Phenopackets Domain Analysis

Phenopackets Domain Analysis Team

Aug 17, 2020

CONTENTS

CHAPTER

ONE

INTRODUCTION

This document is an effort to capture an analysis of the domain modeled in the Phenopackets schema, additional use cases and modeling requirements, how to address interoperability with the FHIR standard, and other relevant content. It is one of the artifacts developed as part of an NIH funded project.

In July of 2019 the NIH issued a notice (NOT-OD-19-122) to encourage investigators to explore the current uses of FHIR in research to ultimately improve methods for clinical researchers to use, standardize, and share electronic health data. The Broad Institute as a host institution for the Global Alliance for Genomics and Health (GA4GH) was one of two sites awarded a contract to pursue these goals under NIH/NLM contract #75N97019P00280, in partnership with the Jackson Lab, Oregon Health & Science University, Oregon State University, and Mayo Clinic. The project goals are to:

1. Coordinate with the appropriate HL7 FHIR work group(s) and develop a FHIR Implementation Guide (IG) based on the data modeled in the GA4GH Phenopackets schema,
2. Identify sets of translational research partners to demonstrate data exchange using FHIR application programing interfaces (APIs) for selected use cases,
3. Develop Driver Use Cases to inform IG development and for pilot-testing,
4. Support pilot-testing,
5. Compile pilot test results into a report,
6. Provide feedback to HL7 to revise relevant FHIR resources, profiles, and/or IGs,
7. Provide feedback to GA4GH on Phenopackets schema and application to use cases, and
8. Disseminate deliverables and obtain feedback for future development.

Note: This site is in early development. Please do not link to any specific page on this site other than the home page. Otherwise, your links will likely break as the initial content is drafted.

CHAPTER

TWO

STAKEHOLDERS

The domain analysis will identify various stakeholders involved in developing standards and other technical solutions for the use and exchange of phenotypic information. The goal of this content is to provide a central resource for all relevant stakeholders, their working groups, significant resources provided by each stakeholder, and how to engage with each of them. The following is an initial list of stakeholders. Others will be added as part of the domain analysis and further information will be collected for each stakeholder.

- GA4GH community
  - All workstreams and driver projects related to the Phenopackets standards
  - The Clinical & Phenotypic Data Capture & Exchange Work Stream
- Phenopackets working group and schema
- HL7 FHIR community
- HL7 working groups developing relevant FHIR resources
  - Biomedical Research and Regulation Group
  - Orders and Observations Group
  - Patient Care Group
- Kids First and the GRIN network

CHAPTER

THREE

USE CASES

This project will develop use cases to help guide the development of the technical artifacts related to the Phenopackets specification. Use cases can be proposed and provided by community members, and they will be collected as a part of this documentation for the benefit of the community and future projects.

- **Driver use cases** guide the technical development for the duration of the project. They represent a small subset of use cases that could be developed within the scientific and clinical domains that are covered by the Phenopackets specification. They are used to scope the work done by the project team in coordination with the project sponsor.
- **Community use cases** are an unconstrained set of use cases related to the Phenopackets specification. They are captured in this documentation to inform future work, including discussions regarding the scope and intent of the specification as well as the development of technical artifacts.

Please see the *driver* and *community* use cases pages for additional details.

3.1 Driver Use Cases

The following driver use cases will guide the technical development during the current project period. Please see the *community use cases* for additional use cases contributed by the community that could guide future work on this project.

A Driver Use Case represents the same content as a Community Use Case, but it is primarily developed by the project team for the purpose of guiding the technical development and testing activities during the project period. These use cases will be based on clinical and research topics developed in coordination with the pilot sites, involve interactions with FHIR and Phenopackets technical components, and specify the data flow, goals, and outcomes of the use case. For this project, the Driver Use Cases will be largely based on existing domain models in Phenopackets and FHIR, and existing terminologies. The primary goal of these use cases is to integrate existing technical solutions rather than evolve any specific one of them. However, in pursuing integration, specific gaps and enhancements will be needed and will be proposed for future development. These use cases will be documented and analyzed within the Domain Analysis Document.

3.1.1 Kids First FHIR interoperability and Phenopackets extraction

This is a Kids First based use case to demonstrate core data interoperability between the evolving Kids First Implemenation Guide (IG), developed separately, and the Core IG that is being developed during this project. Kids First initially attempted to adopt an existing Phenopackets-based FHIR IG to expose existing data through a FHIR API. This effort led to some difficulty with aligning their data, and the development of a new FHIR IG for Kids First. This use case will be based on the scientific and program goals of Kids First. In applying the use case to this project, we will analyze the Kids First IG, one of the Kids First datasets being exposed with this IG, along with the existing Phenopackets IG and the Phenopackets native schema, to specify and demonstrate data interoperability for core entities represented by these different models and FHIR IGs. Specifically, the following will be implemented and tested by this use case:

- We will develop a user story based on the Kids First dataset that illustrates a clinical and/or translational use of the type of data that is supported by the Phenopackets specification.
- We will use the Kids First IG, one of the Kids First datasets exposed with the Kids First IG, and the Phenopackets schema to identify at least 3 core entities that will be modeled in the Core IG.
- The core entities will be thoroughly documented and analyzed in this Domain Analysis Document to better understand the scope of the FHIR representation in the Core IG.
- A demonstration FHIR server instance will be stood up and the Core IG will be loaded into this instance for validation of the IG, and for data validation during subsequent interactions.
- Demonstration of data exchange between the Kids First IG and server and the Core IG and its demonstration server. This could be accomplished with Jupyter notebook examples or other code snippets. This will test and document any issues around FHIR to FHIR interoperability between the two IGs and how to address them going forward.
- Write POC code to extract Phenopackets instance(s) in the native format by querying Core IG based data to test interoperability between the Core IG and the current Phenopackets schema.

This use case is focused on advancing the representation of phenotypic data in FHIR, primarily through:

- Technical interoperability
- Data exchange
- The Core IG being the core common denominator/hub for interoperability between the Kids First IG, the existing Phenopackets IG, and the Phenopackets native schema
- Help us establish and test our technical artifacts and workflow for ongoing IG development beyond the period of this project.
- This use case is focused on data representation and message exchange for few core entities. Higher level domain or application specific use cases will build on the development guided by this use case.

3.1.2 Phenotype-driven molecular genetic diagnostics

This use case demonstrates the value of FHIR interoperability in a clinical setting. The goal is to enhance the communication of phenotypic information in clinical settings to support well-established clinical workflows.

Clinicians and hospitals send blood or DNA samples to a laboratory for molecular diagnostics of Mendelian disorders. Traditionally, little or no detailed information about clinical phenotypes is provided. In the experience of Peter Robinson, paper entry forms were used which allowed submitters to name the suspected diagnosis or other relevant clinical information. In the NGS era, where gene panels, exomes, or genomes are used for diagnostics in cases where the underlying diagnosis is unknown, it would be highly desirable to have more information about phenotypic findings to guide interpretation of NGS findings.

We will demonstrate the exchange of phenotypic information to a hypothetical lab by using our POC Core IG FHIR server. The laboratory request, the Phenopackets file, and any native FHIR representations of the same phenotypic

information for the attached Phenopackets file will be communicated to the POC server. This will simulate the submission to a real FHIR based laboratory/EHR API. We will also demonstrate how the laboratory could retrieve such a request and related clinical information from the POC server to demonstrate how an existing laboratory could still benefit for a third party FHIR server to obtain structured and computable information. We will also simulate the fulfillment of the laboratory request through the POC server. The richness of the data communicated for this use case, and through the POC server, will be limited to the extent it is modeled in the Core IG. However, we'll also explore the possibility of additional data exchange based on the base FHIR spec when possible. We will also demonstrate how to attach any related artifacts (PDFs for example) to FHIR messages to support existing forms of communication in addition to our attempt to provide a structured form of the same information in the form of FHIR resources or Phenopackets instances.

3.2 Community Use Cases

A Community Use Case describes the specific information needs, data types, and any relevant domain or technical aspects that a stakeholder needs to support their goals. The community use cases, as opposed to the *Driver Use Cases*, will primarily be developed by the wider community, not the project team, and will be retained for future use by the Phenopackets project. The project team will establish a workflow for how to collect and evolve these use cases for the benefit of all stakeholders. These use cases will initially be captured as a GitHub issue in the domain analysis repository to facilitate community discussion and development, and will be merged into this document when mature.

CHAPTER

FOUR

PHENOTYPIC INFORMATION

This section will focus on capturing and analyzing examples of phenotypic information independent of any existing data models. This content will also be an input into the process of identifying and defining important domain entities.

CHAPTER

FIVE

DOMAIN ENTITIES

Important aspects of domain analysis include the identification of relevant domain concepts, the documentation of how people in the domain understand and use those concepts, and what data is collected about them or in relation to them. Specific examples of domain entities are the concepts of person, patient, study subject, diagnosis, signs and symptoms, family, and pedigree. People in the domain have a clear and intuitive understanding of these concepts but they also appropriately adjust their understanding and use of these concepts based on the domain context. This frequently leads to difficulties when it comes to data modeling and the often implicit shift of semantic meaning inhibits interoperability.

This section aims to identify these concepts and represent them as modeling entities. This process usually involves capturing the various ways these concepts exist in the domain, giving clear definitions and making clear distinctions, and creating new entities with new labels when a domain concept might be too general or vague and can imply very different entities for different people in the domain or in different domain contexts.

The process for developing this content will likely start as a glossary of the various concepts. Each concept will then be fully defined and contrasted with other concepts such that each has clear semantic boundaries. In addition, specific relationships between concepts will be established, data elements will be defined, and appropriate terminologies and ontologies will be identified to support the technical development efforts that will follow.

CHAPTER

SIX

PHENOPACKETS

Phenopackets is a well established GA4GH schema for capturing and exchanging structured phenotypic data to support comon research and clinical workflows. In order to achieve the goals of this project, the current Phenopackets model will be examined, including the underlying domain it represents and how technical data is representated, which will enable a rigorous comparison to the FHIR standard. Careful analysis of those two specifications, with specific consideration of their semantic models and data representations, will identify new areas of development. We anticipate that the outcome of this process will be the enhancement of both standards through the delivery of technical guidance and tools to support data integration and interoperability.

The Phenopackets project describes it as:

> *The Phenopacket Schema represents an open standard for sharing disease and phenotype information to improve our ability to understand, diagnose, and treat both rare and common diseases. A Phenopacket links detailed phenotype descriptions with disease, patient, and genetic information, enabling clinicians, biologists, and disease and drug researchers to build more complete models of disease (see Disease for the distinction between disease and phenotypic feature). The standard is designed to encourage wide adoption and synergy between the people, organizations and systems that comprise the joint effort to address human disease and biological understanding.*

CHAPTER

SEVEN

HL7 FHIR

The HL7 FHIR specification was mostly developed for clinical systems and is heavily focused on established clinical workflows. It is not as well developed for the research domain, or for capturing more granular clinical facts such as phenotypic level information. However, FHIR does have some foundational building blocks that could be extended (with FHIR extensions) or could be further developed to accommodate both of these areas. The analysis of FHIR to represent detailed clinical phenotypes and related data will be focused on two main areas.

7.1 FHIR representation

The analysis will examine and document how the FHIR framework represents clinical phenotypic information and related data, including which FHIR resources are most applicable to the representation of those data. Particular attention will be paid to identifying gaps in the FHIR specification that need to be addressed to better support our use cases.

7.2 Phenopackets mapping

This part of the FHIR analysis will specify in detail how the Phenopackets schema maps to FHIR resources. This process will provide the mappings needed for technical development and implementation of a FHIR IG, but it will also help identity possible enhancements to the FHIR and Phenopackets specifications to bring them into better alignment in how they represent the underlying domain entities.

CHAPTER

EIGHT

RESOURCES

This page will contain a listing of resources and tools that are relevant to the phenotypic domain.