3. Adding Metadata in the Study Manager - Transcript

Introduction [00:00]

Welcome to the third video about the RECAP Preterm data platform. In this video, I will be showing you how to work with another one of the four main components of a RECAP node, which is the Study Manager.

Now, some of the things I will be showing you in this video are related to previous videos. So if you have not yet watched all of our previous videos, please go back and watch them first, and then you can come back to this one.

And also, as I mentioned before, you can find the link to this wiki, the wiki that I'm using, you can find a link to it in the description of the video.

Creating and Describing a Study [00:38]

So this time, we're going to look at the Study Manager, which is one layer above the Data Repository that we saw in the first video. Recall that on the first video, we created a project and then imported some data into it. Now we're going up one layer to the Study Manager, where we're going to describe a study, create a study and describe the study and its populations and collection events. And basically all the information that is relevant, that would be relevant to a study. But that's all of that information is not data, right? The actual data is on the Data Repository, like we saw on the previous video. Now, we're just adding more information that relates to that data. So we're going to create a study, add a population and a collection events to it, describe all of those things, and then link all of that information to the data to which it relates. So let's click on this section right here. Let's start.

Okay, so just to explain a little bit how the Study Manager works. Let's look first at the right hand side here, we have the Data Repository. And again, this is what we did on the first video, it was all on the Data Repository, we created a project and then added a table with data to that project. Now, in this video, we're going to be on the left hand side here on Study Manager. And as you can see, there's several relevant entities in the said Study Manager. But for this video, we're just going to focus on these four right here. Because there's also some other entities related to harmonisation, but we're going to see that on the next video.

So the first thing we're going to do is create a study. And remember that we are using, as an example, the EPICE-PT cohort, which is one of the cohortes that is collaborating in the RECAP Preterm project. So we're going to create

a study for EPICE-PT, describe the cohort and then we can add multiple populations, we're just going to add one, just as an example, describe the population and then we can create multiple data collection events, and associate those with a particular population. The collection events would be the waves of data collection that cohorts do. So we're going to add just one as an example, again, one collection event for EPICE-PT. And once we have these three entities described on the Study Manager, we have to link that information to the data, the actual data that we have on the Data Repository. And the way we do that is using a collected data set. Now the name is a little misleading, because this is not an actual data set. Because remember, the data is in the Data Repository, not here. This is all metadata, just descriptive information. So this collected data set is not a real data set, it's just a link. It's just the entity that links all of this information to the data, the actual data in the Data Repository.

Now, if you remember, we are using, for these videos, one of our test nodes. I have it open right here. So this is, again, this is the first page, the homepage of the Catalogue, which is the uppermost layer of a node. And here you can see, we have no studies, no data sets. And this is what this is what we're going to change. By the end of this video. We'll have something here, okay. But for now, let's just go into applications here. And then select the study management, right previous video was Data Repository. Now we're going to study management. So I'm going to sign in using my own credentials. Okay, Now, we're in. Now remember, the first thing we wanted to do was create an individual study. Right? If you remember back here, like I said, we want to create a study, add a population to it, add a collection event to it, and then link all of that to the data in the Data Repository using a collected data set. So let's start by creating an individual study. Let's go back here to the Study Manager. Click on studies, up here, then individual studies. As you can see, we have no studies yet. So let's add one, click on Add study. And we get this long form. And we have to fill this in. Not all fields are required, of course, but you can see what I meant by metadata. It's things like, number of participants and other relevant information about the participants, or published papers associated with this particular study. Start Year, EndYear, funding, website, objectives of the study, things like that.

So we're going to need some information to use here. So I'm going to use, again, as an example. So I have a file here. If you're following along, you can also download this file right here. And this link, just right above the before the first section, click on that and download this file. This is a PDF file. I already have it. So I'm going to just cancel that and go to downloads. Yeah, so that's the PDF file to open it. Okay, here we go. So we have all of this information, we can use it to fill in that form. You can see we have three

sections in this document, we have the study information here. And then in the middle, we have information about the population that we can use. Then also, we have information about one of the collection events, we're going to use the perinatal assessment, which was the first collection events of the cohorts of EPICE-PT. So I'm going to use this, I'm going to, let's see this, let's put this side by side here. So basically, I'm just going to copy all of this information that we have here on this document to our form here. So let's see the name of the cohort is “Effective Intensive Perinatal Care in Europe”. That's EPICE-PT. The acronym is EPICE-PT. Oh, we also have a logo here, we can download the logos. Because you can see, I can upload a file here that will serve as the logo for this study. So let's do that. Let's download this file. Okay, you can see it's EPICE underscore logo. I'll save it in my downloads. Just open. See what it looks like. So that's the EPICE logo. Okay, so let's just upload that file to, there it is, EPICE logo.

Okay, so I uploaded the logo. Let's keep going, objectives. I have the objectives here, let’s copy that. What else do we have? We have the start year, we have some funding information there. We also have, we don't have this. Let's see, we have the number of participants here, 74. This is a cohort so select study design, this is a cohort. Marker paper, we have one here, just copy that paste it here, marker paper. By the way, some of these fields that you see here, these are specific to the project, to RECAP Preterm. Because you can see, for example here, number of term born control births included in cohort. So this is specific to RECAP, but this is all adaptable. I can adapt this to some other contexts and just add or remove fields here. Okay, so that's all the information that we have for the study. Because right here, we start the population information. So we haven't gotten that far yet. So let's stop here. For now. Let's just click Save. And this will create the study for us. So it's creating the study right now. Okay. So I think that's it. There we go. There's our logo, and all of that information is saved.

Describing Study Populations and Data Collection Events [11:14]

Now we have our individual study. Now, what was next? Let's see. So we created an individual study, now we have to create a population. So let's go back here. If we scroll down, actually going to just show you, if I go back to the individual studies page, we should see our study now, EPICE-PT, there it is. Okay, so now we want to add a population to it. So we can click on it. And it loads all the information that it already has, that we gave it. And then we can also add members to the study, you know, contacts, or investigators. Let's just add a population. You can see here, this population section, let's add a population. Okay, so this is a little shorter, but we can describe our population here. The ID, it can be whatever you want, it just has to be

different from any other population that already exists. But we don't have any populations yet. So let's just use like, 1 as the ID. And the name, we can call it. Let's just call it “EPICE-PT pop”, for example. Okay, and then the description, we have a description here that we can use, let’s put that there. Countries. So this is Portugal. EPICE-PT is Portugal, here we go. Geographical area, we have, northern region, Lisbon and Tagus Valley region. Now we have some inclusion and exclusion criteria here. I think we only have index group inclusion criteria, so let's copy this. And that would be this field. Okay, and that's it for the population. So let's save it.

Okay, so again, we have all the information about the study. And then down here, we have the population section, we can add more populations, but let's just stick with this one for now. We have the population that we just added in “EPICE-PT pop” with all of that information. And now let's go back here. So we created a study, added a population. Now let's add a data collection events. Let's go back here. Now, you see, right at the end of the page, we'll see the data collection events section. So let's add one, let's add the data collection event. Okay. Again, the ID can be whatever we want, let’s just use 1 again. Let's see, what information do we have here? Right, so we're going to use this collection event. The perinatal assessment, which is the first collection event of the cohort, duration 2011 to 2012. So let's use that. Started in 2011 and ended in 2012, the study did not end in 2012, right? This is just this is related to this particular collection event that took place between 2011 and 2012. Okay, now we need a description for that collection event.

We have all sorts of things that we could use here. Let's see, yes, we can select whether this collection event is the perinatal collection event, which is the first one, or if it's a follow up, in this case, it is perinatal collection events. So if I click that, I'll get more options here that are options that are specific to the collection event that is perinatal collection events. I don't have this information, I don't think. We do, actually we do. It's these numbers right here. Let's see. “Index group numbers survived”. Yeah, so that would be there... “Invited”... Oh, actually we don't have this information. What we do have is here. So “Survived”, “Invited”... Same number, okay. It's all the same number. And again, we can add more information to it. But you know, it doesn't have to be complete for the purpose of just demonstration, but just so you can see what different kinds of information we can put here. And also, we can add, like I said, we can add, however many more fields that we feel are necessary. So let's save that. Save the collection event. And now that we have the collection event, scroll down, you can see a timeline here. Now, this is a very simple timeline, of course, because we only have one collection event. But if you add multiple collection events, and they have start times, end times, we would

get a nice graph here, a nice timeline of all the collection events. And there it is, there's our collection event right there.

Linking a Study to Data in a Data Repository [17:44]

Okay. So now that we have this... so we created the study, population and collection event. Now, this is, again, all of this information that we just created, is in the Study Manager. But it's just like it's isolated, right? It's not linked to that data that we put on the Data Repository on the second video. So what we need to do now is the final piece, which is to create a collected data set that will serve as a link between the data and all of this information that we just created. So we have to create a collected data set. Let's do that. Let's go back here. Let's go to the homepage. Actually, I can close this. We don't need this anymore. Okay, just click up here on datasets. And then collected data sets. Okay, so we need to create a collected data set. So let's click on add data sets. And this we can call it whatever we want. But this, remember that this is going to be the link between a particular collection event and a table in the Data Repository.

Now the collection event that we created was called "Perinatal assessment". So we should call it something similar to that because that's what it's going to be linked to. So let's call it "EPICE-PT Perinatal Dataset", or something like that for the name, and then the acronym can just use the same name. But you could also add a description but I'm just going to save it as it is. Okay, so the dataset is created. And again, it's misleading to call it a dataset. This is a link, this is this part right here, that's going to link this to that. But we've created the collected dataset, but it's not linked to either one of these yet. It's just isolated. At the moment. So the way we link those two is by adding a table. So we can click here. And now we can say to which study, population, and collection event. So to which combination of those three does this collected dataset relate. And we only have one study so it's related to the EPICE-PT study. And now we have to select the population. Again, we also only have one, select that. And then we have selected collection events, which is the perinatal assessment. Click on that. Okay, so now, this part right here, study, population and collection events is this part, right? So now we're linking this to these three. Now, the only thing missing is this part right here. Is linking this to this. The way we do that is down here, "Data source". So this is where we point to the data. So the Study Manager has a connection to the Data Repository. So we can now select a project from the Data Repository. EPICE-PT, and the table "EPICE-PT_Perinatal", again, remember that this project and this table we created in the previous video. So if you haven't watched that, you should watch that first. So basically, once I click save, that connection will be made, the

connection between the study and population and collection event. Between those, and the actual data will exist after I click save. So let's click Save. Okay, so this is done. All of this path here, starting in the study, population, collection event, collected dataset, table and project. All of this is one path now, all of this is linked.

Publishing a Study [22:46]

So, “to what end?”, you may be thinking. Well, the purpose of the Data Repository... So remember, this is an important thing to understand is that these are the different layers, right? Like I explained in the first video, which is the Data Repository where the actual data resides. And then the Study Manager where we describe the study and then the Catalogue. The Catalogue is the only public part of the node and is the thing you see when you first access the node. I'm just going to access our test node here. This is the node we are using. So when you access the node, the first thing you see is the landing page of the Catalogue, which is the public part of the Catalogue and I, anyone can access that page, you don't need to sign in to see the Catalogue. But you do need to sign in to see the Study Manager here. So, you can see here, we have studies and datasets. All of this we just created. But if you look on the Catalogue, there's nothing here yet. And that's because those things that we create on the Study Manager will only show up on the Catalogue, once you publish them. It has to be an explicit act of publishing on our part.

Let's go back here and select our study. So you can see there's our study, and there's this column here: “Published”. And it's not published. If it was then that little star will be filled in. So it's not published. Once we do publish it, it will show up here on the Catalogue, and it will be public. All of that information, all of that metadata that we put in there will be visible here on the Catalogue. So let's do that. Let's publish the study. We can click on it. And it will take us to this page where we were before and now to publish it, we can click up here on ”Draft”. Then “To under review”, this is because there's a workflow here. You can create it and then put it in a draft form and then put it under review. Because there's different roles that I'm not going to go into that much detail here. Just not to confuse you, because it's a lot of information. But there's different roles that users can have here. And some users can be reviewers, for example, and those reviewers would be able to just... any study that is put under review, would have to be approved by someone with the role of reviewer, for example. So this is the so we have before we publish it, we have to... it has to go through all those steps. So right now, once you create a study, it's automatically in draft form. And then you have to put it under review, which is the next stage. And then, only

then, can you publish it. So now, you see there's a new button here. So I'm going to click on Publish. Okay, it's published.

Now, if we go back here to the Catalogue, again, this is the Study Manager, everything we publish here will show up on the Catalogue, the public part of the node. So let's go to the Catalogue. And now if we refresh, we should see. Let's see, what should we see here, we should see one study, which is EPICE-PT, and we should see one data set, let's refresh the page. Okay, there we go. There's one study, which is EPICE-PT, that we just published. And there's no datasets, of course, because we haven't yet published the collected data set, right? We created the study. So this part right here is the study, the population and the collection event. Right? So what we just published was these three entities here, we just published those.

Now, if we publish this, what does it mean to publish this? Remember that this is a link to the data. So by publishing this, does that mean that I'm publishing the data? No, of course not. Because the only thing that this does... once I publish this, the collected data set, there will be a “1” here. And it will not retrieve the data from the repository. Instead, what will happen is it will just pull from that table, it will just pull metadata. So it will pull the dictionary of the table. You know, the list of variables and their types and categories and things like that. And also, summary statistics. So just aggregated data, no real data ever comes out via this link. Let's see what that looks like. Let's go here. So we have to publish. Again, we have to publish this part here, in order to get the metadata from this table. So the dictionary and summary statistics, to get those on the Catalogue, we have to publish this. So this link... so the Study Manager can retrieve that information through this link. And then, because this will be published, that information will show up on the uppermost layer of the node, which is the Catalogue. So let's publish the collected dataset. Let's go here to “Datasets”, “Collected datasets”. There's the one we created. And as you can see, it's not published yet. Let's click on it and do the same thing we did before with the study. Click on “Draft”, “Under review”, now there's the publish button. Let's click on it. And it's published. Okay, so let's go back to the homepage of the Catalogue.

Now, once I refresh, we should see one dataset, It's the one we just published and along with that data set, because that dataset represents a link to the Data Repository, along with that it will retrieve from the Data Repository, the information about the dictionary for the table. So we will actually see a different number of variables, I think it was 11 that we put In the Data Repository on the previous video, so let's hit refresh. And there you go. There's one dataset, which is the link to the Data Repository, and that's why we also see 11 variables there. So let's just see what it looks like here

on the Catalogue. Because there's, it's a different display of the information. So this page will list all the studies that are on this node. Right now, there's only one, this one that we created. EPICE-PT, there's a little summary here. Datasets and variables.

I didn't mention networks, but basically networks are... if you have multiple studies on a node and those studies are related, in some fashion, you can group studies, you can create a group of studies, and that's called a network. So for example, on the central RECAP Catalogue, we have all the cohorts described there, and then created a network called RECAP Preterm, because all of those cohorts are related to each other because they all collaborate in the same project, which is RECAP Preterm. So that would be a network in that scenario.

So let's keep going. So there's our study, you can click here on “Read more”. And basically, we will see all the information that we put on the Study Manager. Right, so the description of the study, all that information that I copied from the document, information about the population, and then the collection event, which is just one in this case. So that's it. And now, also the variables. So 11 variables. So if I click on that, let's see what kind of information we get about the variables. So that took us to the search page, we can, like I showed you on the first video, we can use the search page to search for variables. And we see we have 11 variables here. These are the ones that we imported during the previous video. So all of this information, remember, when I opened the Excel dictionary file from the previous video, all of this information was there, right? So this information is in the Data Repository. And it went to the Study Manager through that link that we created. And because we published that link, now it shows up on the uppermost layer, which is the Catalogue. Let's select, let's see, I can select one of these variables, for example. Let's see this one, sex of baby. Click on it, we should be able to see some summary statistics. Okay, there we go. So we had, remember, we had 50 rows that we imported. So apparently, for this variable, we have three missing values. So we have only 47 out of the 50 have actual values. So we can see a summary of the data here, we get a nice graph there.

Closing Remarks [34:07]

So that's basically it for the Study Manager. Basically, what we did, just in summary: we created a study in the Study Manager, we added a population to it, we added a data collection event to it, we linked all of that information to the data in the Data Repository, then we published all of that, all of that, not the actual data. But you know, we published that and all of

that information shows up here on the Catalogue. And you can browse through it and search for variables. And you even get nice summary statistics, aggregated data retrieved from the actual data. And again, the actual data does not leave the Data Repository. Or at least not this way. There's other ways, we'll explore that in the later videos. But that's it for this one. On the next video. We'll see how to... the next part is about harmonisation. So, we'll proceed to Part C: “Harmonising Data Across Multiple Studies”.