Session 1: Overview of CSPro, Dictionary and Forms

At the end of this lesson participants will be able to:

- Identify different CSPro modules and tools and their roles in the survey workflow
- Create a simple data entry application including dictionary and forms
- Run a data entry application on Windows
- Run a data entry application on Android and retrieve the data entered
- Understand the differences between the new CSPro DB format and the old text format for data files.

CSPro Overview

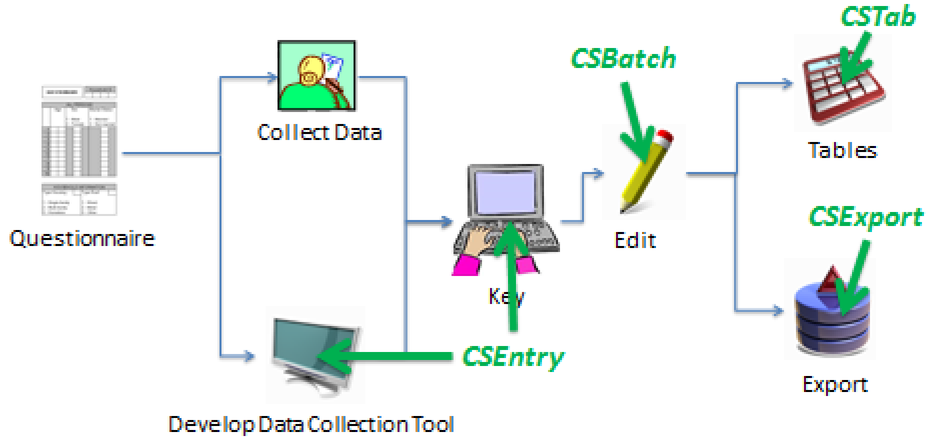
CSPro is a suite of software tools for census and survey data processing that includes modules for data collection, editing, tabulation, and dissemination.

CSPro has a long history. It was first released in 2000 and has been used in over 100 countries worldwide. It has been used for censuses all over the world as well as for many large and complex household surveys including the Demographic and Health Survey (USAID), Multiple Indicator Cluster Survey (UNICEF) and Living Standards Measurement Study (World Bank). The first Android version was released in 2014. CSPro Android has already been used in production for household surveys and population censuses in multiple countries.

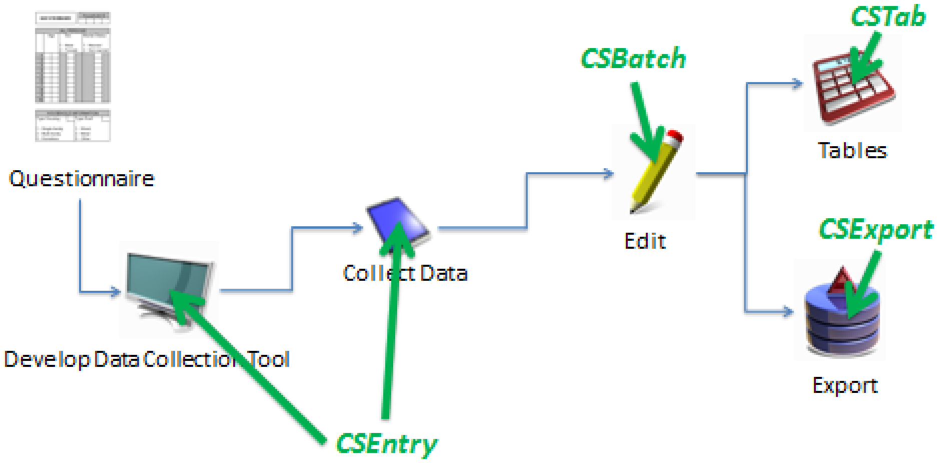
CSPro is free software developed by the US Census Bureau and funded by USAID. The Census Bureau provides free email customer support. You can send questions to cspro@lists.census.gov.

CSPro can be used for both the traditional PAPI (pencil and paper interview) workflow as well as the computer aided personal interview (CAPI) workflow. In this workshop, we will focus on data collection in using a CAPI workflow.

Paper and Pencil Workflow



CAPI Workflow



Using CSPro Android

Before we learn how to build data entry applications in CSPro, let's try doing some data collection to get comfortable with the system.

Group exercise

Split into groups of 3-4 people and use the provided tablets to interview each other using the "Getting to Know You" application. Interview each member of the group so that we have data for all workshop participants. When you are done, tap the sync button (🔄) to upload your results to the server.

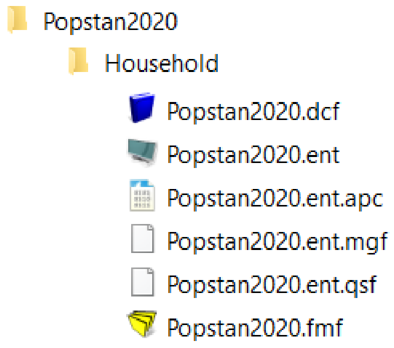
Creating a Data Entry Application

For most of the examples in this workshop, we will be creating a data entry application based on the Popstan 2020 Census questionnaire included with the workshop materials. With CAPI applications, a paper questionnaire alone is not sufficient to define a data entry application. We also a need a specification document that describes consistency checks, skip patterns, text fills, error messages and other aspects of the interactive application that are not defined on a paper questionnaire. Take a moment to review both the questionnaire and the accompanying specification document.

When you launch CSPro you are given the choice of "Data Entry Application" for key from paper (PAPI) and "CAPI Data Entry Application" for electronic data collection using phones/tablets/laptops. The differences are:

- System controlled (tightly controlled path)
- CAPI question text
- Extended controls (radio buttons, checkboxes, date picker, …)

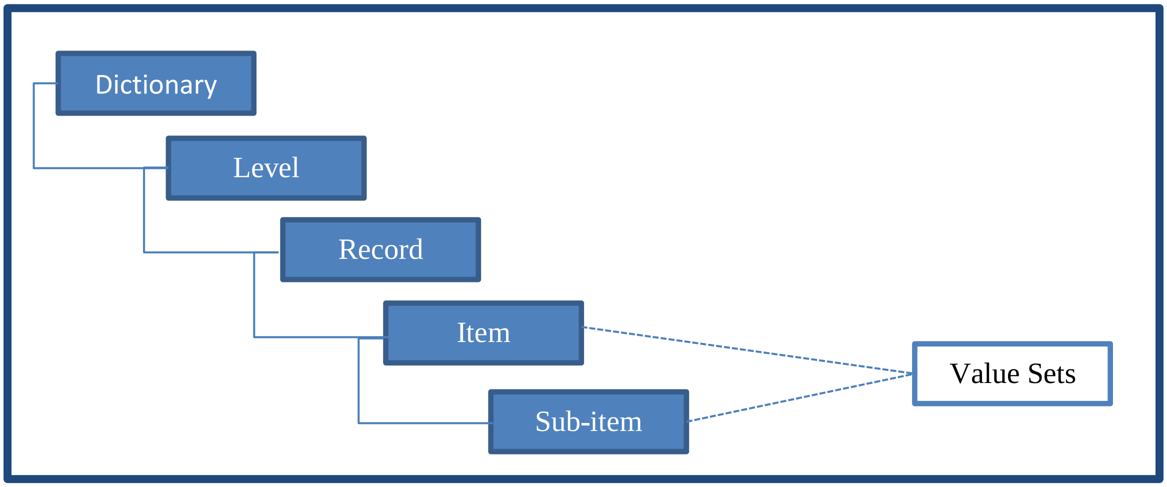
Since we are creating a CAPI application we will choose "CAPI Data Entry Application". We will name the application "Popstan2020" and we will use the same name for the dictionary. Since we will eventually add other applications such as the listing questionnaire and the menu, we will create the following folder structure:



```
Popstan2020
    Household
        Popstan2020.dcf
        Popstan2020.ent
        Popstan2020.ent.apc
        Popstan2020.ent.mgf
        Popstan2020.ent.qsf
        Popstan2020.fmf
```

The Data Dictionary

The first step in creating the application is to define the data dictionary. The data dictionary lists all the data items and possible responses that will be in the application and organizes them in records and levels. The dictionary has the following hierarchy:



Before defining the record and items we first create **ID**-items. **ID**-items uniquely define each case (each questionnaire). Usually these are geographic codes.

What are the **ID**-items for the example questionnaire?

Province, District, Enumeration Area, Area Type, and Household Number.

Why not include GPS, interview date, start and end time since they are in the same section of the questionnaire? Because they are not part of the codes needed to uniquely identify the questionnaire.

Add the id-items to the dictionary

- Province (numeric, length 1)
- District (numeric, length 2)
- Enumeration area (numeric, length 3)
- Area type (numeric, length 1)
- Household number (numeric, length 3)

Properties of dictionary items:

- **Label:** text displayed to interviewers (not the full question text, we will see where to put that in a later lesson).
- **Name:** used to refer to variable in CSPro logic (interviewers will never see this)
- **Start:** column number of first digit (character) of variable
- **Len:** number of characters/digits used to store variable

- **Data type:** alpha for names and other text, numeric for numbers and coded responses
- **Item type:** whether variable is item or sub-item
- **Occ:** number of occurrences for variables that are repeated in record
- **Dec:** for numbers with fractional parts, number of digits after decimal point
- **Dec char:** whether or not to include decimal point in decimal numbers.
- **Zero fill:** for numeric values; whether to add zeros or blanks to left of number when number of digits in value is less than the length of the variable. Using this will avoid problems when dealing with subitems. We can enable zero fill by default on the options menu.

> Tip: Note that you can toggle showing names or labels in the dictionary tree on the left side of the screen using the View menu. You can also select "Append names to labels in tree" to show both at the same time.

What are the records used in our survey?

Record Type	Record Name	Section(s) of the Questionnaire
A	INTERVIEW_REC	A. Non id-items (A6-A10)
B	PERSON_REC	B. Demographics, C. Education, D. Fertility,
E	DEATHS_REC	E. Deaths of Household Members in the Past 5 years
F	HOUSING_REC	F. Housing Characteristics
G	POSSESSIONS_REC	G. Household Possessions
H	AGRICULTURE_REC	H. Agriculture

Note that we could have separate records for education and fertility but instead we will combine them with the person record. This will simplify analysis later on since we will not have to link the records together. Later on, we will see that even with the records combined we can still put education and fertility into separate rosters on our forms.

Let's start by just creating the housing and person records.

Properties of records:

- **Label/name:** same as for variables.
- **Type Value:** to distinguish between different records in the data file.
- **Required:** whether or not the record must be entered for the questionnaire to be complete.
- **Max:** for multiply occurring records the maximum number of occurrences allowed. Generally 1 for singly occurring records (like housing) and a larger number for repeating records (like 50 for household members). Note that CSPro doesn't allocate space in the data file for occurrences that are not used so it is better to err on the side of caution and allow extra occurrences.

What are the properties of the housing record?

- Type value: F (can use anything but nice to use something meaningful like section letter)
- Required: no (we will not collect section F for vacant/refusal)
- Max: 1

for the person record?

- Type value: B
- Required: no (we can have empty households)
- Max: 30 (questionnaire has limit of 10 but no penalty for adding a few extra just in case)

Now add some fields to the person record:

- Person number (numeric length 2)[1]
- Name (alpha length 30)
- Relationship (numeric length 1)
- Sex (numeric length 1)
- Age (numeric length 3)

And to the housing record:

- Number of rooms (numeric length 2)
- Type of main dwelling (numeric length 1)

A note on variable naming

Different people have different styles of naming dictionary variables. Some use a descriptive name such as "PLACE_OF_BIRTH" others prefer to use the question number such as "B07" and others prefer a combination such as "B-7_PLACE_OF_BIRTH". Whichever approach you choose just make sure that it will be easy for users of your application and your data to understand. Will everyone working on the logic for your application know what B07 is?

For each of our variables we need to add the possible responses (value sets). The value set lists all valid responses along with their corresponding labels for coded variables. Without a value set, the interviewer can enter any value (except blank) but with a value set they are limited to the options defined in the value set. Without a values set, users can even enter negative numbers. For this reason, it is good practice to use a value set for all numeric variables.

Define the value sets for some of our variables based on the response codes on the questionnaire:

- Area type (1- Urban, 2- Peri-urban, 3- rural)
- Province (copy/paste from Excel)

[1] The line number is not needed in CSPro itself as there are ways to determine the row number using logic, however, when exporting the data to other packages it is often useful to have it. We will see later how to fill this in automatically during data entry.

- Person number (range: 1-30)
- Name (no value set)
- Sex (1- Male, 2-Female)
- Relationship (see questionnaire)
- Age (use a range: 0-120, plus don't know code 999)
- Number of rooms (use generate value set to generate the codes 1-20)
- Type of main dwelling (use codes from questionnaire)

Dictionary Macros

There are some useful functions for working with dictionaries that you can access by right-clicking on the dictionary in the tree on the left side of the screen and choosing "Dictionary Macros". In particular you can copy/paste all value sets or all item names/labels from the dictionary to/from Excel. This can be used to create codebooks to share with people who do not have access to CSPro. It can also be used to do bulk modifications on dictionary items such as renumbering values in value sets or adding prefixes to item names.

Forms

Before we can enter data, we need to create data entry forms. To start, click on the yellow stack of forms on the toolbar. To follow the look of the paper questionnaire we will create one form for each page of the paper questionnaire.

Create a form for section A: Identification*. Drag and drop the id-items onto the form. Note that we can drag and drop individual items or entire records. By right clicking on the form in the forms tree on left side of the screen we can change the label and name of the form. Let's make the label "A: Identification" and make the name "IDENTIFICATION_FORM".*

Create a form for section B: Demographics. *Drag drop the items from the person record. Let's give the form label "B: Demographics" and name "DEMOGRAPHICS_FORM". Note that when we drop the record we have the option to put the items in a roster or a repeating form. If we drop the items on the household identification form, we can only roster since the household identification isn't repeated. For our example let's use a roster.*

When we create the rosters, CSPro automatically gives them a name that ends in "000", for example "PERSON000". You can see this in the forms tree on the left side of the screen. We can change this by right clicking on the roster in the forms tree and choosing properties. Let's name our roster "DEMOGRAPHICS_ROSTER".

Create a form for section: F: Housing Characteristics. *Drag and drop the items from the housing record.*

For paper and pencil surveys, we would spend a lot of time on the layout, adding additional labels and frames to make the form look exactly like the questionnaire. However, when rendered on Android, the form is rendered one question at a time so making the form look like the paper form is not as important.

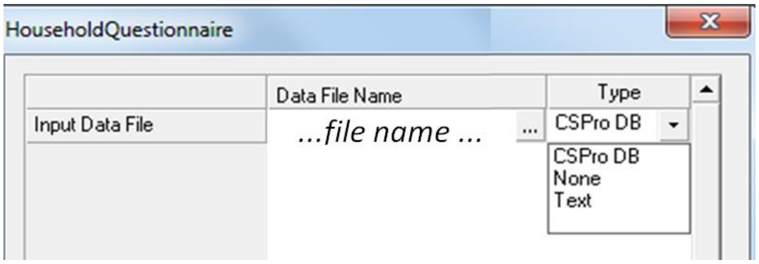
Running the Application on Windows

Click on the green light icon on the toolbar to run the data entry application on your PC. Enter the name of the data file. For data files CSPro uses the file extension ".csdb". Let's name our data file "Household.csdb".

Data File Types

The csdb extension is new in CSPro 7.0 and represents a new file format, the CSPro Data Base File. This file not only contains the data itself but it also contains the notes, the index, the partial save status and metadata used for data synchronization. In earlier versions of CSPro, the notes, index and partial save were stored in extra files that accompanied the data file itself. It was unwieldy to deal with all of these files so the CSPro DB file combines them all into one file. Unlike the text file used by earlier versions of CSPro, this is a binary file that cannot be viewed using TextViewer. For the final CSPro 7.0 release we will have a data viewer tool to see the data contained in a CSDB file which will play the role that TextViewer does for text files.

In CSPro 7.0, when you launch a CSPro data entry application you can select the type of data file that you want to use:

 HouseholdQuestionnaire Input Data File | Data File Name: ...file name ... | Type: CSPro DB 	**CSPro DB**: This is a CSPro Data Base File. It contains the data and metadata **None**: There is no associated data file. This is useful when building menus. **Text**: This is a UTF8 Text file. This is what CSPro has used in previous versions.

While it is still possible to use text files in CSPro 7.0, it is highly recommended to use CSPro DB instead. There are new features such as smart sync and case labels that are not supported using text files.

Controlling the size of the roster

Our program at this point has no way to stop entering household members unless we fill in all 50 rows of the roster. To limit the number of rows in the roster to the number of actual household members we can have the interviewer enter the number of household members and use that to set the size of the roster. Add a new variable, "number of household members", to control the size of the roster. Which record should we add it to? It can't be on the person record since we don't want it to repeat so let's add it to the household characteristics record but put it on the person (section B) form. Right click on the roster, choose properties and set this new field as the occurrence control field.

Try running the application again. Unfortunately, when we get to the person form we go into the roster instead of the “number of household members” field. By default, CSEntry goes in the order in which the fields were dropped on the form. We can change the order by dragging the fields in the forms tree. Move the “number of household members” field up in the tree so that it comes before the roster. Now the roster control field should work. In the next lesson we will use logic to exit the roster without the roster control field.

Running the Application on Android

To put the application on an Android phone/tablet we need to do the following:

1. Publish the .pen file (File menu → Publish Entry Application).
2. Connect the tablet to the PC using a USB cable. The tablet should show up like a USB drive.
3. Copy the files Popstan2020.pen and Popstan2020.pff into the CSEntry directory on the Android device.
4. Start CSEntry on the Android device
5. Enter the data (note the differences between Android and Windows)
6. When you are done, connect the tablet back to the computer and copy the file Household.csdb from the tablet back to the PC. One some tablets (Nexus 7 for example) the first time the data file is created it may not show up when connected to the PC. If this is the case, reboot the tablet and it will.

When we copied the application to the tablet we copied the pen and the pff file. The pff file contains various parameters about how to launch the data entry application including the data file to use. You can modify the pff file by right clicking on it in Windows Explorer and choosing “Edit”. This allows you to modify the name of the data file, force the application to start in add or modify mode and to lock various parts of the user interface.

On Android, the list of available applications on the device is constructed by finding all the pff files in the CSEntry directory and subdirectories so a pen file without a pff will not show up in the list.

Partial Save and Case Tree

It can be a pain when testing a question that is on the third form of a survey to have to reenter all of the data up to the question you are testing. We can enable partial save under the data entry options so that we can exit data entry, modify the application, and come back right to where we left off. While we are in the data entry options we can also enable the case tree on both Windows and Android to make it easier to navigate around the questionnaire while we are testing. The case tree is enabled by default on Android since Android only shows one question at a time but it is off by default on Windows. Note that on Android phones, since there is not enough space to display the case tree and the questions at the same time so you need to tap on the big green “CS” in the top left corner of the screen to bring up the case tree.

Modifying the dictionary after collecting data

Add the variable **F02 number of bedrooms** to the dictionary and form right before F03 tenure status. Don't forget to modify the order in the form tree so that it is asked before F03. Run the application again and modify one of the existing cases. Why is the number of household members blank? Once data has been entered, you should avoid adding variables *in the middle* of a record. It is fine to add to the end of the record but adding in the middle invalidates the start positions of the existing variables. If you must add a variable in the middle of a record, you can use the reformat data tool to adjust old data files to match the new dictionary.

Group exercise

Add a new record to the dictionary for section E of the questionnaire (deaths). Name this record "Deaths". How many occurrences should it have? Don't include E01 and E02 in the new record as they do not have the same number of occurrences as the other variables in this section. Instead, add them to the housing record. Create a new form for section E and add the fields onto it to create a roster. Use E02 as an occurrence control field to the roster to limit the number of rows to the number of deaths in the household. Test the application on both Windows and on Android.

Subitems

Let's add the **Date of Birth** (**B06**) to the application. In order to be able to look at both the date as an 8-digit number and look at the day, month and year individually we can create an item with subitems. Subitems are items that are made up of a subset of the digits of their parent item. Add the item for interview date and the following subitems:

- year of birth (length 4)
- month of birth (length 2)
- day of birth (length 2)

We are putting year first then month and day because this format will work better with other CSPro features that we will see later. Click on **Layout** in the toolbar to ensure that the item and subitem overlap. Add the subitems to the form, add the value sets for each subitem and test the application. Note that when we add the subitems to the form we do not need keep the same order that we have in the dictionary. On the form we can put the day, month then the year.

Linked Value Sets

Note that questions **B07 (place of birth)** and **B08 (residence 1 year ago)** both have the same value set. We could simply copy and paste the value set into both items, however, this would leave us with two copies of the same value set. If later on we change one of them and forget to change the other, then they will be out of sync. This could cause consistency problems later on. Instead, we can create the value set for **B07** and then paste it as a linked value set into **B08**. With a linked value set, CSPro only stores one copy of the value set that is shared between both variables. This way if you edit the value set, the change is reflected in both items.

Multiply Occurring Items

Let's add question **F04** on number of housing units to the application. Since this question asks the number of units of each of five different types we could make five different variables in the dictionary: HOUSING_UNITS_ROUND_HUT, HOUSING_UNITS_DETACHED, HOUSING_UNITS_SEMI_DETACHED etc… To simplify our dictionary, we can instead make a single variable in the dictionary with five occurrences: one for each type of housing unit. When we drag this variable onto the housing characteristics form we will get a roster with 5 rows.

Occurrence Labels

With this approach, our form does not show the housing unit types but we can fix that by using occurrence labels. Select the housing units variable in the dictionary and choose "Occurrence Labels…" from the Edit menu. Add the names of the five types of housing units in the grid that comes up. Note that you can copy from Excel and paste into this dialog. Now when we drag the variable to the form the roster shows the type of housing unit for each row.

Checkboxes

We could also use a multiply occurring item for question **B10**, disabilities, but that can be implemented more easily using checkboxes. Checkboxes offer a friendly interface for multiple response questions by presenting a single screen with a checkbox for each option rather than presenting the options one by one.

In CSPro multiple response questions are implemented as alpha variables whose length is the same as the number of options that can be selected at the same time. The value set has a value for each option which is usually a single letter. The resulting value is a string containing the values for each of the selected items.

For disabilities we will use the following value set:

Visual	A
Hearing	B
Speech	C
Physical	D
Mental	E
Self-care	F

If the interviewer checks the boxes for Visual, Physical and Mental the value for the variable will be "ADE". We will see later how can convert the alpha value into a series of yes/no values to simplify analysis.

Capture Types

If you right click on the disabilities field on the form and choose “Field Properties…” you will see that the capture type is set to check box. There are other capture types:

- Text Box: keyboard data entry
- Radio button: choose one option from many with radio button for each option
- Drop Down: choose one option from many without radio buttons.
- Combo Box: combination of keyboard input and drop down (same as drop down on Windows)
- Check Box: choose multiple responses

When you drag an item on the form, CSPro sets the capture type based on the value set for that item. If there is no value set when you drop the item, the capture type will be set to text box. You can always change the capture using the Field Properties dialog.

Date Fields

Let’s add the interview date to the identification section. Which record should it go on? It could go on any singly occurring record such as housing but let’s create a new record called INTERVIEW_REC to hold the section A items that are not part of the id-items and add it there. We can add an eight-digit item for the interview date. We can also add sub-items for the year, month and day of the interview. Drag the date item onto the form. When dragging don’t use the sub-items, just the items. Now change the capture type for the interview date to be Date and set the date format to be YYYYMMDD to match the order of the sub-items year, month and day.

Filling in Names First

Currently the interviewer has to fill in all the demographic information for the first person before moving on to the next person in the household. In practice, it is simpler to have them fill in the names of all the household members first and then fill in the demographic details only after all the names have been entered. To do this we need to pull just the NAME and PERSON_NUMBER fields (B01 and B02) into a separate roster. Delete them from the existing DEMOGRAPHICS_ROSTER, create a new form before the DEMOGRAPHICS_FORM called NAMES_FORM and drop PERSON_NUMBER onto it to make a new roster. Rename this roster “NAMES_ROSTER”. Drop NAME from the dictionary on top of the NAMES_ROSTER to add it to the roster. Note that when you drop an item from a repeating record on top of an existing roster it gets added to that roster but if you drop it outside the roster it creates a new roster. Remove the NUMBER_OF_HOUSEHOLD_MEMBERS field from the DEMOGRAPHICS_FORM and drop it onto the NAMES_FORM. In the forms tree, move NUMBER_OF_HOUSEHOLD members so that it comes before the NAMES_ROSTER. Set the occurrence control field for NAMES_ROSTER to be NUMBER_OF_HOUSEHOLD_MEMBERS.

Tip: When dragging items onto a form with an existing roster drop the items inside the roster or you will end up with a second roster for the new item.

Exercises

1. Add the remaining fields from the housing section of the questionnaire to the dictionary and the housing form (F5 through F12). Make sure to add the appropriate value sets.
2. Add the remaining fields from the demographics section (B) of the questionnaire to the dictionary and to the demographics form. Make sure to add the appropriate value sets.
   a. For **Occupation(B17)** use only the 4-digit occupation codes. This is hierarchical coding scheme and we only want the last level (level 4) codes. Copy the occupations codes from the Excel Spread Sheet “QuestionnaireAnnexes.xlsx” to the value set.
   b. For **Language(B18)** use checkboxes.
3. Add the fields for section C, Education, to the dictionary. Add them to the person record. Create a new form for section C and drop the education items onto to it to create a roster. Set the occurrence control field for the roster to the number of household members.
4. Add the start and end times of the interview to the identification form after the interview date. Use subitems for the hours and minutes. Add appropriate value sets.
5. Add a new record and form for section G, household possessions. Since G01 (quantity and value) repeat, put these items in their own repeating record but do not include G02. Set the occurrence labels in the roster for G01 to the names of the possessions. Make G02 a singly occurring checkbox field and put it in the housing record since it does not repeat. The value set for G02 should have the possession names as labels with codes “A”, “B”, “C”…

Make sure to test your application on both Windows and Android.