

**Clinical Genomics Domain Information Models**

**Clinical Genomics Statement Model Walk-Through**

**DRAFT – for Informative Ballot**

**Last update: 29July2014**

This document is part of an effort to develop a Clinical Genomics Domain Analysis Model (CG DAM). In the September 2014 cycle, two documents will be balloted separately: Clinical Sequencing DAM and Clinical Genomics Domain Information Model (DIM). These are balloted at Informative level. In a later release, the CG DIM and the Sequencing DAM will be merged into a broader CG DAM.

The CG DIM project addresses the lack of alignment across the current HL7 Clinical Genomics specifications (published v3, v2 & CDA specs, and FHIR resources in development). It is presented to the reader in a 'standards-independent' way, i.e., that is not dependent on any specific HL7 jargon. The CG DIM is a set of conceptual models intended to serve as the agreed-upon semantics of clinical genomics data pertaining to individuals.

For more information about the HL7 Clinical Genomics Domain Information (DIM) Model(s) Project (scope, need, criteria, requirements, etc.), please see here: <http://wiki.hl7.org/index.php?title=Clinical_Genomics_Domain_Information_Model(s)_Project>

This document consists of a walk-through of one of the Clinical Genomics DIM models - the Clinical Genomics Statement (CGS) Model. The model itself can be found in the zip file containing this document and also in the above HL7 project wiki site (see above).

**General comments:**

The CGS model represents a single 'genotype-phenotype' association. It can be utilized in higher-level constructs, e.g., messages, documents, etc. where its subject is identified.

The CGS model is designed with the following underlying principles:

* **Core principles:**
  + Explicit representation of a 'genotype-phenotype' association
  + The association should not be pre-coordinated, that is, a phenotype is not represented as part of some genetic variation representation, e.g., a genetic biomarker is an explicit association of two distinct observations and there is no pre-coordination of the interpretation into a single biomarker code
  + The association is a many-to-many relationship, e.g., a genetic variation can be associated with several phenotypes and vice versa
  + It is possible to populate merely the genotype part, thus serving genetic variation representation in general, even if no phenotypic data is available
* **Genetic variations:**
  + Capture various types of data, from known mutations to sequencing-based variations (e.g., somatic mutations in tumor tissues) to structural variations (e.g., large deletions, cytogenetics) to gene expression and so forth
  + Enabling encapsulation of key raw omics data that is the basis of creating this clinical genomics statement
  + The encapsulated data can have links to the full-blown omics data (e.g., NGS sequencing data)
* **Phenotypes:**
  + Distinguish between interpretive phenotype and observed
  + Interpretation is represented as a distinct observation with its own attributes such as time, method, performer, etc.
  + Links to knowledge sources used for the interpretation
* **Specimen**
  + Molecular specimen and genomic source class
  + Anatomic pathology specimen
* **Other data types:**
  + Optional indications representing the triggers for performing genetic testing or making genomic observations

**Model Walk-through:**

Recent changes (29July2014)

\* Changed the class name GeneticVariation to GeneticObservation, in order to imply the scope of this place holder as not necessarily restricted to nucleotide variation data and when used for variants can also represent wild types

\* Changed the class name AssociatedOmicsData to SupplementingGeneticData, in order not to restrict this placeholder for omics data only and also in order to avoid using the term association except for the main concept in this model, i.e., the 'genotype-phenotype' association

\* Added a class GeneticObservationComponent linked to the GeneticObservation in an aggregation mode, in order to represent data items that are inherent part of the core genetic observation, in a post-coordinated manner; this class does not hold an independent observation and this has merely type-value attribute pair and has no attributes such id, time, etc.

\* Note that the link of GeneticObservation to SupplementingGeneticData was changed from aggregation to simple association

\* Changed class name AssociatedPhenotypicData to SupplementingPhenotypicData

**++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++**

**Class GeneticObservation**

**Description:**

Representation of any type of DNA data, including genetic loci in the size of a gene, e.g., genes, biomarkers, point DNA mutations, etc. as well as structural variations, e.g., large DNA deletions or duplications, etc. Note that omics data related to this variant, e.g., RNA, proteins, expression, etc., should be represented using the linked SupplementingGeneticDataclass.

**Attributes:**

* **id:** a globally-unique id of this GeneticObservation
* **type:** the type of genetic variation, e.g., mutation, SNP, large deletion, HGVS-described variation, etc.
* **value:** the actual variation**,** e.g., if type is **"**DNA Sequence Variation" then value can be "109G>A" using "HGVS nomenclature" (specified as part of the value attribute)
* **performer:** e.g., unique id of the genetic lab that made this observation
* **method:** code and/or description of the methods used to obtain the genetic variation observation
* **observationTime:** the effective time for the subject of this observation

**++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++**

**Class GeneticObservationComponent**

**Description:**

Representation of components of the core genetic observation in a post-coordinated manner, i.e., breaking it down to its primitives in order to disambiguate data held by the GeneticObservation class, e.g., an HGVS string.

**Attributes:**

* type: the type of genetic observation component, e.g., variation location, variation length, etc.
* value: the actual component observation, e.g., if type is variation base length then value is the actual length

**++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++**

**Class SupplementingGeneticData**

**Description:**

Representing any downstream data pertinent to the core genetic observation (including omics data), e.g., RNA, proteins, expression, etc.

**Attributes:**

* **id:** a globally-unique id of this observation
* **type:** the type of observation, e.g., characteristics of the resulting protein
* **value:** the actual observation
* **performer:** e.g., unique id of the genetic lab that generated this data

**++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++**

**Class KeyRawOmicsData**

**Description:**

Representation of key data extracted from mass raw genomic data, e.g., an exon's sequence that is considered key to this clinical genomics statement, is extracted from some larger sequencing (up to whole exome squening) and encapsulated within this statement structure. Another example: certain lines of a VCF file (preferably its 'clinical-grade' version). The encapsulated data should contain references to the full-blown raw omics data.

**Attributes:**

* **encapsulatedData:** a placeholder for the key raw omics data, e.g., DNA sequences, variants calls, expression data, etc.Could be any type of data encoding: binary, text, XML, etc.
* **reference:** a pointer to the repository containing the full-blown data sets out of which these key data items were extracted from
* **type:** designating the notation used in the encapsulatedData attribute, e.g., VCF, FASTA, etc.
* **schemaReference:** a pointer to the schema governing the representation of the data in the **encapsulatedData** attribute

**++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++**

**Class RawOmicsDataSource**

**Description:**

Representing the repository where the full-blown data set could be found.

**++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++**

**Class *Phenotype* (abstract)**

**Description:**

Representation of any data considered phenotype of the source genetic observation. For each genetic observation (e.g., variation), there could be multiple phenotypes, however, each phenotype object should represent a single phenotype.

**Attributes:**

* **id:** a globally-unique id of this observation
* **type:** the type of phenotypic observation, e.g**.,** drug responsiveness
* **value:** the actual observation(should be drawn from dedicated ontologies, e.g., Human Phenotype Ontology, <http://www.human-phenotype-ontology.org/>)
* **status :** describes the 'clinical genomics' status of this phenotype, e.g.:

- Established

- Provisional

- Putative

- Preliminary

**++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++**

**Class SupplementingPhenotypeData**

**Description:**

The SupplementingPhenotypeData class can add details, e.g., drug resisted and context, e.g., age, etc.

**Attributes:**

* **id:** a globally-unique id of this observation
* **type:** the type of associated phenotypic observation, e.g., age, antibiotics taken, etc.
* **value:** the actual observation

**++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++**

**Class ObservedPhenotype (extending *Phenotype*)**

**Description:**

This is a sub-class of Phenotype, representing a phenotype observed in the subject, e.g., if the somatic mutation is known to be the reason of positive responsiveness to some drug, then that responsiveness is an observed phenotype. For example, consider the following case where Kobayashi et al. (2005) reported a patient with advanced nonsmall cell lung cancer having an EGFR somatic mutation who has been gefitinib-responsive for 2 years of complete remission during treatment with gefitinib, but then had a relapse due to a second EGFR mutation (see Kobayashi, S., Boggon, T. J., Dayaram, T., Janne, P. A., Kocher, O., Meyerson, M., Johnson, B. E., Eck, M. J., Tenen, D. G., Halmos, B. EGFR mutation and resistance of non-small-cell lung cancer to gefitinib. New Eng. J. Med. 352: 786-792, 2005. [PubMed: [15728811](http://www.ncbi.nlm.nih.gov/pubmed/15728811)]. At the end of those two years, the first EGFR mutation should be associated with an observed phenotype while the second mutation should be associated with an interpretive phenotype (see below).

**Attributes:**

* **clinicalTime:** the effective time for the subject of this observation
* **ageOfOnset:** use this attribute in cases where dates are not available(e.g., in clinical genomics statements pertaining to a relative of a patient in a family health history)
* **referenceToHealthRecord:** a pointer to an information system (e.g., EHR system) holding the broader contextual data of this observation

**++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++**

**Class InterpretivePhenotype (extending *Phenotype*)**

**Description:**

This is a sub-class of Phenotype, representing a phenotype that is likely to be manifested in the subject, e.g, the somatic mutation is likely to cause resistance to some drug (see above).

**Attributes:**

* **method:** the method by this interpretation was made (e.g., an identified algorithm, a rule engine, a clinical decision support application/service, etc.)
* **interpretation Time:** the time this interpretation was made available for this Clinical Genomics statement instance)
* **performer:** unique id of the entity making this interpretation (e.g., clinical decision support application/service)

**++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++**

**Class InterpretationEvidenceSource**

**Description:**

Representing the knowledge bases that served as source for making the interpretation.

**++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++**

**Class AnatomicPathologyFinding (extending ObservedPhenotype)**

**Description:**

An anatomic pathology finding that is the phenotype of some genetic variation.

**Attributes:**

* **status:** the status of this finding (using AP reporting vocabularies)
* **referenceToAPReport:** a pointer to an AP report consisting of this AP finding

**++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++**

**Class AnatomicPathologySpecimen**

**Description:**

An optional reference to a detailed description of the specimen used for the finding described in the AnatomicPathologyFinding class.

**Attributes:**

* **id:** a globally-unique id of the AP specimen (e.g., using the HL7 Specimen ID)

**++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++**

**Class Indication**

**Description:**

The indication to performing genetic testing, e.g., genetic-related hearing loss diagnostics. Note that this is mainly relevant in clinical environments.

**Attributes:**

* **id:** a globally-unique id of this indication
* **type:** the type of indication (e.g., diagnostic, carrier, etc.)
* **value:** the actual indication (preferably drawn from controlled vocabularies)

**++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++**

**Class MolecularSpecimen**

**Description:**

**Attributes:**

* **id:** a globally-unique id of this specimen
* **type:** the type of molecular specimen (e.g., DNA)
* **genomicSourceClass:** the source of the specimen being analyzed, e.g., germline for inherited genome, somatic for cancer genome and prenatal for fetal genome