1.hml

Description:

Root element of the document identifying it as an HML message. Must contain the version of HML that the modeled data in this document uses.

Children:

- property (optional, qty: 0 or more)
- hmlid (optional, qty: 1)
- reporting-center (optional, qty: 1) - Required for NMDP samples
- sample (required, qty: 1 or more)
- typing-test-names (optional, qty: 0 or more)

Attributes:

- version: Version of HML the document follows (required)
- project-name: Name of the typing project (optional)

Expected to be '1.0.1' to use this version of the HML schema.

2.reporting-center

Description:

This element identifies the entity/organization sending this HML data.

If included, must contain a unique ID identifying the sender as well as
a context which defines to whom the ID is meaningful or the source
of the ID.
This element is required for NMDP transactions and if context is not
included, is assumed to be "NMDP".

Attributes:

- reporting-center-id: (required) Unique id of reporting center
like "789".
- reporting-center-context: (optional) Source of the reporting center ID
like "NMDP". To whom the ID is meaningful.

3.hmlid

Description:

Specifies a unique identifier for this HML document. This id follows
the HL7 standard for uniqueness using a two-part key. 'root' is the
unique organization identifier publicly registered for your organization.
'extension' is the unique document id managed internally for your
organization, but must be unique and identify this specific HML document.
Together root and extension guarantee global uniqueness.

http://www.oid-info.com/faq.htm
http://www.hl7.org/oid/index.cfm

Attributes:

- root: Unique publicly registered identifier for the HML creator's organization.
(ex: NMDP HL7 id is "2.16.840.1.113883.3.1470")
Format is expected to be a string of digits and dot delimiters. (required)
- extension: A unique document identifier managed internally by the organization specified in 'root'. Can be any alpha-numeric format desired by the organization. (ex: "hml-0.9.7-123456789.23a")
(optional and must be at least 1 non-whitespace character long)
NOTE - If extension is NOT included, the unique document identifier is expected to be appended at the end of the root identifier above in accordance with HL7 practices.

3.1root

Description:

Unique publicly registered identifier for the HML creator's organization.
This can be an HL7 compliant field (ex: NMDP is "2.16.840.1.113883.3.1470")
Format is expected to be a string of digits and dot delimiters. (required)

3.2extension

Description:

A unique document identifier managed internally by the
organization specified in 'root'. Can be any alpha-numeric
format desired by the organization. (ex: "hml-0.9.7-123456789.23a")
(required and must be at least 1 non-whitespace character long and
guarantee uniqueness)

4.typing-test-names

Description:

Specifies a list of test names internally referenced by an "sso"
element or an "ssp" element. It wraps a list of "typing-test-name"
elements, which contain the test identifiers.

Children:

- typing-test-name: (required, qty: 1 or more)

Attributes:

- test-id: (required, qty: 1) Reference identifier (unique string)
internal to the document used for referencing the list of
tests contained here with a typing-method.

5.typing-test-name

Description:

Specifies a single test name contained in a referenced "typing-test-names" list. Typing tests may be referenced by other elements including SSO, SSP, etc.

Attributes:

- name: (required) Fully qualified test name (ex: "L999.K1.V1.A9F-S11", "L999.K1.V1.SSP12345")

6.property

Description:

Allows the optional inclusion of key-value pairs (not defined explicitly by the schema) without the need to extend or change the schema. Allows children to be extensible for custom use. Any information contained in this element must be fully understood by the message recipient.

Attributes:

- name: (required) "key" in the name-value pair
- value: (optional) "value" in the name-value pair

7.sample

Description:

Encloses the genotyping data pertaining to a particular sample. It may contain multiple typing elements (for instance, one for each locus).

Children:

- property: (optional, qty: 0 or more)
- collection-method: (optional, qty: 0 or 1) - Free-form text such as "swab", "filter paper", and "blood aliquots".
- typing: (required, qty: 1 or more)

Attributes:

- id: (required) Identifier for the sample (ex: "1234-5678-9", "123456789")
- center-code: (optional) Center code of the sample's origin (donor center, transplant center, etc.)

8.typing

Description:

Encapsulates the primary data from a genotyping method with an optional genotyping result (allele-assignment) determined from the primary data and/or optional consensus sequences.

Children:

- property: (optional, qty: 0 or more)

- allele-assignment: (not required, qty: 0 or more) Also known as interpretation.

- typing-method: (required, qty: 1 or more)

The 'typing-method' element encapsulates methods such as sso, ssp, sbt-sanger, and sbt-ngs.

- consensus-sequence: (optional, qty: 0 to many) Consensus data for the results reported under typing-method.

Also allows an optional "property" element that may have nested/custom use data related to the interpretation.

Attributes:

- gene-family: Represents the gene evaluated in this typing report, e.g. "HLA" or "KIR" (required)

See: http://www.genenames.org/genefamilies for examples.

- date: Typing/testing date for this sample (required)

9.allele-assignment

Description:

Specifies the genotyping call at the most specific level possible. This call can be represented within haploid elements or using gl-resources. When reporting data using haploid, typical use is one or two haploid elements for a particular locus, but possibly more if multiple loci are covered (ex: two DRB1 haploids + one DRB3 haploid).

Children:

- property: (optional, qty 0 to many) Custom use properties.
- haploid: (optional, qty: 1 or more)
- genotype-list: (optional, qty: 0 or more)
- glstring: (optional, qty: 0 ro more)

Allows an optional "property" element that may have nested/custom
use data related to the allele-assignment/interpretation.

Expects at least one of haploid, genotype-list, or glstring.

Attributes:

- date: Date on which the typing was carried out, or on which the
final call was determined. Format can be either ISO-8601 or
"YYYY-MM-DD". (required)
- allele-db: Database or other source for nomenclature used in the
interpretation. (ex: "IMGT-HLADB") (optional, but required for NMDP use)
- allele-version: A specific version of the allele-db (ex: "3.18.0").
(optional, but required for NMDP use)

10.haploid

Description:

Specifies one-half of a full typing at a particular locus. Must

conform to the database specified in allele-assignment/interpretation.

Attributes:

- locus: Locus (ex: "HLA-A", "HLA-DRB1") (required)
- method: Typing method used (ex: "DNA", "SER") (required)
- type: Allele/code level type (ex: "01:01", "01:AB") (required)

11.glstring

Description:

Specifies a resource in Genotype List String (GL String) format for the
interpretation of a typing result, or a URI identifying a resource in
GL String format. For more details about the format and use of GL Strings,
see (http://www.ncbi.nlm.nih.gov/pubmed/23849068)

* glstring is expected to EITHER contain inline character data OR a
URI reference to a location that defines/specifies the glstring data.

Attributes:

- uri: Specifies a URI identifying a resource in GL String format for the
interpretation of a typing result. For more information about the format
and use of GL Strings, see (http://www.ncbi.nlm.nih.gov/pubmed/23849068).
(optional)

Data:

- resource in GL String representation (string, required)

A genotype-list represents a full unambiguous list of possibilities for the typing of a sample (NOTE: This element and its children were deprecated in HML 1.0). The values of the elements in this genotype-list (each allele element) should conform to the nomenclature specified by the allele-assignement/interpretation.

Children:

- diploid-combination (required, qty: 1 or more)

12.diploid-combination

Description:

A diploid-combination element is one possibility value in a genotype-list (NOTE: This element and its children were deprecated in HML 1.0). There may be either one or two locus-block child elements, depending on whether the data provided in this diploid-combination covers one or two chromosomes.

Children:

- locus-block (required, qty: 1 or 2)

13.locus-block

Description:

A locus-block element allows allele-list elements to be grouped
together to mean one allele-list is a possibility if and only if all
others are (NOTE: This element and its children were deprecated in HML 1.0).
This is useful, for example, in the case when listing
HLA-DRB1 alleles next to the corresponding HLA-DRB3 alleles that are
relevant in only some cases (example in comments).
-->

Children:

- allele-list (required, qty: 1 or more)

14.allele-list

Description:

An allele-list element is a representation of the list of allele
possibilities for a genotype (NOTE: This element and its children were
deprecated in HML 1.0). NMDP has historically used allele codes
in combination with allele families to represent this.

Children:

- allele (required, qty: 1 or more)

15.allele

Description:

An allele element specifies a single allele: it should be given in
LOCUS*NAME format and names must be at allele-level resolution (NOTE:
This element and its children were deprecated in HML 1.0). The
value must conform to the nomenclature specified in the interpretation.
-->

Attributes:

- present: Indicates the presence or absence of this allele. A value
of "N" can be used to indicate that a particular allele was tested
for and found not to be a possibility. A value of "U" (untested)
indicates that the given allele was not tested for. The default
value is "Y". [Y|N|U] (required, qty: 1 or more)

15.1present

Description:

An enumerated type indicating the presence or absence of an allele.

N: a particular allele was tested for and found not to be a

possibility

U: the given allele was not tested for

Y: (default) the given allele was tested for and found to be a

possible result

16.typing-method

Description:

Must include at least one of sso, ssp, sbt-sanger, and/or sbt-ngs

for the 'typing-method' element.

17.sso

Description:

Specifies an SSO (sequence-specific oligonucleotide) test that was

done for this sample. Kit information and scores must be identified

to allow for later test reinterpretation. For NMDP, a corresponding

typing-test-names/typing-test-name structure is expected in this

same HML document.

Children:

Allows an OPTIONAL "property" element that may have nested/custom

use data related to this typing-method.

Attributes:

- locus: locus for multi-locus targets (optional)
- test-id: Test ID as registered with the test-id-source.
- test-id-source: A formal or formal test registry location. For example, this could be the NCBI GTR (specified as "GTR"), NMDP for tests registered directly with NMDP (specified as "NMDP), etc. (required if test-id is used)
- scores: The results of the SSO test, specified as one string (ex: "118111100181")

NMDP allows the following test-id-source values:
(Note that this may change in future versions)

* gtr: ID of kit registered with the NCBI Genetic Testing Registry. (Preferred)
* nmdp-refid: ID of kit registered with NMDP. The cardinal sequence numbers of the registered probes in the kit will determine the score order.
* probe-name: Fully qualified probe name. If this attribute is used, the scores attribute must contain exactly one score. (ex: "L0999.K1.V1.A9F-S11")

18.ssp

Description:

Specifies an SSP (sequence-specific primer) test that was performed for this sample. Kit information and scores must be identified to allow for

later test re-interpretation.

Children:

Allows an OPTIONAL "property" element that may have nested/custom use data related to this typing-method.

Attributes:

- locus: locus for multi-locus targets (optional)
- test-id: Test ID as registered with the test-id-source.
- test-id-source: A formal or formal test registry location. For example, this could be the NCBI GTR (specified as "GTR"), NMDP for tests registered directly with NMDP (specified as "NMDP), etc. (required if test-id is used)
- scores: The results of the SSP test, specified as one string (ex: "118111100181")

NMDP allows the following test-id-source values:
(Note that this may change in future versions)

* gtr: ID of kit registered with the NCBI Genetic Testing Registry. (Preferred)
* nmdp-refid: ID of kit registered with NMDP. The cardinal sequence numbers of the registered probes in the kit will determine the score order.
* probe-name: Fully qualified probe name. If this attribute is used, the scores attribute must contain exactly one score. (ex: "L0999.K1.V1.A9F-S11")

19.sbt-sanger

Description:

Describes an SBT (sequence-based typing) that was performed using a

Sanger technique.

Children:

- amplification (required, qty: 1)
- sub-amplification (not required, qty: 0 or more)
- gssp (not required, qty: 0 or more)

Also allows an optional "property" element that may have nested/custom

use data related to this typing-method.

Attributes:

- locus: The locus for which the SBT was performed. (optional)
- test-id: Test ID as registered with the test-id-source.
- test-id-source: A formal or formal test registry location. For

example, this could be the NCBI GTR (specified as "GTR"),

NMDP for tests registered directly with NMDP (specified as

"NMDP), etc. (required if test-id is used)

NMDP allows the following test-id-source values:

(Note that this may change in future versions)

* gtr: ID of kit registered with the NCBI Genetic

Testing Registry. (Preferred)

* nmdp-refid: ID of kit registered with NMDP. The cardinal sequence numbers of the registered probes in the kit will determine the score order.

* probe-name: Fully qualified probe name. If this attribute is used, the scores attribute must contain exactly one score. (ex: "L0999.K1.V1.A9F-S11")

20.amplification

Description:

Identifies the amplification primer used for SBT-Sanger, and the resulting sequence from using it.

Attributes:

- registered-name: Identifies the amplification primer. Must be recognized by the message recipient. (string, required)

Data:

- sequence: IUPAC nucleotide sequence (string, required)

21.sub-amplification

Description:

Identifies sub-amplification primers. These primers are used to resolve ambiguities and may be used either concurrently with or after the amplification step.

Attributes:

- registered-name: Identifies the amplification primer. Must be recognized by the message recipient. (string, required)

Data:

- sequence: IUPAC nucleotide sequence (string, required)

22.gssp

Description:

Describes the Group Specific Sequencing Primer used.

Attributes:

- registered-name: Identifies the amplification primer. Must be recognized by the message recipient. (string, optional)
- primer-sequence: PCR primer sequences used to amplify a polymorphic region of sequences. (string, optional)
- primer-target: If the primer sequence is proprietary (or otherwise unable to be explicitly specified), the primer sequence can be imputed from the gssp result. This imputed primer sequence is specified as the

primer-target. (string, optional)

Data:

- Resulting nucleotide sequence from the GSSP used. (string, required)

23.sbt-ngs

Description:

Describes an NGS (next-generation sequencing) event that was performed.

Children:

- property: (optional, qty, 0 or more)
- raw-reads (optional, qty: 0 or more)

Also allows an optional "property" element that may have nested/custom
use data related to this typing-method.

Attributes:

- locus: The locus for which the SBT was performed. (optional)
- test-id: Test ID as registered with the test-id-source.
- test-id-source: A formal or formal test registry location. For
example, this could be the NCBI GTR (specified as "gtr"),
NMDP for tests registered directly with NMDP (specified as
"NMDP"), etc. (required if test-id is used)

NMDP allows the following test-id-source values:

(Note that this may change in future versions)

* gtr: ID of kit registered with the NCBI Genetic Testing Registry. (Preferred)

* nmdp-refid: ID of kit registered with NMDP. The cardinal sequence numbers of the registered probes in the kit will determine the score order.

* probe-name: Fully qualified probe name. If this attribute is used, the scores attribute must contain exactly one score. (ex: "L0999.K1.V1.A9F-S11")

24.sequence

Description:

The DNA alphabet consists of primary nucleotides (A, C, G, T).

Wildcard IUPAC nucleotides (M, R, W, S, Y, K, V, H, D, B, X, N) may be used if they are acceptable in the context in which they appear. The default is to use all upper case letters.

The full specification of the IUPAC codes may be found here:

(http://nar.oxfordjournals.org/content/13/9/3021.short)

Cornish-Bowden A. Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. Nucleic Acids Res. 1985; 13:3021-3030.

Attribute:

- xs:anyAttribute: Custom use attribute for additional sequence information. (optional)

Data:

- Sequence in the DNA alphabet (string, required)

25.simple-bases

Description:

A string of simple nucleotide bases. These are A, C, G and T (DNA). Used to define reference bases and alternate bases for sequence variants.

26.iupac-bases

Description:

Nucleotide bases representing sequence ambiguity.

Primary nucleotides: A, C, G, T (DNA).

"Wildcard" nucleotides: M, R, W, S, Y, K, V, H, D, B, X, N.

Wildcard nucleotides may be used if they are acceptable in the context in which they appear. The default is to use all upper case letters.

The full specification of the IUPAC codes may be found here:

(http://nar.oxfordjournals.org/content/13/9/3021.short)
Cornish-Bowden A. Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. Nucleic Acids Res. 1985; 13:3021-3030.

The bases of the sequence string are restricted to the upper and lower case versions of the nucleotides specified above.

Data:

- Nucleotide sequence in DNA alphabet (string, required)

27.consensus-sequence

Description:

Describes a sequence that is the result of an alignment or assembly of shorter sequence reads generated by an NGS platform.

Wraps one or more reference database definitions and one or more consensus-sequence-blocks.

Children:

- reference-database (required, qty: 1 or more)
- consensus-sequence-block (required, qty: 1 or more)

28.consensus-sequence-block

Description:

A consensus sequence contains a sequence of IUPAC nucleotides and novel variants.

The nucleotides are specified in the DNA alphabet.

The DNA alphabet consists of primary nucleotides (A, C, G, T).

Wildcard IUPAC nucleotides (M, R, W, S, Y, K, V, H, D, B, X, N) may be used if they are acceptable in the context in which they appear. The default is to use all upper case letters.

The full specification of the IUPAC codes may be found here:
(http://nar.oxfordjournals.org/content/13/9/3021.short)
Cornish-Bowden A. Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. Nucleic Acids Res. 1985; 13:3021-3030.

Children:

- sequence: (required, qty: 1) Nucleotide data for the consensus block.
- variant: (optional, 0 or more) If region-match is false, variant is expected to refer to the novel-variants.
- sequence-quality: (optional, qty: 0 or more) A score for a sub-sequence specified by start and end (includes 'start', excludes 'end') that indicates the quality of the read.

Attributes:

- reference-sequence-id: (required) Reference to a unique reference-sequence
defined in this document under "consensus-sequence". IDREF
must exactly match the ID for the reference-sequence.
- start: (required) Start position of a targeted region on contig,
0-based or space-counted coordinate system, closed-open range
- end: (required) End position of a targeted region on contig,
0-based or space-counted coordinate system, closed-open range
- strand: (optional) String value (eg. one of "-1", "1", "-", "+");
defaults to "+" if unspecified
- phasing-group: Phasing group identifier - DEPRECATED. Use "phase-set".
- phase-set: Phase set identifier (string, optional)
- continuity: (optional) True if this represents a continuous read, false
if not continuous.
- expected-copy-number: (optional) Integer for how many copies of
the sequence block were expected (0 to n).
- description: (optional) Text description of the targeted region, like "HLA-A exon 3"
- xs:anyAttribute: Custom use attribute for additional sequence
information. (optional)

29.position-type

Description:

30.sequence-quality

Description:

Defines the quality for a range within the consensus sequence block.

Attribute:

- start: (required) Sequence start position for quality - inclusive
- end: (required) Sequence end position for quality - not inclusive
- quality-score: (required) Value indicating the quality of the consensus sequence.

31.quality

Description:

Used to indicate a quality-score for a variant/consensus sequence block.

32.reference-database

Description:

A database reference for the consensus sequence blocks included in the
HML document.

Each reference-database may have 1 or more reference-sequence definitions.

Examples:

Genome Reference Consortium:

description="Genome Reference Consortium (GRC)"

version="GRCh38.p1"

availability="public"

curated="true"

uri="http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human">

IMGT/HLA:

description="IMGT/HLA Database"

version="3.18.0"

availability="public"

curated="true"

uri="http://www.ebi.ac.uk/ipd/imgt/hla">

KIR:

description="IPD KIR Database"

version="2.5.0"

availability="public"

curated="true"

uri="http://www.ebi.ac.uk/ipd/kir">

Children:

- reference-sequence: (required, qty: 1 or more) Reference sequence for this database.

Attribute:

- name: (optional) Name for this database.

- description: (optional) Description of this database reference.
- version: (optional) Version of this reference database.
- availability: (required) Defines how this reference database is available ("public", "private", "none").
- curated: (optional) "true" if curated, "false" otherwise.
- uri: External reference for this database.

A sequence reference for the consensus sequence blocks included in the HML document, associated with a reference-database.

The reference-sequence id must be document-unique and is referenced by consensus-sequence-block elements via the required reference-sequence-id attribute.

Examples:

Genome Reference Consortium:

```
name="HSCHR6_MHC_MCF_CTG1"
start="0"
end="4827813"
accession="GL000254.2"
uri="http://www.ncbi.nlm.nih.gov/nuccore/GL000254.2" />
```

IMGT/HLA:

```
name="HLA-A*01:01:01:01"
start="0"
end="3053"
accession="HLA00001"
uri="http://www.ebi.ac.uk/Tools/dbfetch/dbfetch?db=imgthla;id=HLA00001" />
```

Attribute:

- id: (required) Unique reference for this
database/sequence combination which is referred to
in each consensus-sequence-block.
*Note: XML 'ID' type must begin with a non-symbol,
non-digit, alphabetic character.
- name:
- description: (optional) Description of this database reference.
- start: (required) Start position of a targeted region on contig,
0-based or space-counted coordinate system, closed-open range
- end: (required) End position of a targeted region on contig,
0-based or space-counted coordinate system, closed-open range
- accession: (optional)
- strand: (optional) String value (eg. one of "-1", "1", "-", "+");
defaults to "+" if unspecified
- uri: External reference for this database.

33.variant

Description:

A variant needs to be included for a sequence if consensus-sequence-block
doesn't match a known database, meaning there is some ambiguity in the
submitted sequence of nucleotide bases.

Children:

- variant-effect: (optional) Effect of this variation from the published sequence.

Attribute:

- id:
- name:
- start: Variant sequence start position - ('0' based).
- end: Variant sequence end position.
- reference-bases: The nucleotide bases from which the reported sequence differs. (required, A/G/C/T string)
- alternate-bases: The nucleotide bases to substitute for the reported sequence. (required, A/G/C/T string)
- quality-score: Quality of the variant (optional).
- filter: Values 'PASS' or 'FAIL' as used in VCF format.
- uri: External reference for this variant.

34.variant-effect

Description:

A child of "variant", defines the effect of this variation from the published sequence.

Sequence Ontology (SO) variant effect terms are specifications of the sequence_variant (http://sequenceontology.org/browser/current_svn/term/SO:0001060) term.

HGVS (Human Genome Variation Society) information can be found here:

http://www.hgvs.org/mutnomen/disc.html

Additional attributes may be used to provide more information on the effect, for example severity, POLYPHEN prediction, SIFT score, etc.

Attribute:

- term: (required) Sequence Ontology (SO) term describing the effect, e.g.
"synonymous_variant" (http://sequenceontology.org/browser/current_svn/term/SO:0001819)
"missense_variant" (http://sequenceontology.org/browser/current_svn/term/SO:0001583)
- hgvs: (optional) Human Genome Variation Society variant effect description, e.g.
ENST00000288602.3:c.83T>A
ENSP00000288602.1:p.Val28Glu
- uri: (optional) External reference for this variant effect.

35.raw-reads

Description:

Reports the raw sequence reads generated by an NGS platform. Because various platforms report reads in various formats, the platform must be specified. Since this data is quite large even for relatively small regions of the genome, this information must be linked to using an external URI.

Attributes:

- uri: An external link to the raw reads. (required)
- format: Identifies the format of the data located at the URI. (required)

- paired: true/false (default) (required)
- pooled: true/false (default) (required)
- availability: public|private|permission (optional)
- adapterTrimmed: true/false (default) (required)
- qualityTrimmed: true/false (default) (required)