

APRIL 4, 2023

ADDI DATA PREPARATION GUIDE

ADDI TECH TEAM

Table of Contents

Data Preparation	2
Introduction and Motivation	2
Data Preparation Checklist	3
Catalogue	3
Dictionaries	4
Fields	6
Lookups	7
Dataset Creation and Upload	8
Glossary	11

Data Preparation

Introduction and Motivation

Data curation and preparation is a critical step in making data usable within a broad research community. Taking additional steps to prepare data and accompanying documentation so that external users can decipher variables and the information contained in them ensures data are used properly and efficiently and avoids errors in processing or analysis. ADDI is committed to working with data contributors so that their data is represented in the best way possible, setting up the data consumers for success.

The ADDI Data Preparation Guide is intended to assist with gathering the required documentation for onboarding data to the ADDI portal. The required documents are each outlined below in detail. This document should be used in conjunction with the 'FAIR Metadata Template' spreadsheet document. A brief description of the tabs in spreadsheet is given below:

1. **settings:** This tab is used to define basic settings related to dataset approval workflow, visibility, and cohort settings. Other than “visibility” and “workflow_key” all other settings can be left untouched as default values. Visibility can include one of the two values – **private** or **internal**. Private sets the visibility as private which means the dataset is not visibly to anyone apart from the owner and the people dataset is shared with. Internal means dataset is visible to all FAIR users. Settings can be modified from the FAIR website once the dataset has been created
2. **catalogue:** The catalogue collects high-level information about the dataset such as “title”, “description”, and “keywords” etc.
3. **dictionaries:** This tab contains information about data tables in the. “code” is used to identify the tables in the dataset, it can only contain letters, numbers, and underscores. The “code” defined here is referenced in the “fields” tab where we define the variables in the table.
4. **fields:** The variables in the tables need to be defined here. One row per variable per data table. The “dictionary_code” should be one of the “code” defined in the dictionaries tab.
5. **lookups:** Lookups are used to define the code-book (if any) for the variables.

We refer to a dataset as a distinct study that has a unique entry in the ADDI portal. By logging in to the [ADDI portal](#) and navigating to the FAIR search page, you can familiarize yourself with how these

fields will be displayed to the user for your dataset. The following sections describe each of these in detail.

Data Preparation Checklist

The following items are required by ADDI to fully ingest a dataset to the ADDI portal:

1. **catalogue:** one entry per study/dataset
2. **dictionaries:** name, label and description for each data table included in the study/dataset
3. **fields:** name, label, description and type for each variable within each data table included in the study/dataset; for variables that are coded or constrained, accompanying variable lookup tables must also be supplied

Catalogue

Figure 1 below identifies the fields that will be populated for the dataset. Please use the 'CatalogInformation' tab in the 'Data Prep Templates' spreadsheet to populate these fields for your study. The following information will be displayed with the dataset on the FAIR Search page:

1. **Title:** a short name to identify the dataset; displayed at the top of the dataset page in bold
2. **Description:** a paragraph description of the study, including details such as study design, number of participants, recruitment site(s) and strategy, outcomes measured, and any links to further study resources or prior publications
3. **Creator:** the author, PI of the study, or creator of the dataset; some examples are 'Centers for Disease Control and Prevention' or 'Jane Smith'
4. **License (optional):** license for the dataset, such as the creative commons license described here: <http://creativecommons.org/licenses/by-sa/3.0/> (field not shown in Figure 1)
5. **Version (optional):** the dataset version, such as 1.0
6. **Keywords:** a list of keywords to identify the study such as AD, Alzheimer's, MMSE questionnaire, family history
7. **Identifier (optional):** a digital object identifier (DOI) for the dataset
8. **Access rights (optional):** access rights for the dataset, such as 'public' (field not shown in Figure 1)
9. **Publisher:** the organization or person publishing the data; could be the same as 'creator name'

10. **Publisher Url (optional):** website URL for the dataset or creation organization (field not shown in Figure 1)

11. **language (optional):** language for the dataset (field not shown in Figure 1)

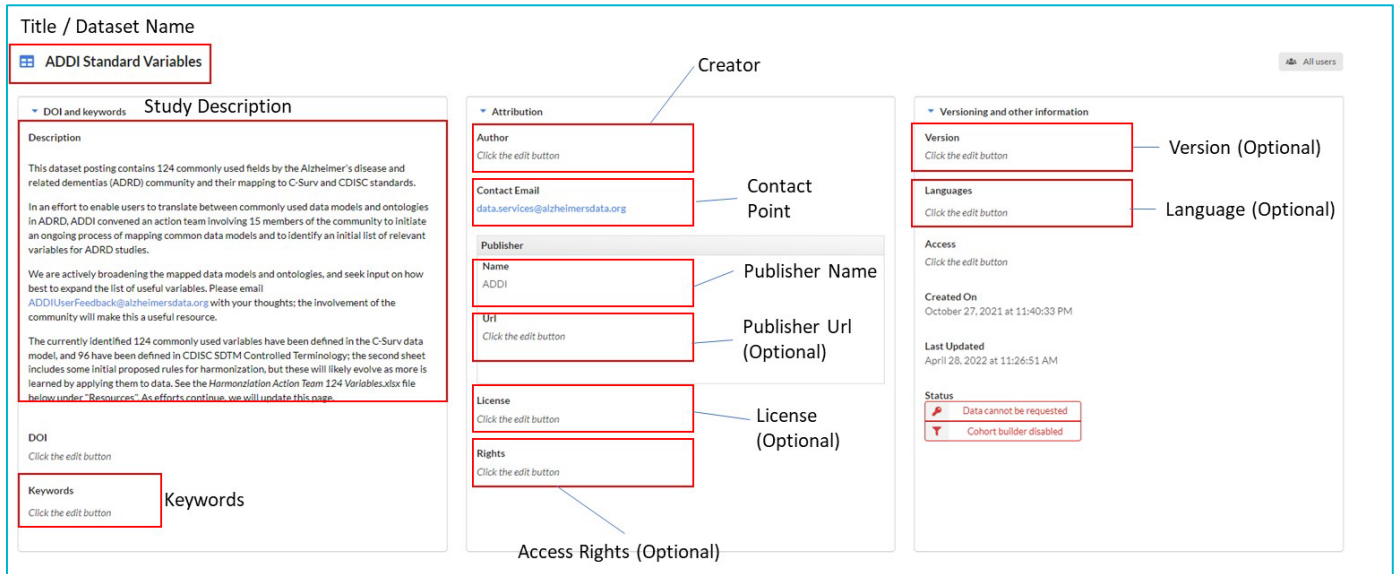


Figure 1. Annotated Fields for Catalogue

Dictionaries

Oftentimes the data for a given study are separated into more than one data table. For each data table, we require a brief name, label and description; if there is just one table then a single name, label and description will need to be supplied. Figure 2 shows how this information is displayed in the ADDI portal dataset catalog entry. When requesting access to a study/dataset, a user has the option to request either the full set of data tables or any given subset, and this list is indexed by the Data Table 'Name' field. Please use the 'dictionaries' tab in the 'FAIR Metadata Template' spreadsheet to populate these fields for each data table in your study. There should be one row per data table. The particular information is:

1. **Code:** an alphanumeric code to index the data dictionary. The code should be unique and can only contain letters, numbers, and underscores. The "code" defined here is also referenced in the "fields" tab as "dictionary_code"/

2. **Name:** a short name to identify the data table, displayed at the top of the data dictionary; this is the name that will be presented to the user in the Data Access Request form
3. **Description:** a phrase or sentence describing the information and variables stored in a given data table (field not shown in Figure 2)

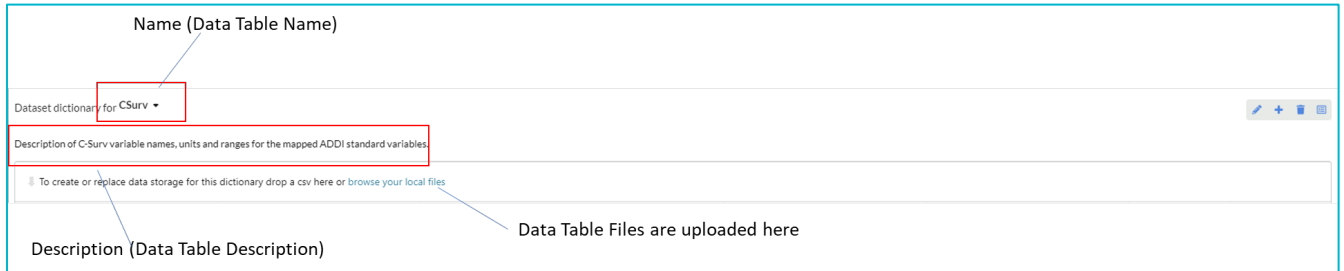


Figure 2. Annotated Fields for Dictionaries

In the case of a dataset that includes omics, as well as tabular clinical data on the study participants, we can help design how the data tables are presented. Figure 3 shows an example of how omics data could be displayed: a table of gene expression values where rows correspond to the probes and columns correspond to the study participants. Three total tables are defined in the FAIR posting (<https://fair.addi.ad-datainitiative.org/#/data/datasets/gse46579>): one table for the expressions, one for the probes, and one for the participants.

Dataset dictionary for **expression-table** 📄

Table of expression values, where rows=probes and columns=targets. Please see additional annotation files for information.

Field name	Field label	Type	Description	URI	Entity
Probe	PROBE	Text	Probe; please see probe-annotations for more details		<input type="checkbox"/>
Target	TARGET	Text	Target; please see target-annotations for more details		<input type="checkbox"/>

Dataset dictionary for **probe-annotations** 📄

Annotation file with information on probes

Field name	Field label	Type	Description	URI	Entity
Id	ID	Integer	Unique probe identifier		<input type="checkbox"/>
Probe	PROBE	Text	Name of probe		<input type="checkbox"/>

Dataset dictionary for target-annotations ▾

Annotation file with information on targets

Field name	Field label	Type	Description	URI	Entity
geo_accession	GEO ACCESSION	Text	Unique target id		<input type="checkbox"/>
target	TARGET	Text	Identifier of control or Alzheimer's Disease brain		<input type="checkbox"/>

Figure 3. Data Table Definitions for Example Omics Study

Fields

For each distinct data table, we require an accompanying metadata (also referred to as fields). The 'fields' tab lists each variable as it appears in the data, a label, a variable type, and a brief description of what information is conveyed with the variable. Figure 3 shows the data dictionary annotated with the required fields.

Please use the 'fields' tab in the 'FAIR Metadata Template' spreadsheet to supply these details for each variable in your dataset. There should be one row per variable per data table.

The details for the required information are:

- dictionary_code:** This should be one of the "code" values defined in the 'dictionaries' tab. It represents the data table for which variables are being described
- name:** the exact name of the variable as it's stored in the data table, also commonly referred to as 'variable name'; for example, 'sysbp'. Please note that field names can only contain letters, numbers, and underscores and they cannot start with a number.
- label:** a brief label for the variable, oftentimes an expansion of the field name, such as 'systolic blood pressure'
- type:** the type of variable; must be one of: **text, integer, float, date**, date with timestamp, or **boolean** (True/False)
- constraints:** for variables that have constrained vocabulary, the name of the lookup table; the lookup tables are defined in 'Field Lookups' (see next section for more details); please leave blank for non-constrained variables
- description:** a phrase or sentence describing the information stored in a given variable, such as 'systolic blood pressure measured at baseline'
- uri (optional):** an external web uri that further describes the variable
- entity (optional):** a flag to indicate whether the variable can be an entity in the database. It can take TRUE/FALSE as values. Leave blank if not sure

Data Table Name

Dataset dictionary for **CSurv**

Description of C-Surv variable names, units and ranges for the mapped ADDI standard variables.

To create or replace data storage for this dictionary drop a csv here or browse your local files

Field name	Field label	Type	Description	URI	Entity
ABUSEPRT0r18	ADDI_var1	(Y/N) Text	physical abuse by parent during ages 0-18 (Y/N)	Click the edit button	<input type="checkbox"/>
ABUSEPRT19r25	ADDI_var2	(Y/N) Text	physical abuse by parent during ages 19-25 (Y/N)	Click the edit button	<input type="checkbox"/>
ABUSESEX18r25	ADDI_var3	(Y/N) Text	sexual abuse during ages 18-25 (Y/N)	Click the edit button	<input type="checkbox"/>
ACCTYPE	ADDI_var4	ACCTYPE_ENV Text	type of accommodation	Click the edit button	<input type="checkbox"/>
ACTMODDYWK	ADDI_var5	Integer	moderate exercise per week (hours)	Click the edit button	<input type="checkbox"/>
ACTVIGDYWK	ADDI_var6	Integer	vigorous exercise per week (hours)	Click the edit button	<input type="checkbox"/>
ACTWALKDYWK	ADDI_var7	Integer	walking per week (hours)	Click the edit button	<input type="checkbox"/>
ADASTOT	ADDI_var8	Integer	Alzheimer's Disease Assessment Scale-cognitive (ADAS cog) total score	Click the edit button	<input type="checkbox"/>
ADDX	ADDI_var9	(Y/N) Text	Alzheimer's Disease (AD) diagnosis (Y/N)	Click the edit button	<input type="checkbox"/>

Field name Field Label Field Type / Field Constraints Field Description

Figure 4. Annotated Fields for Fields

Lookups

Oftentimes variables are coded, and their meaning needs to be defined. For example, if sex is stored as 1 and 2, a user needs further definition. This is what we call a 'constrained' or 'coded' variable. For each of these variables, the 'constraints' entry in the 'fields' tab must be non-blank (see previous section) and the corresponding entries in the 'lookups' table must be populated. For each constrained variable, there must be one row in 'lookups' for each of the possible values the variable might take. In the sex example, two rows would be populated in 'lookups'. Please use the 'lookups' tab in the 'FAIR Metadata Template' spreadsheet to supply these details for each variable in your dataset. The details that must be supplied are:

1. **lookup**: the exact name of the constraint as it appears in the 'fields' tab, such as ACC_TYPE (field shown in Figure 4)
2. **name**: the values allowed for the constrained variable, such as M and F (these will be two rows in the table)
3. **description**: a definition of what the constraint means, such as Male and Female (these are entries in the two rows described above)
4. **uri (optional)**: an external web uri that further describes the lookup

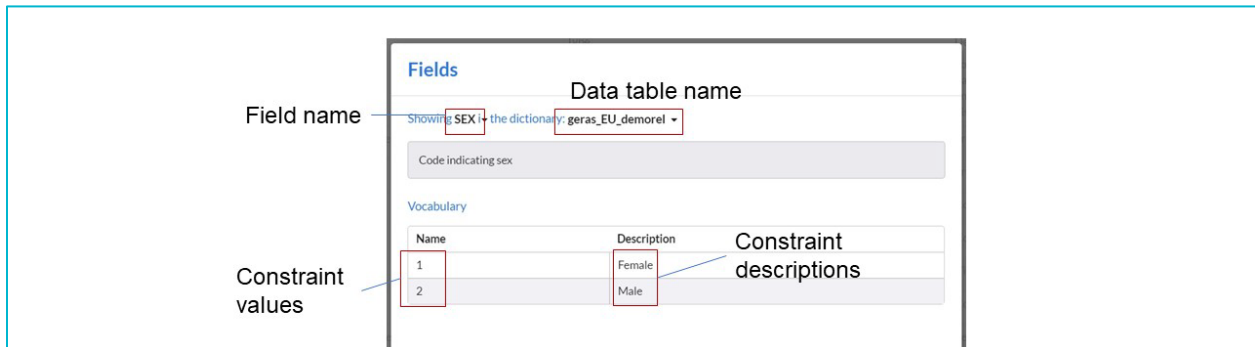


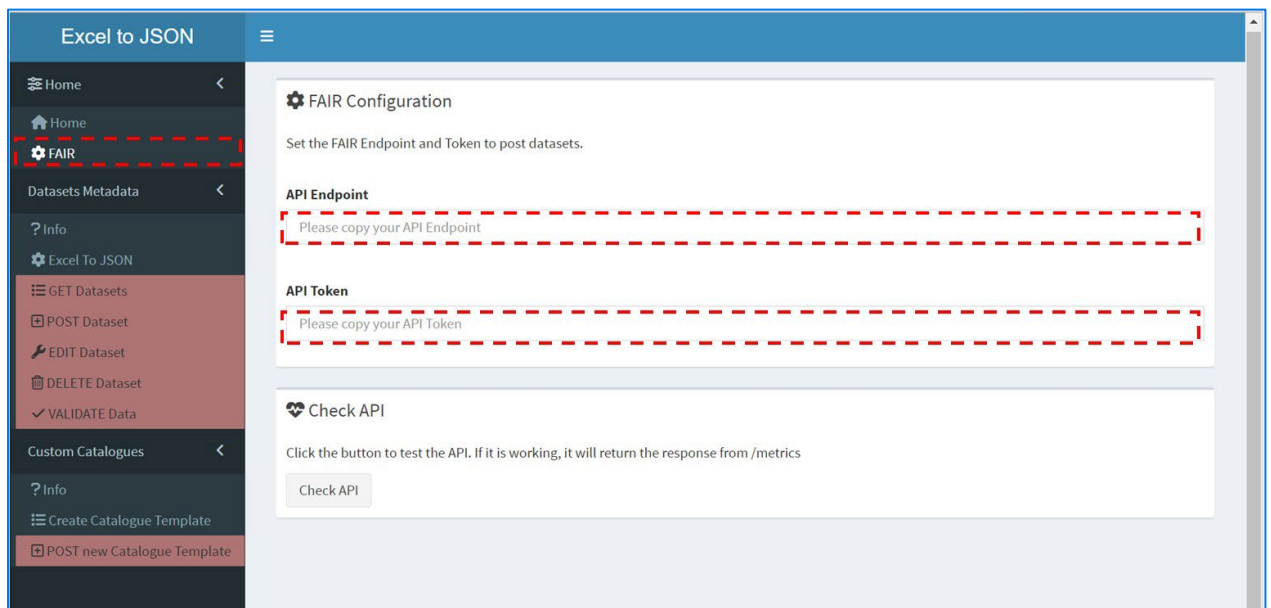
Figure 5. Annotated Fields for Lookups

Dataset Creation and Upload

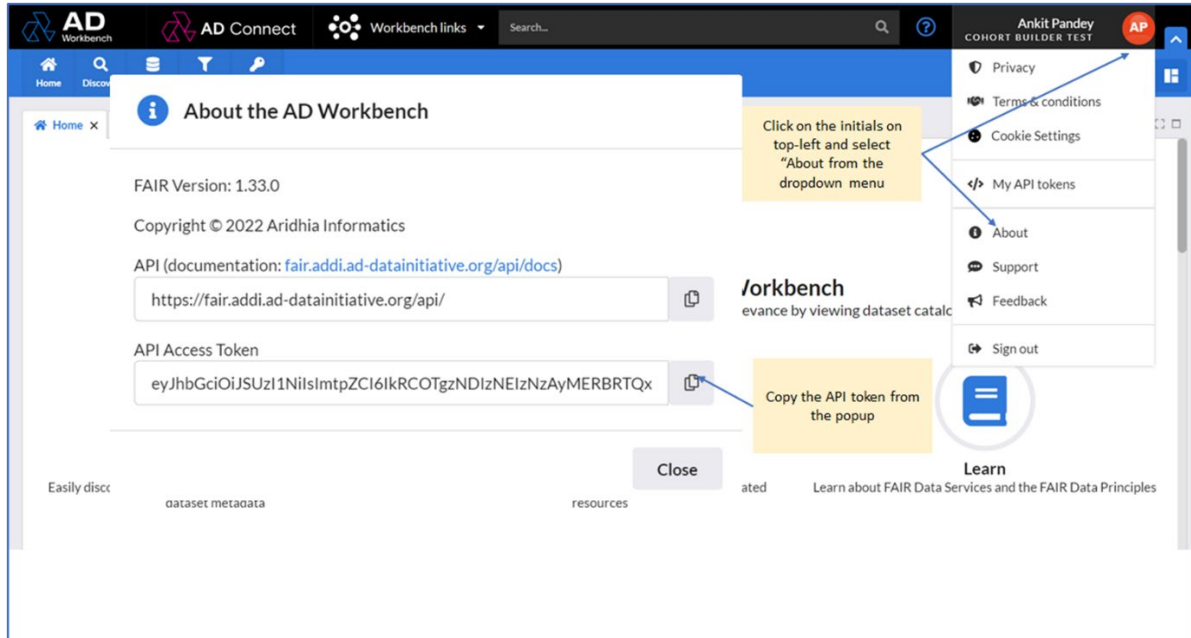
The next step in data curation and preparation is to create a dataset entry in FAIR and upload dataset files*. For creating dataset entries in FAIR, we have two options – first is a low-tech option using the [shiny app](#) and second using [FAIR APIs](#). We recommend using the shiny app for creation of dataset entries for all data providers. For onboarding a dataset, it is essential to have “Data Steward” access in FAIR. If you don’t already have Data Steward access, put in a request at servicedesk@aridhia.com

The steps for creating dataset entry using the shiny app are listed below:

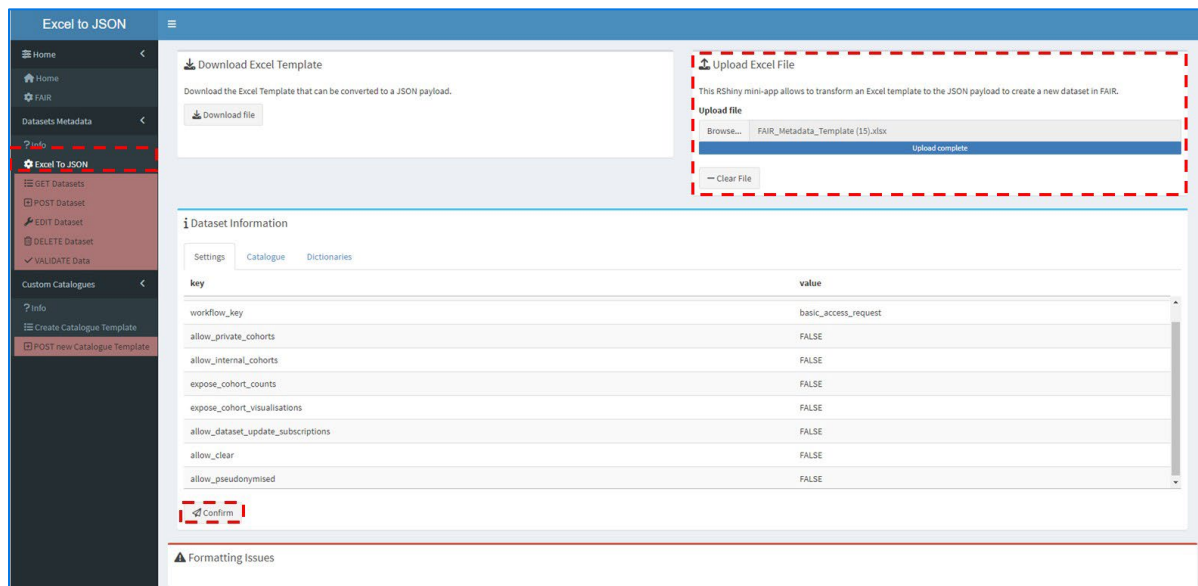
1. Enter FAIR API endpoint and FAIR API tokens in the shiny app as described below.



2. Get the API token from FAIR as shown below.

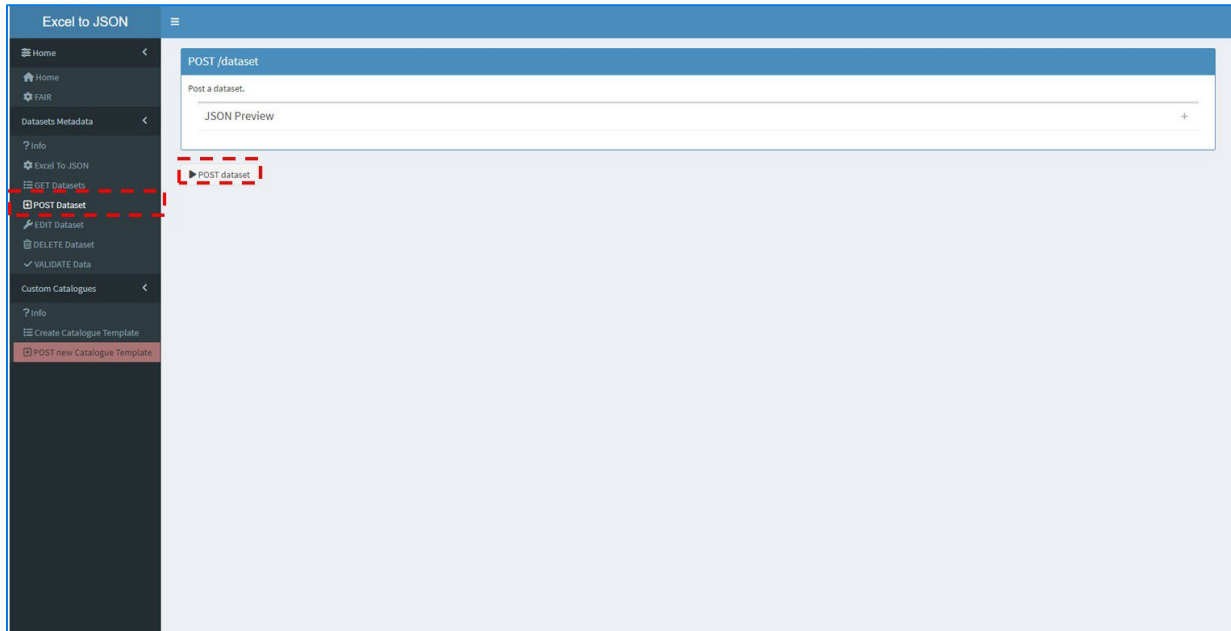


3. Upload the FAIR Metadata excel template and look for “Formatting Issues” at the bottom of the page. If nothing shows up, you can press “Confirm” to generate the dictionary.

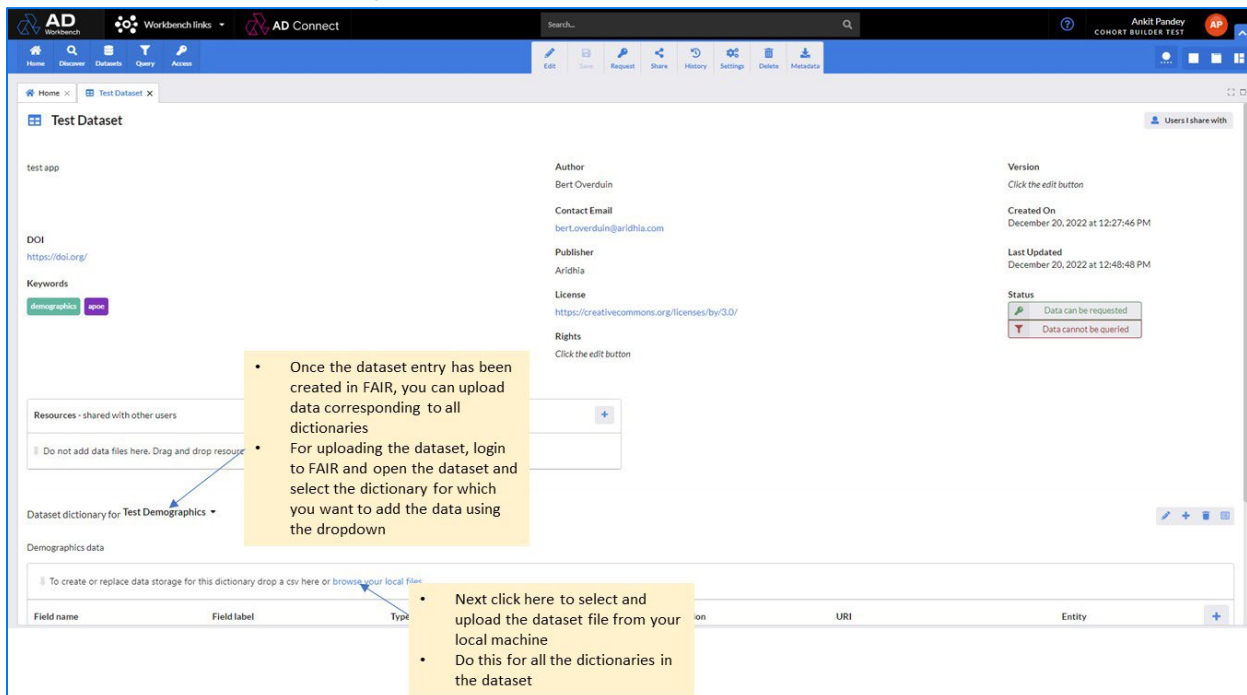


4. Post the created dictionary to FAIR. If you see a successful response from the API then it means the dataset has been successfully posted to FAIR. You can then login to FAIR to

check the newly created dataset entry. If you get an error response from the API, try to correct the issues described in the response in the excel template. If you are unable to identify the issues, share the metadata template with ADDI tech team and they will help you create the dataset entry in FAIR



5. Upload dataset files against the dictionaries in FAIR as described below



Once the dataset entry has been created in FAIR, you can upload data corresponding to all dictionaries

For uploading the dataset, login to FAIR and open the dataset and select the dictionary for which you want to add the data using the dropdown

Next click here to select and upload the dataset file from your local machine

Do this for all the dictionaries in the dataset

Glossary

1. **Constrained variable:** a variable that is coded or only has limited values, such as SEX coded as M or F, or recruitment site coded as 1, 2, 3, where 1=London, 2=Seattle and 3=Tokyo
2. **Data access request (DAR):** a request from an ADDI user to a data provider to access their data; oftentimes a user must provide information such as an analysis plan or estimated timeline.
3. **Data table:** a table containing variables which is part of a larger study or dataset.
4. **Dataset:** a discrete set of data, usually coming from a single study and oftentimes consisting of numerous data tables
5. **Field:** variable
6. **Metadata:** a set of information and descriptions about a list of variables or datasets; oftentimes used interchangeably with data dictionary
7. **Variable:** a single measurement collected across a set of study participants