Executive Summary

To meet the immediate data management needs of the vertical research topics, the Alzheimer's disease (AD) topic in particular, the consortium will develop a first prototype of the "Private Remote Research Environments" (PRREs). This PRRE prototype is based on tranSMART, an open source web platform for translational research. The concept driving the work over the last year (2013) is described in the Description of Work (Annex 1 of the Grant Agreement):

"The fourth design decision [of the Platform] is to create the possibility that individual data sources may want to make, in the context of a specific research question, additional data available for analysis, but need to restrict access to that data to predefined users and/or for specific uses. In the EMIF- Platform, The consortium introduces the notion of the Private Remote Research Environments (PRREs) to enable such possibility. A prime example of this situation is the need of the current EMIF Research Topics, which heavily rely on specific cohorts for their investigations – therefore, detailed data needs to be available, but with strict access measures limited to the research groups involved. This enables the EMIF-Platform to prototype PRREs in the early stages of the project."

Although the initial work in tranSMART focused on the needs and requirements of the AD vertical, the learnings and solutions are applicable to the Metabolic vertical as well.

In the first 2 cycles (first 12 months) of the EMIF project, the team has focused on the "orange track" as described in the Description of work and its first deliverable: "PPREs – data management solution for verticals". Hence, the team has worked on the implementation of tranSMART primarily through 5 work-streams achieving the data management solution.

- Developing a taxonomy – per definition, a taxonomy is a structural framework for organizing information and represents knowledge as a set of concepts within a domain, using a shared vocabulary to denote the types, properties and interrelationships of those concepts. For tranSMART and the AD data cohorts, the taxonomy focuses on the structure and definitions of the variables coming from different sources. The same principle and approach of developing an Alzheimer's disease specific taxonomy can be applied to other disease areas.
- tranSMART – This workstream has two components. a. Loading the data into tranSMART and b. Implementation of the tranSMART system on a central server.
- Security features – Security and related workflows is a critical aspect of the Private Remote Research Environment (PRRE). On top, data owners, during interviews, have expressed their preference of having full control over who will get access to what data. The actual process of approving accounts and of requesting, granting and revoking data access is role based and manual. In the current implementation, user authentication is quite elaborate and is based on the Custodix Identity Access Management solution.

- Data de-identification – Data cohorts can contain data privacy sensitive data. The de-identification of the data can take place at the source before the data is submitted to be loaded into tranSMART or can take place at Custodix, before the data is made available to tranSMART data curators. In both cases, a control on data de-identification will be performed at Custodix based on a number of approaches and algorithms described in this document.
- Additional requirements – Version 1.1 of the tranSMART open source web platform has a number of functionalities. From the onset of the EMIF program, we already knew that cross-cohort data harmonization functionalities would be a much required feature. The team conducted a number of interviews with potential users of the system and drew up a set of requirements based on the current practices of the data owners, epidemiologists and bioinformaticians.

Contacts

EMIF-Platform: Johan van der Lei – Nigel Hughes
j.vanderlei@erasmusmc.nl - nhughes@its.jnj.com