2. Setting Up Data in the Data Repository - Transcript

Introduction [00:00]

Hello, and welcome to the second video about the RECAP Preterm data platform. My name is Gonçalo and in this video, we'll be showing you how to work with one of the main components of a RECAP node, which is a Data Repository.

Now, if you're not yet familiar with what a RECAP node is, and the different components that it includes, I suggest you go back and watch our first video about the RECAP data platform. So you can come into this one with a bit more context and hopefully have a better understanding of the things that I'm going to talk about. So if you have not yet watched the first video, you can find a link to it in the description. The video is entitled, overview of the RECAP Preterm data platform.

Now before we start, I just want to mention that throughout this video, and also the remaining videos, I will be demonstrating how to work with a recap platform by essentially following the steps described in a wiki that we had created for this project. So if you'd like to follow along, you're welcome to try it for yourself, you can find the link to the wiki in the description of this video as well. So, let's get started!

Creating a Project [01:17]

This is the wiki I just mentioned, this is the one we're going to be following the little introduction here, just to remind you of the four different main components of a node, like we discussed in the first video, but for this video, we're going to focus on this part, the Data Repository. So what we're going to do is import data into the Data Repository of a node. So for that, we're going to be following this first section here, setting up data in the Data Repository, and then there's four more sections, so one for each of the four remaining videos, but for this video, we'll just cover this, this first part. So let's click on that.

You can see here, there's an index, this is basically a two step process. So the Data Repository, organises data in projects, so we have to first create a project and then, inside that project, you can create multiple data tables. So if you want to put data in our node, we first have to create a project. So that's the first step here. And then once we do that, we can move on to importing data into that project. Basically, there's two ways of doing that. It depends on the format of your data. So we'll do both, we'll cover both of

these. So the first one will be using your raw data format such as a CSV file.But because it's a raw data format, by the way, what I mean by raw is that the file only contains the data itself, there's no metadata. So, no variable labels or types or categories or anything like that. It's just the actual data, the raw data. Because it's raw data, you have to complement it with the dictionary file.

We'll use a spreadsheet. So this dictionary file for that, where we describe the variables, you know, the labels and the types of each variable. I'll show you that in a second. So that's the first way of doing it. Then, the second way is if you have your data in Stata, or SAS or SPSS format, you can just use that single file and then just import your data using that single file. So let's start by creating a project. So we're going to, of course, have to do this on a node. So we'll use one of our test nodes for this purpose, we have a test node right here. And this is the first thing you see when you access your node. This is the main page, the home page or the node and you can see there's nothing here yet. So, all zeros, but we'll change that in the course of these videos. So again, this is the Catalogue. So this is the main page of the Catalogue and then if we click up here on applications, you can see the other three components that we mentioned. But now we're interested in the Data Repository. So I'll click on that.

Now I have to log in, I have my credentials here. So you would have an account and you would just type your own username and password, and sign in. Okay, so this is the homepage of the Data Repository. The first thing we'll need to do is to create a project, as I mentioned, so we can click up here on projects. You can see there's no projects yet. So let's add a project. Let's click on “Add project”. Now we have to give a name to the project.

So throughout these videos, we're going to be using as an example, one of the cohorts that participates on the RECAP Preterm project, which is EPICE-PT. So we'll use a lot of examples from that cohort. For this particular video, we'll be uploading data from that cohort. It’s not actual real data, it's just pseudo data that we generated for the purpose of this video. But yes, so the data is from EPICE-PT and also in the other videos, we're also going to be using examples from that same cohort. So, this is to say that, I'm going to use the name of that cohort for the name of this project, because that's the data that I'm going to be uploading, is from that cohort, so it makes sense that I just use the name of the cohort as the name of the project. Now, I can just use the same name for the title of the project and then we can just click “Save”. Okay, so this is our project, we're inside of the project now, I can just go back just to show you, if I click on ”Projects” it will appear the page where we were before, but now we have our project that we just created.

Importing Data Using a CSV Data File and a Dictionary [07:04]

So now that we have the project, we can upload data into it. So let's see how we do that. So let me just go back to the wiki for a second. So, we've done that, the first step was very easy. So let's go to the second step: importing data. So the first thing we're going to do is, this option of using a CSV data file plus an Excel dictionary data file. So I'm going to... if you're following along, if you're doing this yourself, you can click here as well, just as I'm going to do now. So this will download a file, a zip file and if you extract that file, let's see, I'm going to extract the file.

So now we have two files. We have a CSV data file and then an Excel dictionary, like I mentioned, and just so you see what I meant when I mentioned this is raw data, if you try to open it with just a regular text editor you'll see that it's just the data itself. There's no metadata. This is the actual data that's in here. So to complement that, we can use an Excel dictionary file. So, I'm just going to open the file, I'm going to show you what it looks like. But I'm not going to go into much detail about the format of the dictionary, because it has to be a specific format, that the platform is ready to interpret and knows how to interpret. So it has to be a specific format. So I'm going to show you what it looks like. But if you want to know in more detail, the structure of the dictionary file, there's another video from one of our partners at work package 3 and I'll put a link in the description. The video is called Data Mapping and Harmonisation. So if you want to know in more detail how these dictionaries work you can watch that video, but I can show you just briefly what it looks like. Here we go. So it has two sheets, so one for the variables and one for the categories. So the variables, we have a list of variables here we have just 12 variables, the cohort, EPICE-PT itself has many more variables, but we just selected a subset, just as an example. These are all perinatal variables. So this was the first data collection event from that cohort. We selected just a subset of 11 variables. You can see, we have the value types for each variable, integer and dates, decimal variables, but we also had units and the labels of variables. So this is what I mean, when I say metadata. The CSV file, like I showed you, has the actual data. But then this file complements that with this metadata. Then also for those variables that are categorical, such as, let's see, for example, this one, "a8", if you look at the label, it's "Sex of baby". So this is a categorical variable, and for those variables, we can find the categories in the second sheet. So if I click on that, you'll see, so there's the "a8" variable. But you can see it has four categories, 1, 2, 3, and 9, which correspond to male, female, undetermined and missing. So, this is how you describe categories of a variable and then the variables themselves are described here. Okay, so I'm going to close the dictionary now.

For this step here, we're using these two files. So let's go back to the Data Repository, and let's use those two files to create a data table here inside our project. Let's click on the project, we can see we have no tables currently, but if we click on this second tab, on the left hand side, “tables” tab. This is where we will see our tables, we don't have any at the moment, but let's add one. So using this first approach, you know, the CSV plus dictionary file, we first have to create the table itself just using the dictionary, there is no data yet just the dictionary. So we can click on “Add table”, then “add /update tables from dictionary”. So this is what we want to do, we want to use our Excel dictionary to create a table. Alright, so we'll need to select the file here, we can click on “Browse”, and now, right now I have the files on my computer, right. But I have to upload the files here first, so I can then use them. So I'll just click on “Upload” here and choose files. Okay, and now I have those files in the Downloads folder. Then inside this folder. Okay, here they are. So I can select them both. I'll just upload them both right now.

Click on “Open”, and then ”Upload”. Okay, so the two files are in the Data Repository now, but we haven't done anything with them yet. So they're there. But we haven't used them yet. So now, remember what I was doing... let me just go back here for a second. Let me cancel this. What we were doing was creating a table using a dictionary, so let's click on that. And now we have to select the dictionary. Let's click on “Browse”. Now we do have our dictionary here. So I'm going to select it. This is the Excel dictionary file, select that, file selected, then click “Next”. Okay, so now it gives us a summary of what the dictionary has, and like we saw, it has 11 variables. 11 new variables. So, that checks out. So, let's finish. Just wait a few seconds, and there's our table. Okay, so this is the table we just created using a dictionary file. It's called a “EPICE-PT_Perinatal” because as I mentioned, the data or the pseudo data that we're going to be using is from from variables that were collected at birth or not at birth, but you know, during the perinatal period, so you know, just right before or right after birth. So, there's 11 variables, and our entities, so entities will be the participants, so the participants in the cohort, so each line of our table will be an entity. And currently, there's no data in there. So there's no entities. We have, so the table is created, we can click on it, let’s click on it, see what it looks like and we can see, all of this information came from our dictionary, right? There's the names of the variables here, then the labels, the value types and then for those variables that are categorical, such as this one, we have the categories. Even for variables that are not categorical, but do have a specific code that represents missing values, we can just say that, for example, the age, this is the variable for age of discharge, which is a continuous variable. But in EPICE-PT, there's a specific code for missing

values, which is 9999. So we can specify that as a category for those missing values. Okay, so this is what the table looks like, but this is just metadata, right? There's no data, the way we would see the data is up here. So we have a lot of tabs here, right now we're looking at the dictionary, but we could click on values and we can see that there's no values yet.

So that's where the CSV comes in, the actual data. We have to upload our data now, into this table. The table is empty, but we're going to import the CSV into this table. So let me go back here. Right, so I'm inside the project and I can see our table right there. So now I'm going to click on “Import”, is this button here, the Import button. Here is where we select the data format. So like I said, first, we're going to use CSV. I should select CSV here and then I can just click “Next”. Okay, so now I have to select the CSV file, I can click on “Browse” and there's our CSV file that we uploaded earlier. We can select it. Okay, then the file is selected. Now we have to specify, by default, it just uses the name of the file . So here, it should be the name of the table into which we want to import this data and which is that table right there. Right, the “EPICE-PT_Perinatal” table. By default, it just uses the name of the file, of the CSV file. So we're going to remove that. As we start typing the name of the table that we actually want, it will suggest the table so we can just click on it. So that's the table where we want to put this data, the data that's on the CSV file.

Okay, so the file is selected and I want to put that data into this table. So that looks good. So now let's click on “Next”. Then there's some options here, that you would may need for, you know, in the future, for example, if you already have a table with data, and you're just adding more data to that table, you could select this, for example, “incremental import” means that there's data already there, and I'm just adding more data to that same table, but we don't need that right now. We're just starting so let's just leave it with all the default options and click “Next”. Okay, again, it gives us a summary of what's going to happen. So “unmodified variables” 11. So what this means is that... so remember that before this, we created a table just using the Excel dictionary, which had 11 variables described there. Now you're taking the CSV and importing the data in the CSV into this table that we created. So, what the Data Repository does is match the names of the variables that were in the dictionary with the names of the variables that are on the CSV, and because we have 11 “unmodified variables”, that means that it was able to match all of those 11 variables. Let's say you had an extra variable on the CSV that you did not have on the data dictionary. So, that would be 11 here, and then just one one new variable. So that would mean that there's one variable on the CSV that is not in the dictionary. So, this means that there were 11 matches to the 11 variables.

So that looks good. So let's just click “Next”. Now we have a preview of what the data is just, this is just a subset of what the table will look like, once the data is imported in there, so that looks okay, it looks good. So we have our IDs. Then, you know, all the other variables, you can scroll. So that looks good. I can click on “Finish”. Now the data is being imported into the table, we can check the progress of the import by clicking on this icon here. This is the “Tasks” tab, you can see all the tests that are currently happening or have happened in the past. So, currently, we only have one test, which was the import we just started. We can actually see that it just finished and it succeeded. There's a green dot right there. So our import succeeded. So let's go back to the tables, which is the second tab on the left. There's our table. And now we can see, there's 50 entities, which is the number of lines on our CSV file. So we just imported 50 lines into this table. Now if we click on it, and then try to see the values, I'm already there, the table is selected from before. Now we can see the values. So this is the data that was in the CSV, and now that we've imported that data into this table. There it is so. That's definitely as easy as that. So remember, this way we used the dictionary to create a table, an empty table, so just a description of the variables that are going to be there and then we used the raw data file, CSV file, and imported that data into the table. So that's one way of doing it, of importing data into a node. The other way is just use a format such as Stata, or SAS or SPSS, like we saw here, let me go back here. So we just did that. We just did 2.1. So now let's try 2.2 using a STATA, SAS or SPSS data file.

Importing Data Using a STATA, SAS or SPSS Data File [23:48]

Again, we have an example here, an example dataset. In our case, we're going to use SPSS but we could use any of the other formats. So I'll download that. So again, this is an SPSS file, which has the extension “.sav”. Save it. Let me just check here on downloads. There it is. There's my file. Again, this is also pseudo data. This is exactly the same data that was on the CSV, but it's just in a different format. This is a single file, right? So in this file, we only need a single file, because this is not a raw data format. It has the data but also has all the metadata associated with those data, so we only need a single file. So let's use that file. Let's try to import that file, let me go back to the Data Repository. Now there's our table that we've created before with a CSV plus the Excel dictionary file. So let's remove that because I'm going to create another one using a different approach. Okay, so the table is gone. Now I'm going to import the SPSS data. So now, whereas before, we had to create a table using a dictionary and then import the CSV into it. Now we don't have a dictionary, we do have a dictionary, but it's incorporated into the data file. So we only need to do the import. Just

select instead of CSV here, we'll just select SPSS. You can see there's other formats here. Let's just use SPSS because that's the format we have, then click “Next”.

Now we have to select our data file. Oh, and I haven't uploaded the file yet. But let's do that. Now let's upload our SPSS file. Let's go to downloads. There it is. So that's our SPSS file, we're going to upload that into the repository. So there's the file. Again, when I say upload, I'm just putting the file in the repository. But we haven't done anything with it yet. We haven't created a table with it. So that's what we're going to do now. So I'll select that file. Okay, so the data file is selected, we can just click “Next”. There's an option here to specify which column has the IDs. If we don't specify, if we just leave it empty, it will assume that the first column of the file is the “IDS” column, which it usually is, and in this case it is. So I'll just leave it empty and just click “Next”. Again, we also have the same options we had before, but we don't need that right now. So let's just click “Next”. So it's reading the file, the SPSS file. There we go. So again, it read, it found 11 variables. So let's click “Next”. Again, it also gives us a preview of what the table will look like. Click “Finish”. And now the import has started. We can again, we can check here on the ”Tasks” tab we can go here. Okay, so now we have our latest tasks, which was an import that we just started and we can see that it also just succeeded. So let's go back to tables. And see that there's our table right there. 11 variables and 50 entities and we can click on it. There we go, there's the dictionary. So remember, we only used a single file. Because it's not a raw data format it also has all these metadata in there. So it has the data and also all of this information. So if you click on “Values”,you can see the values as well. So this is just a different way of importing data into a Data Repository.

So that is the end of this video. Let me just go back here to the home page of the wiki. We just covered this first part, setting up data in the Data Repository, we learned how to create a project and how to import data into it. In the next video, we're going to see how to add metadata using the Study Manager, which is the next layer, if you will, on a RECAP node. So you'll have the link on the description to the next video.