Experience using Phenopackets schema for phenotypic data submission of common EGA use cases: Phenopacket exercise: GCAT example

Claudia Vasallo

European Genome-phenome Archive

claudia.vasallo@crg.eu

July 24, 2020

Overview

1. Phenopackets purpose in EGA?
2. GCAT use case
   - Use cases, data types
   - Definition Schema
3. To be discussed

Phenopackets purpose in EGA

A) A static substitute for current Metadata Model, with prepackaged definition of use cases (1 or n), allowing better representation and new Sample elements including Individuals, Extended Pedigrees, Measurements, etc > Submissions are aided to be convertible into Phenopackets (e.g. intermediate template (csv, xml, json), web-mediated or programatic submission)

B) A flexible substitute for current Metadata Model, allowing, in addition to prepackaged objects, some bespoke objects to be submitted (dump for all possible use cases even if meaning of objects cannot be forced and hence they might not be widely interoperable) > Prepackaged and Customary convertible submissions + any home-made phenopacket accepted without converting to EGA Phenopacket data model(s)

EGA use cases

- 1 genomic file + 1 time point observation
- 1 genomic file + n time points/ longitudinal observations
- n genomic files (tumors, expression data) + n time point interventions/observations

Use cases not representable with phenopacket schema v1

- **Timecourses (longitudinal), many timepoints per individual, extended events**

e.g. Treatments, Exposures, Events like hospitalizations

> PhenopacketSets or phenopackets bundle with *TimeElements*

- **Intervals for age at timecourse events**

e.g. Exposure occurred between age 41 and 42

> Extend Interval message in *TimeElement* to include age at start and at end of Exposure, Treaments etc

Use cases not representable with phenopacket schema v1

- **Behavior/lifestyle/sociodemographic and other correlate data types**

> Use *Exposure* mesages

- **Quantitative measures**

e.g. Lifestyle (exposures), antropometric, medical, laboratory measures

> A new 'Measurement' message based on *Quantity* and *OntologyClass*

Use cases not representable with phenopacket schema v1

- **Non standard units, categories and scores**

e.g. Scores and categories based on quantitative data such as measures or test results

> Messages based on *Quantity* and *OntologyClass*

- **Other bespoke messages when needed**

e.g. scores and categories based on quantitative data such as measures or test results

> Build Flexible blocks?

Reuse of v1 messages and bespoke messages

message EncounterSet - based on *PhenopacketSet*

- Resuses *Individual*, *Hts.file*, *Metadata* from Phenopackets (avoids repeating this info in every phenopacket)
- Includes bespoke messages Encounter similar to *Phenopacket*

message EncountersSet
string set_id = 1;
string description = 2;
org.phenopackets.schema.v1.core.Individual subject = 3;
repeated Encounter encounters = 4
repeated org.phenopackets.schema.v1.core.HtsFile hts_files = 5;
org.phenopakets.schema.v1.core.MetaData metadata = 6;

* files for Individual when only 1 time, otherwise hts_files messages in every Encounter?

Reuse of v1 messages and bespoke messages

message Encounter - based on *Phenopacket*

- Resuses *Individual*, *Hts.file*, *Metadata*
- Includes any number of bespoke messages *Finding*, *Diagnose* and*ExposureExtended* that occurred at a certain date

```
message Encounter
string encounter_id = 1;
repeated org.phenopackets.schema.v1.core.Age age_at_encounter = 2;
google.protobuf.Timestamp time_at_encounter = 3;
repeated Diagnose diagnoses = 5;
repeated Finding findings = 4;
```

* 1 date: any disease or findings present (or absent) ongoing at encounter time (disease "Tietze's disease", findings: phenotypic feature "obesity", measurement "weight", clinical finding "Hemorrhage"

Reuse of v1 messages and bespoke messages

message Diagnose - based on *Disease*

- Reuses all Disease, including date, age (duplicated with Encounter Set)
- Includes bespoke message Origin for Medical Unit

```
message Diagnose
org.phenopackets.schema.v1.core.Disease disease = 1;
Origin origin = 2;
Method origin = 3;
```

* diseases only, accompanied by associated origin and method "Tietze's disease" (ontology)

Reuse of v1 messages and bespoke messages

message Findings - based on *PhenotypicFeatures, Exposures*

- Reuses messages, including date, age (duplicated with Encounter Set)
- Includes bespoke messages based on *OntologyClass*

message Findings
one of
repeated ClinicalFinding clinical_findings= 1;
repeated ExposureExtended exposures = 2;
repeated Measurement measurements = 3;
repeated Score scores = 4;
repeated Category categories = 5;

Reuse of v1 messages and bespoke messages

message ClinicalFinding - based on *PhenotypicFeature*

- Reuses all PhenotypicFeature)
- Includes bespoke messages Origin for Medical Unit and Method

message ClinicalFinding
org.phenopackets.schema.v1.core.PhenotypicFeature term = 1;
Origin origin = 2;

* any finding, sign found by clinician, e.g "Hemorrhage"

Reuse of v1 messages and bespoke messages

message ExposureExtended - based on *Exposure*

- Reuses all Exposure, including date, age (duplicated with Encounter Set), type, severity, evidence
- Includes also TimeElement for interval in age at exposure instead of timestamp
- Includes also messages Quantity, Frequency

```
message ExposureExtended
org.phenopackets.schema.v1.1.core.Exposure exposure = 1;
org.phenopackets.schema.v1.1.core.Quantity quantity = 2;
org.phenopackets.schema.v1.core.OntologyClass frequency = 3;
org.phenopackets.schema.v1.1.core.TimeElement exposure_time = 4;
```

* any exposure either environmental or lifestyle that is present/ accounted for to have occurred at *Encounter* time, e.g. ”smoking” (3 packs daily for 3 years), ”physical activity” (9 hours weekly), ”mediterranean diet”

Reuse of v1 messages and bespoke messages

message Measurement - based on *OntologyClass* and *Quantity*

- Reuses building blocks
- Includes bespoke messages Origin for Medical Unit and Method

message Measurement
string description = 1;
org.phenopackets.schema.v1.core.OntologyClass parameter = 2;
org.phenopackets.schema.v1.1.core.Quantity quantity = 3;
Origin origin = 4;
Method origin = 5;

* any measure that is taken/ accounted for to have occurred at *Encounter* time, e.g. "waist circumference" (77 cm), "systolic blood pressure" (130 mm Hg)

Reuse of v1 messages and bespoke messages

message Category - based on *OntologyClass*

- Reuses building block

message Category
string description = 1;
org.phenopackets.schema.v1.core.OntologyClass type = 2;
org.phenopackets.schema.v1.core.OntologyClass classification = 3;

* any categorical value that doesn't fit in *PhenotypicFeature* e.g. "obesity" - "class II obesity" (WHO classification associated to BMI), "risk of disease" - "high" (based on parameters not being stored)

Reuse of v1 messages and bespoke messages

message Score - based on *OntologyClass*

- Reuses building block

```
message Score
string description = 1;
org.phenopackets.schema.v1.core.OntologyClass type = 2;
double value = 3;
```

* any numerical value that doesn't fit in *Quantity* e.g. "epic score" - 4 (for alcohol consumption behavior)

To be discussed: Implementation and Interoperability

- Phenopackets purpose? - all phenotypic data, computable only, phenome but not epidemiological (e.g sociodemographic or lifestyle correlates)
- Meaning of objects. "One of" structure is meant for 1 schema with shared meaning among partners, for EGA we would have to make them compatible
- Redundancy, e.g *Score*/ *Category* and other derivatives worth storing? vs *Finding* (redundancy vs interoperability)
- Different codings for the same thing: e.g. *PhenotypicFeature* (e.g "Hypertension") quantified as *Category* (severe) vs *Measurement* (Systolic Blood pressure) with *Quantity* (260 mm Hg))

To be discussed: Implementation and Interoperability

- Unknown data, such as *TimeElements* (e.g. past events e.g smoking without known timeframe, family history of disease without pedigree (self-reported))
- Non standard units, custom scores (e.g. predicted risk?), or codes such as ICD > codes remapping? define in metadata? store both original and remapped for interoperability?
- Interoperability at API level (EGA Beacon, EGA Data Portal): ontology mapper, converters, etc, instead of impossing an input model