Work package planning status

WP1: Federated Data Discovery & Queries

CINECA kick-off meeting

24th January 2019

Jonathan Dursi for the WP1 team

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825775.

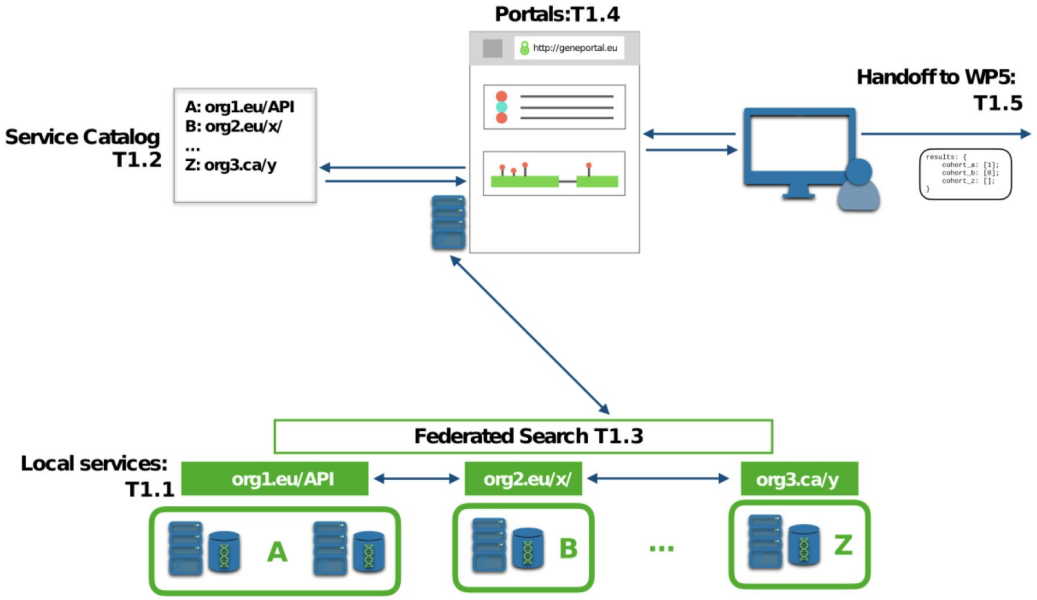
WP1 Participants

- Active (funded person-month) participants:
  - CSC
  - Centre for Genomic Regulation
  - DNAStack
  - EMBL-EBI
  - Erasmus MC
  - HES-SO Genève
  - SickKids, McGill - CanDIG project
  - UMCG
  - University of Tartu
- But **need** to work closely with
  - WP2 - Interoperable AAI
  - WP4 - Federated analyses
- And interact regularly with
  - WP3 - Metadata
  - WP6 - Outreach
  - WP7,9 - ELSI
- So please join in!

WP1 Main goals

- WP1 will provide APIs (and portals that use those APIs) for programmatic or interactive **discovery** of relevant datasets in our federated network, and complex **queries** over those data sets.

- Complex interactive-speed queries: “What is the prevalence of stop-loss mutations in gene X in subjects with/without condition Y?”, **as vs batch analyses**



WP1 Impact

- WP1 will make the technical aspect of searching and querying across distributed genomics and clinical/phenotypic datasets straightforward
- Data stewards will ensure their data has the largest impact
- Clinical + fundamental science researchers will be able to search and query larger datasets and across cohorts
- Queries can be answers in and of themselves, or inputs to other analyses (*e.g.*, running workflows on query-defined cohort)
- Impacts will be highly visible - the bulk of the work (APIs) will hopefully be largely invisible to most users
- Portals may be of use to WP6, outreach

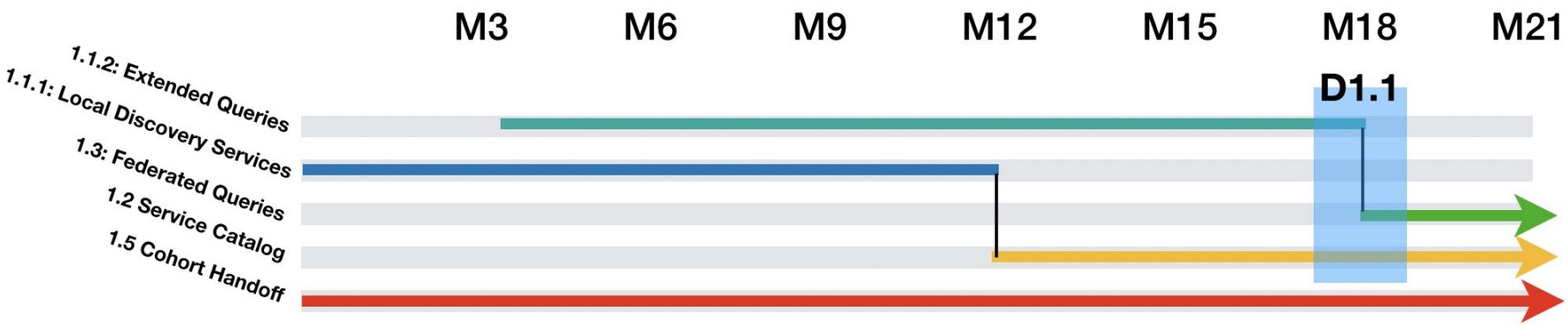
WP1 Deliverables

- Deliverables:
  - Programmatically queryable service catalog (distributed? centralized?) listing all services: **Month 18**
  - Federated discovery and complex queries across datasets: **Month 24**
  - Distributed cohort portals: **Month 48**
  - Interoperability with batch analyses of WP4: **Month 48**

- No blocking dependencies on other WPs
  - To be as useful as possible will require the work of WP2, ensuring interoperable authentication and authorization
    - Can work around this during development
    - Will coordinate with WP2 to ensure no technical barriers
  - For federated query results to be meaningful will require the work of WPs 3 & 5, ensuring semantically compatible metadata
    - Can work around this during development
  - Will coordinate with WP4 on dynamic cohort data models for “handoff” to batch analysis

WP1 Timeline first reporting period

- First reporting period - largely independent of other WPs
- Will be working on discovery APIs & services, extended queries, and inventory of available services.
- Working together via GA4GH Discovery workstream, to ensure APIs get used elsewhere



WP1 Risks and action items

- Risk: Difficulty coming to agreement on **API definitions/standards** (e.g., failed Discovery Search API effort)
  - Mitigation: Start simple (MVP) and iterate between partners; don't attempt standardization until well along path
- Risk: problems with **interoperability of implementations**
  - Mitigation: API-first approach (OpenAPI), build from there
  - Mitigation: Continuous interoperability testing during development (aided by API-first)
- Risk: New data types/services emerge which are **not yet supported** by APIs
  - Mitigation: lightweight search/query APIs that "wrap" low-level queries
  - Mitigation: focus on extensibility from begining; very compatible with start simple + iterate

WP1 User/stakeholder engagement

- Primary “audience” for WP1 deliverables in first reporting period are other genomics projects interested in interoperable query APIs (and possibly implementations)
- WP1 will engage other projects interested in interoperable querying by interacting through existing GA4GH workstreams
- Will plan hackathon to coincide with upcoming GA4GH meetings

WP1 Initial plans for sustainability

- WP1 will aim to work with GA4GH to ensure its APIs evolve into standards
- Standards will continue to be maintained by GA4GH workstreams
  - For querying and discovery,
- Implementations will continue to be developed by partners
  - WP1 will encourage open-source implementations
  - CanDIG will continue maintaining and using its implementation for forseeable future