“EUCAN-connect meta consortium: federating the data federations”

“My dream with EUCAN-connect is to create a long-living platform for large and valuable research datasets, and methods and tools to analyse them, accelerating multi-center cohort studies to improve citizens' lives.” Two years into the EUCAN-connect project, its principal investigator Professor Morris Swertz from the University Medical Center Groningen and MOLGENIS open-source software project evaluates progress towards realizing this dream.

Motivation

In the last decade we have been part of many NL and EU consortia and experienced many great achievements in terms of data generation, integration, and analysis towards promises of personalized approaches to improve health. I think we are fortunate to live in regions with 1000s of biobanks, cohorts and other data banks that have been collecting data for decades. Also, we are in an era with many novel large-scale profiling methods, measuring genomics and environmental exposures on an unprecedented scale. This all provides a wealth of data that often enables us to go “back to the future” and identify the subtle factors from the past that, when combined, do impact the development and health of us and our children today.

But I have also witnessed that, even with the best intentions, impressive data and methods made available in these projects tend to fade away when projects have ended - and when the next project starts, much of the work starts from scratch again. We found ourselves in such a situation a few years ago, when we became responsible for the data infrastructure for the EU LifeCycle project. What was special about the LifeCycle project was that we planned to use a very ambitious approach, that is, to implement a ‘federated analysis’. In this approach, no personal data is shared centrally but instead the analysis is sent to the data sources to be executed behind lock and key, and only summary-level data is exchanged.

While today there are more projects working to implement the federated approach, I had the good fortune to be a young postdoc in one of the first successful implementations of this approach, in the EU BioSHaRE project as far back as ten years ago - a project which was very timely in preparing for today's GDPR legislation. However, I quickly learned that 10 years later, the methods we developed are by no means ‘turnkey’ yet. Of course, we all know that research projects nowadays should minimally allocate 10% of their budget for ‘data infrastructure’. However, that is only sufficient for operational support, not for innovation. We found ourselves at a crossroads: either we should give up, or we should dramatically raise our ambitions and not focus on individual projects but try to synergize the larger community of cohort analysis projects. Obviously, we chose the latter.

Meta-consortium

The result is EUCAN-connect, in fact a ‘meta’ consortium that joins the efforts of large existing consortia working on federated analysis in parallel. Now we are two years into the project and we have made great progress. For each of the main ‘pillars’ that federated analysis projects depend on, we have at least three independent teams working together, i.e., for interfacing and connecting data catalogues to make full metadata findable and accessible on all connected cohorts; harmonisation methods to bridge the data heterogeneity between cohorts and to enable pooled analysis; operational tools for cohorts creating the local access nodes; the analysis toolbox for federated analysis; and ELSI issues surrounding all this. (Readers might recognize the FAIR principles in these pillars, i.e. EUCAN-connect aims to deliver on FAIR cohorts, i.e. findability, accessibility, interoperability and reusability).

What is special about the approach of EUCAN-connect is that we do not aim to build one software tool or system for federated analysis. Instead, we are defining interoperable interfaces that allow different software developing groups to integrate their tools. This enables healthy innovative competition where alternative solutions can be developed and the best solutions for each analysis can emerge. For example, DataSHIELD is the protocol we use for implementing the federated analysis. Now we have three independently developed implementations of DataSHIELD 'servers' that prove that DataSHIELD has properly defined its interoperability interfaces, so it can be implemented beyond the original software developers and, for example, be adopted by the many electronic capture systems out there. In the wake of this development, DataSHIELD has been enriched with a new framework that enables vast datasets (such as genomics, imaging or for machine learning) to be used inside a DataSHIELD network. Similarly, we are now interoperating catalogues and harmonisation tools, all involving open-source such as Opal, MOLGENIS, Mica, Armadillo and various R packages.

Another special feature of EUCAN-connect is that we don't only serve researchers inside the consortium (of course we do have a great research work package to validate our developments internally) but we also serve a growing network of other research consortia: i.e. we serve a vast user base of all researchers in LifeCycle, RECAP, InterConnect and the Canadian Reach project, and have strong links to large research infrastructures such as BBMRI, ELIXIR and Canadian Maelstrom initiatives. As a consequence, all of the services are in routine operations and being used by this diverse use base, firming up the software, manuals, training materials and supporting teams.

Networking the networks

I see it as a great compliment to all involved that EUCAN-connect practices are now being embraced by new projects such as LongITools, Athlete, members of the EU Exposome Network (EHEN). This pushes EUCAN-connect infrastructure to new levels of maturity, and will provide a proving ground for the challenges we have set ourselves for the coming two years, aiming to further grow this federation of data federations. To promote sustainability of all this work, we have designed the system without central points of failure. Each node is fully autonomous in terms of controlling access, and can be set up to be the coordinator to operate a federated analysis network; BBMRI, ELIXIR and Maelstrom provide safe havens for keeping basic software and data artifacts available in the long term. Thus we expect EUCAN-connect's federated network to be resilient to changing projects and funding, in order to serve many more research projects to come.