NMDP HML Gateway Programmatic Validation

09-April-2020

HML Gateway's /validate endpoint performs schema, internal schema, and semantic validation. Incoming HML undergoes an initial sanity check, a structure/schema validation, and a rules/semantics validation. Internal schema validation ascertains gene family consistency.

Sanity Check

- The submitted HML cannot be null or empty
- The submitted HML's HML ID cannot be null
- The submitted HML's HML ID root value cannot be null or empty
- The submitted HML must have at least one sample
- The submitted HML version must be the currently supported version, 1.0.1

Schema Validation

This isn't simply binding the incoming HML to a java object, HML gateway will produce a list of schema validation errors. This list can be used without having to read the schema file manually to figure out where the HML is malformed.

Internal Schema Validation

The HML's 'gene-family' typing element attributes must all have the same value

Semantic Validation

- The HML must be correctly bound to HML Gateway's Java class. The Java HML object cannot be null
- The HML reporting center value cannot be null or empty
- The HML project name must be valid (see section on project validation)
- The HML version cannot be null or empty
- The HML ID cannot be null or empty
- The HML must have no duplicate sample IDs
- An HML sample's center code must be lower or uppercase alpha, or numeric, and have no more than 50 characters
- All HML submission samples must have the same gene family
- Each reference sequence start position cannot be greater than the reference sequence end position
- Each variant start position cannot be greater than the variant end position
- Each HML sample must have at least one typing
- Each typing date cannot be null
- Each typing date cannot be set in the future of the submission time
- Each typing method cannot be empty or null

- Each typing method must be either SSO, SSP, SBT-Sanger, or SBT-NGS (i.e. must be one of the following lowercase values: "sso", "ssp", "sbt-sanger", "sbt-ngs")
- If the typing method test-id is present, test-id-source must be also
- If the typing method test-id-source is present, test-id must be also
- Finally, each HML instance must pass additional semantic rule validation for HLA, or KIR

Project Name Semantic Validation

- The HML submission project name cannot be null or empty
- For NMDP Form 22 submissions, the project name must be "REG_CONFIRM_DNR"
- For NMDP Form 117 submissions, the project name must be "REG_CONFIRM_RCP"
- For CIBMTR form 2005 submission, project name is a research project name, i.e. it has a "RES_" prefix and is comprised of only lower and uppercase letters, or numbers (0-9), or the "-" character. (see separate instructions for CIBMTR Form 2005 submissions).

HLA Semantic Validation

- An HML sample requires at least one allele assignment
- An HML sample's allele assignment date cannot be null
- An HML sample's allele assignment date cannot be set in the future of the submission time
- An HML sample's haploid type cannot be equal to "0000". Will be reported as "FAILED-RESULT"
- An HML sample's GL String cannot represent more than one HLA locus
- An HML sample's locus cannot be reported more than once
- An HML sample's GL String cannot contain both a text value and a URI
- A GL String must be Uppercase alpha, numeric, or one of the following characters: +, ^, ~, |, *, :, -, (or a character that does not match the \w regex character)
- A GL String must contain at least one fully qualified allele name (i.e. this GL String must contain at least one "*" character)
- A GL String cannot start with a +, /, |, ~, or ^ character
- A GL String cannot end with a +, /, |, ~, or ^ character
- A GL String allele name must match the following regex:
  "(HLA\-)(A|B|C|DRB1|DRB3|DRB4|DRB5|DQB1|DPB1|DQA1|DPA1)\s?(*)\s?(\d{2,5})(:\d{2,5}[A-Z]?)+"
- An allele name cannot be less than one character
- Genotype lists are not supported, so they are ignored
- An allele assignment must have at least one haploid or GL string
- A non-even number of haploids will generate a warning
- An HML sample's haploid type cannot be blank
- HML sample haploids cannot have more than two alleles for the same locus even across multiple typings
- A GL String cannot have more than two alleles for the same locus (i.e. for any substring surrounded by | characters, no more than one + or ~ character delimiting 2 allele substrings)

KIR Semantic Validation

- Allele assignment haploids are not allowed for gene-family KIR
- A KIR GL String cannot contain both a text value and a URI
- Typing allele assignment properties must have valid supply order line IDs