LRG XML schema 1.8

This new schema is scheduled to be released by the end of October

The LRG schema is defined in RelaxNG format in the file "LRG.rnc" - this is the latest version. This RelaxNG file is used as the reference schema when building new LRG XML files, and is also used to validate existing LRG files. The order, hierarchy and content of the elements within the schema must be preserved in the LRG XML files. Elements not described in the schema will not be valid in an LRG.
All published LRGs should validate successfully against this schema.

The LRG XML schema is divided into two main sections:

Fixed annotation

The fixed annotation is, as its name suggests, fixed and permanent. Once an LRG is created, it is intended that the fixed annotation section remains unaltered for the lifetime of an LRG. If significant changes must be made, then a new LRG will be created. The idea is that this section contains the core information about the LRG. Here is some info about the key elements in the fixed annotation section:

- **id –** this is the identifier for the LRG, and should be in the format LRG_[number], e.g. LRG_1, LRG_24
- **sequence_source** – the accession number of the RefSeqGene on which the LRG is based. The genomic reference sequence of the LRG is identical to this RefSeqGene
- **organism –** all LRGs will be for human at first; this element allows for future expansion into other species
- **source –** this is an instance of the standard source element and contains source information for the sequence, along with contact details
- **mol_type –** a description of the type of sequence e.g. "dna"
- **creation_date –** date of the creation of the first version of the LRG
- **sequence –** the agreed genomic DNA sequence for the LRG. By default this sequence will have 5000bp 5' and 2000bp 3' of the main transcript sequence. This will be extended if necessary to include all transcripts encoded by the gene specified in **lrg_locus** or if requested by the community. It should contain no characters other than [ACGT]
- **transcript –** the fixed annotation section will contain one or more transcripts, each with their own element and child elements. Each transcript must have a name (e.g. "t1", "t2"), and a start and end coordinate. All coordinates, unless otherwise stated, are relative to the LRG genomic sequence, and 1-indexed - i.e. start=1 would be the first basepair of the LRG genomic sequence
  - **coordinates –** coordinates of the specified sequence relative to the LRG. The strand of the specified feature is also included. The gene specified by **lrg_locus** is always transcribed by the positive strand, specified as strand=1.
  - **cdna –** contains a sequence element as above that holds the unprocessed, spliced cDNA sequence (UTRs included, introns removed)
  - **coding_region –** this element represents the coding region (or CDS) of the cDNA. The **coding_region** element has a child element which contains the translated protein sequence. Each transcript element can have zero or more translated proteins. The **coding_region** element can also contain the optional child element **translation_exception** which will contain any exceptions to the usual codon usage (for example see LRG_417 - <translation_exception codon="523"><sequence>U</sequence></translation_exception>)
  - **exon –** there then follows one or more exon elements, between each of which will be an intron element. The RelaxNG schema is designed to unambiguously define exons in a transcript. Each exon is defined by start and end coordinates on the LRG, rather than any numbering system, which could differ between laboratories. Exons have LRG specific labels. Each exon element can contain three sets of coordinates, containing the coordinates relative to the LRG sequence, the cDNA sequence and the peptide sequence.
  - **intron –** these are found in between exon elements. They have only one attribute, phase, which can take a value of 0, 1 or 2. This represents the location of the exon/intron boundary in terms of its location in the three basepair codon, according to the following system:
    - **0** - intron falls between codons
    - **1** - intron falls between 1st and 2nd base of codon
    - **2** - intron falls between 2nd and 3rd base of codon

Updatable annotation

The updatable annotation contains one or more annotation set elements, each of which contains a set of annotation from one source. Currently the updatable section of each LRG contains 3 annotation sections. One contains the mapping information of the LRG genomic sequence to the current genome assembly, and then one each for Ensembl and NCBI containing feature annotation and other elements.

- **source –** as above, but this should be the source of the annotation in this annotation_set
- **modification_date –** obvious; this should be changed when the annotation is updated, either automatically or by hand
- **mapping –** this element contains the mapping by sequence alignment of the LRG sequence to a specified reference sequence, for example the current genome assembly. The reference sequence is specified by the attributes **coord_system**, **other_name and other_id.** Multiple instances are allowed such that we can store mappings to multiple reference sequences. The mapping element has the following attributes:
  - **coord_system**
  - **other_name –** Usually the same as the **coord_system** for the transcript mappings
  - **other_id**
  - **other_start** and **other_end** - these coordinates represent the most 5' and most 3' coordinates covered by all of the contained **mapping_span** elements, i.e. the total span of the mapping
  - **mapping_span –** each mapping span represents a largely contiguous area of alignment between the LRG sequence and the reference. The mapping element can have one or more child mapping_span elements. The reason for this is if the LRG maps into several large separate segments. The **mapping_span** element has the following attributes:
    - **lrg_start** and **lrg_end** – LRG coordinates
    - **other_start** and **other_end** – Coordinates from an other system (e.g. GRCh37, Ensembl transcript, RefSeqGene transcript)
    - **strand**

    And the **mapping_span** element can contain the following element:
    - **diff –** Each mapping_span element can then contain one or more diff elements - this allows us to precisely describe any single- or multiple-basepair differences between LRG and the reference, as well as any insertions or deletions. The **diff** element has the following attributes:
      - **type** - can be mismatch, lrg_ins (representing an insertion in the LRG) or genomic_ins (representing an insertion in the reference, or to think of it another way, deletion in the LRG)
      - **lrg_start** and **lrg_end** - describes the coordinates of the difference in LRG coordinates
      - **other_start** and other_**end** - describes the coordinates of the difference on the other system (e.g. GRCh37, Ensembl transcript, RefSeqGene transcript)
      - **lrg_sequence** - the sequence in the LRG - used for mismatches and lrg_ins instances
      - **other_sequence** - the sequence of the difference in the other system (used for mismatches and insertions)
- **fixed_transcript_annotation –** this information contains updatable information relating to the transcripts included in the fixed section, for example exon numbering
  - **other_exon_numbering** – this element contains alternate exon naming for each exon within the transcripts included in the fixed section.
    - **exon**
      - **coordinates –** Exon coordinates in LRG coordinate system
      - **label –** Alternative exon naming
- **lrg_locus** – Main gene symbol associated with the LRG
- **features** - this element contains the genomic annotations found in the region mapped to by the LRG. It is distinct from the information found in the fixed section in that it may change, and may contain more genes and transcripts than the fixed layer. These annotations should not beused as the reference point for further annotations - only the features described in the fixed section should be used for this
  - **gene** - represents a gene, and has, along with attributes for coordinates and strand (all relative to the LRG sequence), an attribute "symbol" - this should be the HGNC standard symbol for the gene. Gene, transcript

and protein_product elements can all contain optional elements from the following set:

- **symbol -** Symbol of the gene (usually, this is an HGNC symbol)
  - **synonym –** Other name(s) used to define the gene
- **coordinates –** In LRG coordinate system
- **partial** - indicates that the feature only overlaps the region mapped by the LRG partially, and can be '5-prime' or '3-prime'. 3-prime indicates that the 3-prime end of the transcript lies outside of the LRG, while 5-prime indicates the 5-prime end
- **long_name** - a long name for the feature, e.g. collagen, type I, alpha 1
- **db_xref** - an element representing an identifier for this feature in an external database. These identifiers can currently only come from a limited set of sources: GeneID, HGNC, MIM, GI, RefSeq', Ensembl, CCDS, UniProtKB
  - **synonym –** Other name(s) liked to the external identifier
- **transcript** - can be zero or more transcript elements per gene. It contains information about the protein product if it codes for a protein.
  - **coordinates –** In LRG coordinate system
  - **long_name** - a long name for the transcript
  - **db_xref** - an element representing an identifier for this feature in an external database. These identifiers can currently only come from a limited set of sources: GeneID, HGNC, MIM, GI, RefSeq', Ensembl, CCDS, UniProtKB
    - **synonym** – Other name(s) liked to the external identifier
  - **exon –** List of the exon names in the source
    - **coordinates –** In LRG coordinate system
  - **protein_product** - each transcript element can have zero or more protein_product child elements, with attributes representing the coding start and end in LRG coordinates
    - **coordinates –** In LRG coordinate system
    - **long_name** - a long name for the protein
    - **db_xref** - an element representing an identifier for this feature in an external database. These identifiers can currently only come from a limited set of sources: GeneID, HGNC, MIM, GI, RefSeq', Ensembl, CCDS, UniProtKB
      - **synonym** – Other name(s) liked to the external identifier

Changes from LRG schema 1.7

- New attribute "**label**" in the "**exon**" tags (fixed section).
- Changes on the tag "symbol":
  - Only one tag "**symbol**" is allowed per "**gene**"
  - The symbol name has been moved from the content of the tag "**symbol**" to a new attribute "**name**".
  - The tag "**synonym**" (0 to many) has been added as a child of the tag "**symbol**".
  - Moved up the tag "**symbol**" at the top position in the tag "**gene**".
- Change the pattern of the LRG transcript and protein coordinate systems (i.e. the attribute "**coord_system**" in the tag "**coordinates**") by removing the character "_" between the LRG name and the transcript and protein name, e.g.:
  - **LRG_1t1** instead of LRG_1_t1
  - **LRG_1p1** instead of LRG_1_p1
- Added an optional tag "**comment**" in the tag "**other_exon_naming**"