1. Overview of the RECAP Preterm Data Platform - Transcript

Introduction [00:00]

Hello, and welcome to the first video of the WP4 module for the RECAP Preterm Summer School. My name is Gonçalo and in the course of the series of videos, in this module, I will be talking to you about the RECAP Preterm data platform. So the next five videos are actual hands-on demonstrations of how to do things on the platform. But in this first video, I will just do a presentation with a brief overview of the RECAP platform. So you can become a bit more familiar with what it is and how it works and also just to provide a bit of context for what's to come on the following videos. So let's start.

What is the RECAP Preterm Data Platform? [00:38]

So what is the RECAP Preterm data platform? Well, first of all, it was developed, as I'm sure you know, in the context of the RECAP Preterm project, which is a EU funded project under the Horizon 2020 programme and the aim of the platform is to bring together population-based cohorts of children born very preterm, or with very low birth weights from 13 different European countries. And also along with cohorts, National Health Data registers from four of the Nordic countries.

And then, also some of the main features of the platform is... it facilitates the harmonisation of data hosted across multiple sites or nodes, as we call them. So each partner has a node as we'll see in a minute. So the data is distributed and the platform allows for harmonisation across different nodes, physically distant, nodes. And also, perhaps more importantly, the platform also allows for distributed and non-disclosive statistical analysis. So, particularly this last point about analysis and also harmonisation, we have specific videos for this. Harmonisation is our fourth video. And then analysis is our last video, which is the sixth video of the WP4 module. So we'll talk about this in more detail on those videos. This is just an overview. But those are basically the main features that the platform provides.

Architecture of the RECAP Preterm Data Platform [02:30]

So let's look now a bit at the architecture of the RECAP data platform. So how is it built? And how does it work? Well, the first point is that it is a federated infrastructure that comprises 13 nodes hosted across Europe. As you can see here on the map, we have 13 nodes. So the nodes are basically

just servers where you can put data, and also do lots of other things that we'll talk about in a minute and also on the following videos. But basically, the data is distributed across these nodes. And you'll see the second point here is that the nodes host data from their respective cohorts, of course. That's the whole point is that the data is hosted within the partners premises, so the data is hosted there, and it doesn't have to leave in order for you to use it for your analysis. But we'll see how that works in a minute.

And also, among these 13 nodes, there's a central node, which is here in Portugal, if we zoom in a bit here, you can see that there are three nodes here. If we zoom in, here are three nodes. And this one right here, this is the central node. Now, you may be thinking, how do you have a central node and then also say that it's a federated infrastructure? Well actually, having a central entity in an infrastructure doesn't make it centralised. It would be centralised, if the other nodes were not able to talk to each other. But because they are and they can also communicate with the central node, that's a federated infrastructure. That's what makes it federated.

And so how do the other nodes communicate with the central node and what's the central node used for? So the other nodes share data dictionaries and summary statistics with the central node. So they send it, or rather, the central node is able to gather, retrieve data dictionaries and summary statistics with specific permissions from the nodes involved, of course, it's able to retrieve the information and assemble all of that information into what we call the central RECAP Catalogue. So all of that information, data dictionaries from all the nodes, all the cohorts, all the variables, and then also summary statistics for those variables, are sent to the central node by all the other nodes. And then, the central node uses that information to build the central RECAP Catalogue.

We have an illustration here, so, like I said, the other nodes all send that information. Remember that information, when I say that information, I don't mean data, the data is on the nodes. And it doesn't leave, unless the owner of the data wants the data to leave. But what happens here is that the central node is able to retrieve only data dictionaries and summary statistics and then it uses that information here to build what we call, again, the RECAP Preterm Catalogue. You can see a snapshot of one of the pages of the Catalogue (there’s many pages), and you can see each of these squares here are cohorts that we've described and listed here on the Catalogue. For each cohort, you have a description of the cohort of the populations of their cohort and data collection events and then you have dictionaries of the datasets that were collected within the context of that cohort. You could click, you could search for variables, you can click on the variables and see summary statistics for those variables and all of that

information comes from the nodes, the distributed nodes, and then they send the information to the central node to build the RECAP Preterm Catalogue. This is sort of a zoomed out overview of what happens on the platform.

Components of a RECAP Node [07:08]

But let's now zoom in a little bit in one of these nodes. They're all structurally the same, so it doesn't matter, let's just zoom in on a RECAP node and see what it looks like inside. So there are four main components of a RECAP node. There's the Authentication Server, the Data Repository, the Study Manager, and the Catalogue. So all of those nodes that you saw on the map back there, all of them have these four components. Now let's see, briefly, what each of these components is used for.

So the first one, the Authentication Server, and this is just where you can go to manage your users, your local users. It's used for user authentication, managing user accounts on your node, and then also email notifications. Because you can have... users can register on your node. If you allow that, if you allow registration requests, then there's email notifications involved with that and you can either approve or reject new registrations, or you can create accounts for yourself or for other users. Basically that's the role of this Authentication Server. It's just to manage users on your node.

The next layer, if you will, on a RECAP node is the Data Repository, which is sort of the main, I would say the main component, or at least the probably the most important component, because it's where the actual data is hosted. So in the Data Repository, you can import and export data there in multiple formats: CSV, SPSS, STATA or SAS. Then it's also in the Data Repository, that you harmonise the data. So the harmonisation on the platform works a little bit differently from what you may be used to. And we'll cover that in one of the following videos. But this is where the harmonisation takes place on the platform, is inside the Data Repository. And then the other thing is data access management. So you'll have your data on your Data Repository and then again, remember in the previous component, you can create user accounts. So then, you may have, let's say, you have multiple datasets on your Data Repository, you can grant access to a particular user to a specific dataset, and not only can you can you give access to a specific dataset, you can also decide the level of access that user will have to that dataset. Because it's not a binary thing, like either the user has or does not have access to that dataset. If the user has access to the dataset, you can decide which level of access. So there are multiple levels, we'll see that on the next videos, but mainly, you can either grant

access to the actual data (individual level data), or you can just grant access to aggregated data. So just summary statistics and dictionaries of those datasets. So that's data access management. Then finally, data analysis. So this is where the analysis takes place. It's also in the Data Repository and again, we'll do one of our videos, it's the last video, it's specifically about data analysis. So we'll go into that further on that video.

The next layer is the Study Manager. So, say you have data on your Data Repository, that data may have some context, right? They may have. So the variables, it has a dictionary. So it may have variable labels, or types, or categories. So there's some context, but how was that data collected and in what context and for what purpose? So that's why we have this layer here. So this layer allows you to have structured descriptions of the studies and populations and data collection events. So you can describe, let's say, a cohort, you can describe the cohort here in the Study Manager, and say, we have these populations, and we made these collection events and those collection events resulted in... and then you would point to the data in the Data Repository. So this allows you to provide context for the data that's there, it just further describes the data and the context in which those data were collected. Also, you can manage data access requests. So the Study Manager also allows users to make data access requests, and then you, as an administrator here on the Study Manager, can either approve or reject those data access requests and also... so you describe your cohort, let's say, here on the Study Manager, but that is sort of hidden behind authentication, because the Study Manager, the Data Repository, and the Authentication Server, all of those components require you to log in, in order to access them.

But then there's this part, the Catalogue, which is the only public part of the node. When you go to a node, the homepage of the node will be the Catalogue. That's the first thing you'll see when you access a node, and the Catalogue is, as I said, it's the public part of the node, and the information you will see there is whatever information is in the Study Manager that has been published. So you can have all of that information that I just mentioned, in the Study Manager, and that information will only show up here on the Catalogue. So it will only be public, if you actively publish it on the Study Manager. So once you're happy with it, you can publish that and then it will show up on the Catalogue. So you can see here, it's a public Catalogue containing content that has been published in the Study Manager. So that will allow you to browse the studies, datasets, variables, variable summary statistics, so all of that information becomes available on the Catalogue once you publish it on the Study Manager.

So, let's now look at how these four components interact with each other. So hopefully, this will make sense after my description of each of the components. This is, let's say, the base component. It's the Data Repository where the actual data is. And then you have the next layer, if you will, which is the Study Manager where you describe those data. So there's a link there because the descriptions you have here can be linked to data that is hosted here. Then whatever you publish here will go up to the Catalogue, which is the public part of the node. So when you publish something here, it will become public in the Catalogue. And sort of parallel to all of this, there's the Authentication Server that, when you want to log in, in any of these components, you can use your account. And when you try to log in, it will just authenticate you in the Authentication Server. So that's just a background thing that happens when you log into any of these components. What the platform is doing is authenticating you and making sure that account exists, and that the password is correct, using the Authentication Server.

So all of this, all of these interactions of these four components exists inside a node, a RECAP node, because remember, we zoomed in on one of those nodes, doesn't matter, they all have this. So remember, when we looked at the map, we had 13, nodes, and all of them, all of those 13 nodes have these four components. So all of this that we just saw, is happening in each of these 13 nodes. And so the information, all the information, like I said in the beginning, that exists on these nodes, that information is sent to the central node, which is hosted in Portugal. And that's how we are able to have the central Catalogue, which is sort of a display of what we have, data-wise on the RECAP project.

What’s Next? [17:05]

And so that is basically it for the overview of the platform. So what's next, what am I going to talk about in the next five videos? So the next video will be about setting up data in the Data Repository. So basically, we’ll do, it's not a direct mapping, but we'll do basically one video for each of those components. And so the first video will be how to use the Data Repository, how to put data in there. And then the second video, the third video rather, will be how to describe a cohort, or study in a more general sense, a study in the Study Manager, and how to link that information that metadata to the data, the actual data that you have on the Data Repository. And then we'll look at how harmonisation works on the platform. So I think I mentioned in the beginning that harmonisation works in a distributed way, so each node does its own harmonisation of the same variables, then it all comes together on the central node. We'll be able to see on the central node, the status of

the harmonisation across all the cohorts, the participating cohorts. And then we'll see how to manage users and permissions, how you can create users on your node, and how you can grant those users access to some specific data set on your Data Repository.

Then finally, we'll look at how analysis works on the platform. Because there's basically two ways you can work. The first way is, if you have been granted access to individual level data, then you can just do whatever you can just use whatever statistical analysis programme you like. Because if you have access to the data, you can just pull the data, all of the data together, then just do your analysis as you normally would. But then there's also the most interesting part of this list, which is the remote and distributed, non-disclosive analysis using DataSHIELD, which is something we'll talk about on our last video. So basically, that will allow you to still perform analysis using distributed data without actually having access to the individual level data. So we'll see how that works on our last video.

And in case you have any questions, please feel free to contact us with this email address (recap-dev@inesctec.pt). And with that, I thank you and I hope to see you on the next video.