Phenopackets in EGA: some points to discuss

1 'Phenopacket' as entity

1. Substitute for our Sample object altogether? (this is, since Phenopacket is individual-centered we adopt new data model where all Family/Cohort/Invidivual/Biosample data is in phenopacket format and all sample(s) data is naturally included there as Biosample(s) object and we get rid of our current (ENA) Sample object). This way Phenopackets HtsFile object allow to link to associated files to Biosamples.

2. Linked to our current Sample object, Phenopackets would be adding only info of Individual and top-level elements such as Family/Cohort where Sample belongs while keeping SRA sample schema for sample?

2 'Phenopacket' format

1. *Phenopacket* bundle doesn't fit all purposes, for instance Pedigree or Cohorts. *Phenopacket schema* allows designed flexibility and still integration among resources and global interoperability of common blocks.

2. *Phenopacket schema* implementations are naturally project (study type) -specific, so EGA Data Model definition should precede *Phenopacket schema* implementations to handle all of our use cases.

3. Or can we come up with a *Phenopacket schema* that fits all purposes and then just leave empty fields and use only those applicable fields for each cases and validation rules applied per case type? Is this possible allowing by default Cohort and Family schemas to co-exist? Interpretation/association vs other studies?

4. A minimal set of *Phenopacket schema* (and thus, of "phenopacket pre-formats") implementations need to be defined to accommodate all study types in EGA (matching our current or future 'study type' field, e.g cancer genomics or an extended version of it)

5. *Phenopacket* bundle is Individual-centered, *Phenopacket schema* blocks would allow to design a biosample-centered implemetation to fit our data model, but it's a better idea to shift to Invididual-centered data model to allow better representation of use cases with case-control individual each having case-control (normal-abnormal/tumoural, healthy-unhealhy, etc) samples, which are not easy no represent in current model.

3 Phenopacket format validation

Data Model driving *Phenopacket schema* definition will allow QC (content validation of phenopackets)

3.1 Study unit

Family (n members), Cohorts, Groups, Individual(s) This will define if *Phenopacket schema* top level element as Cohort(s), Family or Individual (Phenopacket), so if a Pedigree object or Groups need to be defined, as well as the meaning of objects (the problem of semantics in phenopacket schemas).

3.1.1 Family and Cohort Studies and expected linked files

Family and Cohort Studies will require a Family or Cohort top-level element gathering Phenopacket sets. Aggregate data types such as "Cohort" and "Family"are expected to contain aggregate HTS file data i.e. merged/multi-sample VCF at the level of the Family/Cohort, but each member Phenopacket can contain its own locally-scope HTS files pertaining to that individual/biosample(s).

3.2 Study Focus

This will define if *Phenopacket schema* op level element as Phenopacket bundle or Interpretation, if only one or many disease objects need to be defined or if phenotypic feature objects. Focus can be Analysis e.g GWAS which would require an Interpretation schema. Focus of study can be a Disease diagnosis/ Phenotypic feature observation/ Event (stoke, death, side effect)/ Outcome (progression/ survival), Response to treatment. Focus can be combined with intervention in both possible combinations e.g chemotherapy and survival

3.2.1 Interpretation/Association studies and meaning of objects

Interpretation studies will requiere a top-level element Interpretation gathering together Phenopackets or Families where genomic interpretation comes from. Needs genomic interpretation object. e.g 'Genes' object within Interpretation/ association studies mean suspected/putative causative/ risk genes (the difference in meaning should be defined in implementation), as opposed to for instance non-Interpretion studies e.g cancer molecular subtyping and response to treatment, where 'Genes' object could mean markers (the difference in meaning should be defined in implementation)

3.3 Study Design

timeline Longitudinal vs Crosssectional

approach Observational vs Interventional This

direction Prospective vs Retrospective

This would define if one or a series of phenopackets (Phenopacket set) should be entered per individual and their grouping

3.3.1 Longitudinal studies and timelines representation

Studies with timelines will require a set of Phenopackets or date-associated objects as mandatory.

3.3.2 Focused observational studies and groups

Need to specify observation focus (as negative or positive findings) and Comorbidities (or history of disease) as Groups or History of disease, not as the same object of newly diagnosed diseases - or maybe using some distinction - in order to maintain meaning.

e.g cardiac events/ coronary disease diagnoses in British population of different risk groups (Comorbidities), covid19 severity in different risk groups

3.3.3 Interventional studies and groups

Need to specify Intervention/Medication object, timeline of intervention and observations/measures, need to define different phenopacket sets for groups?

eg. diet and cardiac event, preventive vaccine and immunity, therapeutic intervention and tumour progression

4 Use case exercises

Uses cases could be drivers to develop minimal set of *Phenopacket schema* implementations needed to handle all EGA use cases.

To start with we could assess which the most commonly submitted study types are or a few representative

types and identify in each case whether *Phenopacket* bundle suffices or what number of *Phenopacket schema* prototipes would be needed to handle them.

1. CGAT: electronic health record data (including diagnosis, clinical findings, medical interventions and treatment prescriptions, medical/lab measurements), pharmacy data, environmental and socio-demographic survey data

   What is better suited phenopacket schema for this¿ 1. all patients EHR ¿retrospective, observational, individual-centered, many diseases/ phenotypic features 2. interventional records + EHR + lifestyle/socio survey + follwow up: does a certain intervention (lifestyle, medications, interventions) affect a phenotypic outcome in genera/ in genetic risk group

2. Covid-19 human pilot CODIV-19 phenotype data includes baseline, one-time data (events/interventions, with or without a date associated) and longitudinal data (symptopms, lab tests, viral load test, medication), which will required a set of timelined phenopackets per individual

   (web app, hospitalization, longitudinal study) Tooling: - HPO terms extended to include temrs for covid19 - Beta version of phenopacket elements for covid19 (including medical actions including exposure, pharmaceutical treatments with timecourses , doses, and intervals, test results, etc), we have opportunity to test and give feedback, for example, how to represent some survey data on physical lifestyle that are planned to be collected or comorbidities.

3. Hipothetically Covid-19 prevemtive vaccine/novel antiviral clinical trial in different ethnic backbrounds

4. Other exercises

5 Phenotypic data findability

Individual level data available only to authorized users- for download.
Summary/ dataset level could be allowed to all registered/ all users? - for user to find relevant datasets through Data Portal

- How many datasets match a certain Disease in study focus, patients of a certain age group, and have WGS data available, for example
- How many datasets match a certain Disease in study focus and certain ancestry group

How is this implemented? could be API on phenopacket protobuf db schema?