient visit	INT	FALSE
COVID Table 01	COVID Study X	Dry Cough	dry cough symptom	VARCHAR	FALSE
COVID Table 01	COVID Study X	Wet Cough	wet cough symptom	VARCHAR	FALSE
COVID Table 02	COVID Study X	Symptom1	persistent cough	VARCHAR	FALSE
COVID Table 02	COVID Study X	Symptom2	sore throat	VARCHAR	FALSE
COVID Table 02	COVID Study X	Symptom3	headache	VARCHAR	FALSE

Figure 3: Example of Structural Metadata Template contents

Step 5: Data Partner sends Summary and Structural Metadata Templates to the CO-CONNECT Team

Complete Summary and Structural Metadata Templates should be sent directly to the CO-CONNECT Team using the following email address:

- CO-CONNECT@dundee.ac.uk

The CO-CONNECT Team will acknowledge receipt by return email.

Step 6: CO-CONNECT Team assess Data Partner metadata

The completed Summary and Structural Metadata Templates will be assessed by the CO-CONNECT Team, who will check the structure of the metadata and will review any missing descriptive information or metadata items.

Step 7: CO-CONNECT Team send queries to Data Partner

Once the supplied metadata is assessed, the CO-CONNECT Team will contact the Data Partner with a set of queries.

Depending on the information supplied, further information metadata information and associated descriptions may be needed.

Step 8: Data Partner replies to CO-CONNECT Team queries

Further information requested by the CO-CONNECT Team should be sent by the Data Partner.

Step 9: Uploading dataset metadata to HDR UK Gateway

Metadata is needed to assist researchers and public health groups to discover the data. It can also support governance applications needed for further data analysis or research. An example metadata entry representing the Oxford Healthcare Workers data set in the HDR UK Gateway is provided here:

- Oxford Healthcare Workers *(within the HDRUK Gateway)*

Data Partner checks / creates HDR UK account

Before commencing, a Data Partner must check whether or not their institution has an existing account within the HDR UK gateway. If one does not exist, the Data Partner must create one for their institution.

HDR UK account members

To onboard and manage metadata on the HDR UK Gateway, Data Partners must first be a member of a team. To create a team, email: support@healthdatagateway.org,

or raise a support ticket at https://hdruk.atlassian.net/servicedesk/customer/portal/1.

When logged on, within the members tab section, Data Partners can add new members to the team, assuming individuals have an account on the HDR UK Gateway.

Metadata Upload Options

There are two options to upload the Summary and Structural Metadata to the HDR UK Gateway (created in 'Step 4' of this guide), these are:

Option 1: The Data Partner uploads dataset metadata to the HDR UK Gateway

The Data Partner uploads the content of the Summary and Structural Metadata Templates to the HDR UK gateway. Guidance for this process can be found here:

- How to add a dataset to the HDRUK Gateway

Option 2: The CO-CONNECT Team uploads dataset metadata to the HDR UK Gateway

The Data Partner must advise the CO-CONNECT Team in writing (by email) to give explicit consent for the CO-CONNECT Team to upload metadata to their institution's HDR UK gateway account.

HDR UK Gateway dataset tagging

Each CO-CONNECT associated dataset that is created and uploaded into the HDR UK gateway must be tagged with the following compulsory tag set by the per:

Compulsory tag set
- COVID
- COVID-19
- SARS-CoV-2
- Coronavirus
- CO-CONNECT
- Serology

Additionally, there are a suggested list of optimal tags that can be added to a dataset entry:

Optional tag list
- Data Partner Name
- Dataset Name
- Health Care Workers
- Healthcare Workers
- HCW
- Antibody
- Antibodies
- Seroprevalence
- Prevalence
- National Core Studies
- NCS

HDR UK Gateway Walkthroughs and Help

For additional support, links are provided below to help Data Partners understand the HDR UK Gateway and associated metadata uploading processes:
- HDR UK Data Custodian Webpage: How-To and Help
- HDR UK Data Custodian Playbook.pdf (download)
- HDR UK How to videos and guides
- HDR UK FAQs