Deciphering federated analysis:
What is it and how far it can take us

Date: Monday 10 June, 11am in Sal X [link]

Chairs:

María Chavero-Díez
Laura Portell-Silva

Abstract:

Biomedical research is witnessing a paradigm shift driven by the growing need for collaboration across stakeholders from different domains. In this era, secure data access and analysis are fundamental requirements; however, we are confronted by an inherent dilemma - the need to harness the power of collective analysis while safeguarding data privacy and complying with data protection regulations. This is the challenge that the use of federated analysis seeks to address.

The usage of federated analysis is linked to several ongoing developments associated with ELIXIR; including GDI, EUCAIM, BY-COVID, EOSC4Cancer and technology-oriented projects like EuroScienceGateway and EOSC-ENTRUST. These multiple parallel developments create an imperative need to establish unifying guidelines to harmonise these efforts, guaranteeing synchronised development and coherence in the evolution of the different projects.

This workshop objective will be to identify federated analysis key developmental technologies, assess their maturity level and their suitability for different data types through a collective discussion, enhancing the knowledge exchange between the ELIXIR members working on different projects, platforms and communities.

Within the context of this workshop, we will:

- *Interactively discuss, e.g. through a World-cafe set-up, and aim to define common terminology and requisites for federated analysis and its underlying federated infrastructure, considering commonalities and differences with other approaches.*
- *Pinpoint synergies and gaps across ELIXIR platforms, communities and projects encouraging debate between different teams on tackling them.*
- *Discuss which elements should be considered for evaluating the technology maturity levels.*
- *Identify relevant initiatives, projects and organisations to engage with regarding secure federated analysis.*

- [*Joint Slide deck*]

Agenda:

Time (CET)	Subject
11:00-11:05	**Title**: Introduction **Speaker**: Maria Chavero-Diez (Barcelona Supercomputing Center) *Link to slides or documents*
11:05-11:20	**Title**: The many faces of federated analysis within ELIXIR: The universe of EUCAIM & GDI **Speaker**: Carles Hernandez-Ferrer (Barcelona Supercomputing Center) *Link to slides or documents*
11:20-11:35	**Title**: How Galaxy can be used for federated analytics **Speaker**: Björn Grunning (Albert Ludwigs University of Freiburg) *Link to slides or documents*
11:35-11:50	**Title**: The role of EOSC-ENTRUST in secure federated data analysis **Speaker**: Tim Beck (University of Nottingham) *Link to slides or documents*
11:50-12:30	**Title**: World Café **Moderators**: Laura Portell-Silva & Maria Chavero-Diez (Barcelona Supercomputing Center) *Link to slides or documents*

Notes (collectively taken)

- Introduction
    - Projects include EOSC4Cancer, EUCAIM, EuroScienceGateway, BY-COVID, EOSC Entrust
    - Privacy concerns
    - Challenge of Regulations
- The many faces of federated analysis EUCAIM/GDI
    - EUCAIM imaging focused e.g. cancer
    - GDI genomics focused
    - Example of federated processing. Initially needed to pull all data to the same place for analysis. Within single organization.
    - This might have been scaled up with multiple compute node, but still same org
    - Visiting analysts can use the compute node across multiple organisations. Data outside.
    - Can we bring in multiple data sources and collaborators into the same environment
    - Federated Learning as a specialisation of Federated Analysis
    - EUCAIM architecture. Core services incl. AAI, storage, model repository. Then federated data nodes. Together this forms the EUCAIM Federation.
        - A "FP Dashboard", Manager etc. can coordinate FP (Federated Processing??)
        - There's an FP Daemon at each federated data node, contacted by FP Manager through queues
    - Moving this into GDI to do federated learning and on health data?
    - Q: Is there an AI repository? We will have one w ?? model. Private decision to be made here.
    - Q: ??? Giving data to a frame is one option. Decision on how to access then is common. Or become your own data node and make decisions yourself.
    - ???? Yeah
    - Q: From Excelerate and GA4GH work like WES and CWL that can be used here?
        - Not yet fully compatible. Considering transforming FP Manager to WES,
        - ??: Yes.
    - Q : (to be addressed later, not posed in the room) If the huggingface structure is envisioned for the EUCAIM AI Repo, is there a plan to ensure that the respective metadata are compatible to relevant efforts (such as NFDI4DS, the RDA FAIR4ML and the DOME)?
        -
- Galaxy and Fed analytics
    - Architecture is similar to EUCAIM. Compute nodes are called Pulsar Nodes.

- ELIXIR AAI, job scheduled across distributed nodes including carbon efficiency (or different kinds of efficiency), quite dynamic
- More on this in Compute session
- Pulsar server and network is doing the heavy lifting. Now supported by EuroScienceGateway project. Uses TES.
  - Australian BioCommons use the same setup. Similar set up in US for the Galaxy head node
- Where we struggle perhaps is storage. Different kind of storage, reference data pretty static but distributed to each compute node. Remote files e.g. to bioimaging archive. More dynamic data that has to be uploaded from the user. Also take into consideration long term archive. All Storage!
- Next: Federated Storage. API and command line.
  - Can choose in UI between different storage options. Temporary, huge, permanent etc.
  - Storage engines could be local, s3 buckets, you can connect your own AWS buckets per workflow.
  - Templates for Bring Your Own Storage. Assume some instutitional storage solution. This can then be added into your own template. including credentials. This then can be used for your own workflows rather than central storage.
  - EGI can also provide storage as a commodity. Fill in a form and connect.
  - This relates not just to federated storage but also federate compute. Smart Job scheduling now next, for instance put jobs close to the storage
- Object Store may be the whole history as well exported [as RO-Crate!]
- Different options
  - Deferred data == only pulled when/where job is running. E.g. scheduled job in Italy can pull data only there.
  - Additional scratch space, beyond 250 GB of default quota. Can now have 1 TB but only for 4 weeks. For persistence, plug in your own store.
  - Making data leave Galaxy, InvenioRDM (incl. zenodo?) integration to publish into a long term archive as RO-Crate and get a DOI
  - File storage protocol, s3, irods etc
  - BYOS per user. Can be group-based.
- Q: Pushing responsibility of storage choice to the user, then we need to go further into AAI.. EGI EOSC storage providers. Many providers to deal with could make it difficult
  - We have presented the framework, but how we deploy it is another aspect, e.g. contract with EGI and recommendation to users. What was before is not everything you should do.
  - The admin can for instance override and set the default storage.
  - But want to at least give the user a chance to connect external storage
  - Ideally every University assigns some storage. If then every user could connect this as a private object store. This do mean extra AAI requirements particularly for the job level.
  - John Chilton did lots of this work, see also GCC2024 material in a few weeks.

- EOSC-ENTRUST
  - Also part of Health Data Research UK Federated Analytics programme.
  - Part of EOSC. Coordinated by ELIXIR.
  - Making a tight knit network of TRE (= operators across Europe.
  - Less about tech, more about community and common blue print. Terminology important. E.g. TRE: Trusted Research Environment. Also called Secure Processing Environment (SPE)
  - Project brings together Drivers in use cases, and different providers
  - One area is using RO-Crate in federated workflow processing.
  - Making federated architecture. TRE keeping data secure and private. But moving away from the 1:1 relationship between user and data custodian, allow many researchers access many data sources
  - What is RO-Crate? Community effort supported by ELIXIR on packaging research data with metadata. http://researchobject.org/ro-crate/
  - Include metadata such as authors, how resources relate to each other.
  - RO-Crate profiles adds conventions for additional metadata and types of resources

    - Five Safes Framework is about making safe access to data. Safe Data, Safe Project, Safe People, Safe Settings, Safe Outputs.
    - Come with a governance process, e.g. authorising specific people for specific projects
    - RO-Crate Five Safes Profile https://trefx.uk/5s-crate
    - TRE-FX Architecture for federated analytics
    - https://trefx.uk/#project-outputs
    - How the RO-Crate is processed within architecture.
        - Check phase, Validation Process, Workflow Retrieal Phase, Sign-off phase, Workflow execution phase, DIsclosure Phhase, Publishing Phase, Receiving Phase
    - We tried this with Bitfound and Datashield
    - Now for the video! JSON visible on screen…
    - Input crate for the workflow run. Now inspecting the metadata file
    - References WorkflowHub and input files
    - Person who requested it is included.
    - Now we are asking to run it. The query details uses a particular OMOP codde.
    - Workflow in WorkflowHub with CWL. This consumes that input file in query.
    - Now in results we have expanded the initial crate, we added additional phases and the output files. Here we have an output response file from the query.
    - Within EOSC ENTRUST we are proposing how this can work on a larger scale.
    - Full demo https://youtu.be/Ax86GwTRWzA
- World Cafe
    - Three f2f groups, one online group
    - Three questions, timer per topic.
    - For online participants: slido.com 3019 114
    - (room now reorganizing)