SwissEGA API analyses

How well the API and service would serve our queries

- Informed by WP5 needs: https://github.com/cineca-wp1/project_management/blob/master/queries.md , allowing a wide range of non-prescribed queries
  - M12 examples - simple discovery query examples
  - M18 examples: More complex return values including histogram of counts, min/max, etc., and Include RNA expression, include capabilities for treatments, outcomes, lab work, references to other data

- **Answer**: Return list of datasets matching query concepts (plus Boolean logic)
  - Ok (M12): Samples corresponding to patients with disease X and variants in gene Y
  - Ok (M12): Patients with disease X and variants in gene Y that have samples of tissue Z available
  - Ok (M18): Patients with disease X and mutations in gene Y with RNA expression data available in gene Y
  - Ok (M18): List of patients in registry with a specific rare disease

  - https://github.com/biocaddie/WG3-MetadataSpecifications/blob/master/doc/v2.2/DataMedDATSspecificationv2.2-NIH-BD2KbioCADDIE.pdf

Top down – Competency queries

Internal bioCADDIE code	Competency question
BGUC1-1	Search for disease x data of all types across all databases (Note: these first three use cases are linked; also there is a Common Data Element for the disease x [HD])
BGUC1-2	Search for data type x related to disease x and disease y to compare behavioral studies (HD and ADHD)
BGUC1-3	Search for data on diseases c, d, e, and f that mention disease x or the disease x gene
BGUC2	Search for organism x in biological process y (apoptosis) at scale z with an estimate of the reliability of the annotations
BGUC3-1	Search for new drug x to predict and track biological process y (cardiotoxicity)
BGUC3-2	Search for data type x ('omics correlates) of biological process for drugs related to drug x
BGUC3-3	Search for data types a, b, and c (EHR data, self-report, sensor) to determine *natural history* of patients given drugs similar to drug x
BGUC3-4	Track responses to treatment to ensure detection of biological process x
BGUC3-5	Find patient data *"like these"* with similar treatments, responses to treatment, genetics

SPUC1	Search for birth cohort x (adolescents) with combination of imaging data types a-z to identify phenotypes a-z predictive of disorders x and y (alcohol and drug use)
SPUC2	Search for data type x (imaging data), *across the lifespan*, with deep phenotyping and data type y (GWS data)
SPUC3 PRE3	Search for birth cohort data that are harmonized on variable x (educational attainment) to understand *historical* impact on biological process y (adult mortality)
SPUC4	Query broader and updated phenotypic categories for generalized enrichment analysis on data type ('omics)
SPUC5	Create virtual networking environment, linking data types x and y and literature to understand biological process (molecular biology of carcinogenic pathway), which is accessible to medical professionals and patients.
SPUC6	Search for constraints of genotypes a-z and phenotypes a-z
SPUC7-1	Search for EHR data to monitor side effects of drug x with condition/context y, data quality z, prevalence of medication use, etc.
SPUC7-2	Link EHR data with knowledge bases a-z (e.g., SemMedDB, DrugBank, etc.)

How well the API and service meets WP3 needs

- Loose coupling between API + metadata fields
  - Allow metadata (both names and values) to evolve rapidly over the next 1-2 years without requiring changes to code
- **Answer**: SwissEGA: Information retrieval approach : ranked list of best results
  - Querying a data retrieval system produces exact/precise results or no results if no exact match is found
  - Querying an IR system produces multiple results with ranking. Partial match is allowed
  - E.g.: bioCADDIE Dataset Retrieval Challenge (https://biocaddie.org/biocaddie-2016-dataset-retrieval-challenge-registration)
- Clear place to implement mapping when these mappings are defined by WP3
- **Answer**: MeSH/UMLS mapping + query expansion (word2vec, ClinicalBERT)

How well aligned is this API/service with the GA4GH (CINECA has a commitment to be aligned with GA4GH standards)

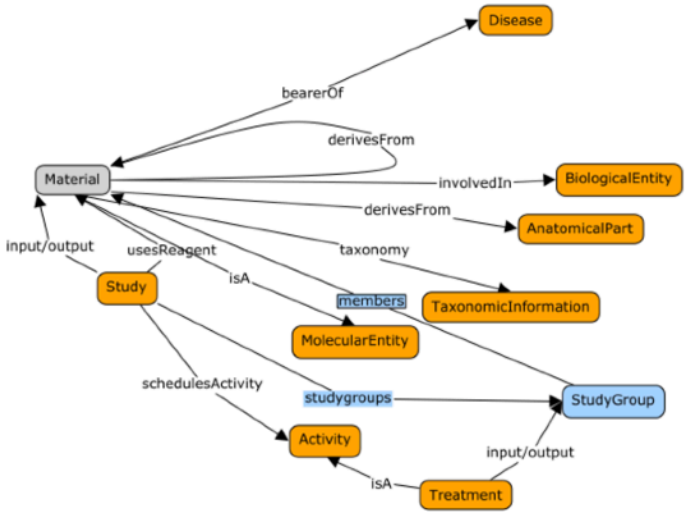
- **Answer:** Working on the alignment with EGA metadata model

Demonstration of the feasibility/usefulness of this API/service

- **Answer**:
  - datamed.org:
    - 2,336,403 datasets, 75 dataset repositories (e.g., UniProtKB, dbGap, etc.)
  - SwissEGA: MyHealthMyData, SPHN
  - Model validation tools:
    - https://github.com/datatagsuite/dats-tools/tree/master/dats

How easy is it to implement this API/service upon existing infrastructure/APIs

- **Answer:**
  - Dataset metadata needs to be converted to the DATS format
  - Plenty of documentation:
    - https://wg3-metadataspecifications.readthedocs.io/
    - https://github.com/datatagsuite/examples
  - It is not designed to connect to existing structure directly (prod data, service load, etc.)
  - It should use a lightweight information retrieval engine to host metadata (elasticsearch, terrier, lucene, etc.)



Is this API/service already on the site’s roadmap

- **Answer:**
  - Yes, being extended in the context of SPHN
  - Planned to expose CoLaus/PyCoLaus metadata

  - Demo: http://candy.hesge.ch/SwissEGA/