Representing phenopackets in EGA - Technical design RFC

# Summary

*Describe the vision of the RFC in 2-3 sentences. Consider this section as your "elevator pitch" or "release note" to the EGA team.*

This document proposes how to incorporate phenopackets into the EGA.

# Author(s)

*Add your first and last name if you are contributing to this RFC.*

Mallory Freeberg, Coline Thomas,

# Motivation

*Describe the user need or technical need in detail with alignment to the EGA roadmap priorities where possible.*

Phenopackets are emerging as a community standard for storing and sharing controlled access (private) phenotype information (metadata) about an individual ([link](https://www.ga4gh.org/news/phenopackets-standardizing-and-exchanging-patient-phenotypic-data/)). “The Phenopackets standard aims to facilitate communication between the research and clinical genomics communities by creating an ecosystem of interoperable tools and resources that can use phenotypic data with fewer barriers.” The EGA would benefit from aligning with and implementing this standard to support submission of, access to, and searching of private phenotype metadata. To maximise the usefulness of phenopackets in the EGA, there needs to be careful thought and planning about how to incorporate into the current/future EGA metadata standard, including the data model. This RFC proposes a plan for such integration and outlines key steps to get there.

## User Stories

|  |  |
| --- | --- |
| **User** | Data submitter |
| **Story** | I want to be able to submit controlled access phenotype metadata so that my data can be more useful to others to support their research aims. |

|  |  |
| --- | --- |
| **User** | Data requester |
| **Story** | I want to be able to search through controlled access phenotype metadata so that I can more easily find which datasets might be useful to support my research aims. |

|  |  |
| --- | --- |
| **User** | EGA team member |
| **Story** | I want to be able to implement the phenopacket standard so that the EGA can be an example for adopting GA4GH and other community standards. |

*Add more stories as needed...*

# Detailed Design

*Explain the design in sufficient detail such that the implementation and (if appropriate) the interaction with existing EGA services are both reasonably clear. Include diagrams and tables to improve clarity.*

## Background

The scope of this RFC specifically covers how phenopackets will be processed and represented in the EGA and what the phenopacket format will be. It does not cover design for phenopacket conversion tools or services to access phenopacket information, although the requirements for these types of tools/services should be considered in the design here.

## Proposal

### Phenopacket journey [whiteboard a diagram]

1. Phenotype metadata submitted encrypted to EGA
   1. Phenotype metadata **is not** in phenopacket format
      1. EGA converts to phenopacket OR asks submitter to convert (to develop or re-use an existing tool/service to do this)
      2. We could imagine a “phenopacket pre-format”. May be in dsv (tsv, csv..) format following a specific template(s). We could use attributes common in EGA or ENA (for example: sample attributes), but they would be clearly mapped to the phenopacket schema (see CRG [bio-ega slides](https://drive.google.com/file/d/1Uu0fG1KzL2YeCzav_ZzznaHKE1TcgDse/view)). That would give us a format easily used by submitters, and easily translated in phenopacket later on for us.
   2. Phenotype metadata **is** in phenopacket format
2. Phenopacket file is linked to an Individual or Sample object
3. Phenopacket file is re-encrypted and archived
4. Contents of phenopacket file are loaded into a dedicated DB/table in the Vault
5. Phenotype metadata accessed by approved users
   1. Phenopacket files are downloaded via standard EGA protocols
   2. Phenopacket information is accessible and searchable in DB/table (currently no tool/service exists for this)

### Phenopacket format

What is th native format? JSON? YAML? Others?

What do we do if we are given phenotype information that we can’t format into phenopacket format? Make PR against [phenopacket GH repo](https://github.com/phenopackets/phenopacket-schema)?

### Phenopacket validation

How is the format validated? E.g. if in JSON, can use a json schema validator in any number of languages: <https://json-schema.org/implementations.html>

### Phenopacket file association

**Proposal**: Phenopacket file is linked to Sample (now) or Individual (future) objects similar to how data files are linked to Run, Analysis, and Array packet objects. Does this require a major change to the ENA Sample object - how do we implement this?

<?xml version="1.0" encoding="UTF-8"?>

<SAMPLE\_SET>

<SAMPLE alias="CACO2" center\_name="TEST\_SUBMISSIONS" accession="ERS853135" broker\_name="EGA" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">

<IDENTIFIERS>

<PRIMARY\_ID>ERS853135</PRIMARY\_ID>

<SUBMITTER\_ID namespace="TEST\_SUBMISSIONS">CACO2</SUBMITTER\_ID>

</IDENTIFIERS>

<TITLE>CACO2</TITLE>

<SAMPLE\_NAME>

<TAXON\_ID>9606</TAXON\_ID>

<SCIENTIFIC\_NAME>Homo sapiens</SCIENTIFIC\_NAME>

<COMMON\_NAME>human</COMMON\_NAME>

</SAMPLE\_NAME>

<DESCRIPTION>This is a colorectal adenocarcinoma cellline sample.</DESCRIPTION>

<DATA\_BLOCK>

<FILES>

<FILE filename="CACO2\_phenopacket.json" filetype="phenotype\_file" checksum\_method="MD5" checksum="cc013849d561286939f2361043e15992" unencrypted\_checksum="d79971c9dad3b6945f00df483798850e"/>

</FILES>

</DATA\_BLOCK>

<SAMPLE\_ATTRIBUTES>

<SAMPLE\_ATTRIBUTE>

<TAG>subject\_id</TAG>

<VALUE>CACO2</VALUE>

</SAMPLE\_ATTRIBUTE>

<SAMPLE\_ATTRIBUTE>

<TAG>gender</TAG>

<VALUE>male</VALUE>

</SAMPLE\_ATTRIBUTE>

<SAMPLE\_ATTRIBUTE>

<TAG>phenotype</TAG>

<VALUE>Colorectal adenocarcinoma</VALUE>

<UNITS>EFO:0000365</UNITS>

</SAMPLE\_ATTRIBUTE>

<SAMPLE\_ATTRIBUTE>

<TAG>ENA-CHECKLIST</TAG>

<VALUE>ERC000026</VALUE>

</SAMPLE\_ATTRIBUTE>

</SAMPLE\_ATTRIBUTES>

</SAMPLE>

</SAMPLE\_SET>

**Proposal**: Phenopacket file is archived via standard EGA protocol into FIRE object store.

### Phenopacket in DB

**Proposal**: Contents of the phenopacket file are parsed and loaded into a dedicated DB/table in the Vault. Information is only accessible by approved users. How do we do this?

## Roadmap for implementation

1. Accept, validate, and archive phenopacket files (partially supported already).
2. Archive a phenopacket file linked to a sample or individual object.
3. Create and populate a phenopacket DB.
4. Allow controlled access searching of phenopacket information in DB.

## Acceptance Criteria

*Acceptance criteria are the conditions that a RFC must satisfy to be accepted by users or other stakeholders.*

This RFC will be accepted when there is agreement between CRG and EBI on how the EGA will store and share phenopackets. The design must support the following requirements:

* align with standards set out in the GA4GH phenopacket format
* enable phenopacket file to be downloaded by approved users
* enable phenopacket file contents to be queried by approved users
* allow submission and release of datasets if controlled access phenotype information cannot be represented by phenopacket format
* allow for curation over time of existing controlled access phenotype information into phenopacket format

## Unresolved Questions

*What aspects of the design do you expect to clarify further through the RFC review process?*

*What aspects of the design do you expect to clarify later during iterative development of this RFC?*

*What are the drawbacks and limitations of the current RFC?*

*Why should this RFC not be implemented?*

1. Can Sample and Individual objects reference files like Run, Analysis, and Array packet objects already do? If not, how do we modify this for EGA use case (w/o depending on ENA/BioSamples)?
2. How do we enforce controlled access to phenopacket information stored in a DB/table in the Vault?
3. How do we flatter phenopacket information such that it can be stored in a 2D table in the Vault?
4. A phenopacket can describe both an individual and the samples that were taken from them (e.g. some unrelated to genomics like biopsy or blood test, that are purely phenotyping data, and some that correspond to EGA kind of samples). So we need to reflect on which object (individual or sample) we will link the phenopacket. If it describes both the individual AND a sample recorded in EGA (or maybe several), to which one should we link it?

## Prior Art

*Share references to prior art to deepen community understanding of the RFC, such as learnings, adaptations from earlier designs, or community standards.*

Solve-RD as the first example of storing phenopacket files in EGA [expand on this].

## Alternatives

*Highlight other possible approaches to delivering the value proposed in this RFC. What other designs were explored? What were their advantages? What was the rationale for rejecting alternatives?*

Phenopackets as a first-order object in the EGA data model, like Sample?