HL7 v3 Substance Model DRAFT

Gunther Schadow, Larry Callahan

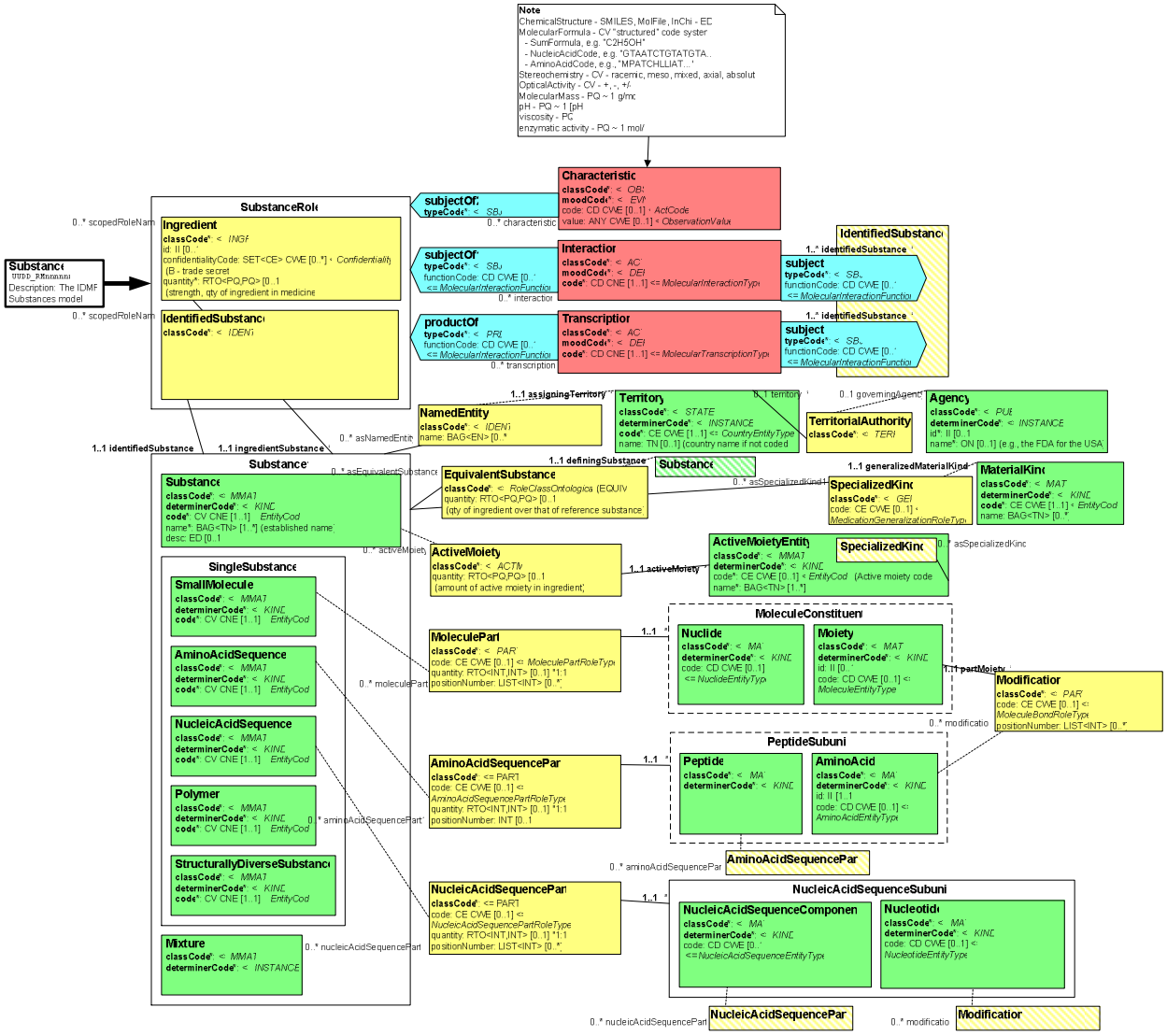
Table of Contents

Introduction

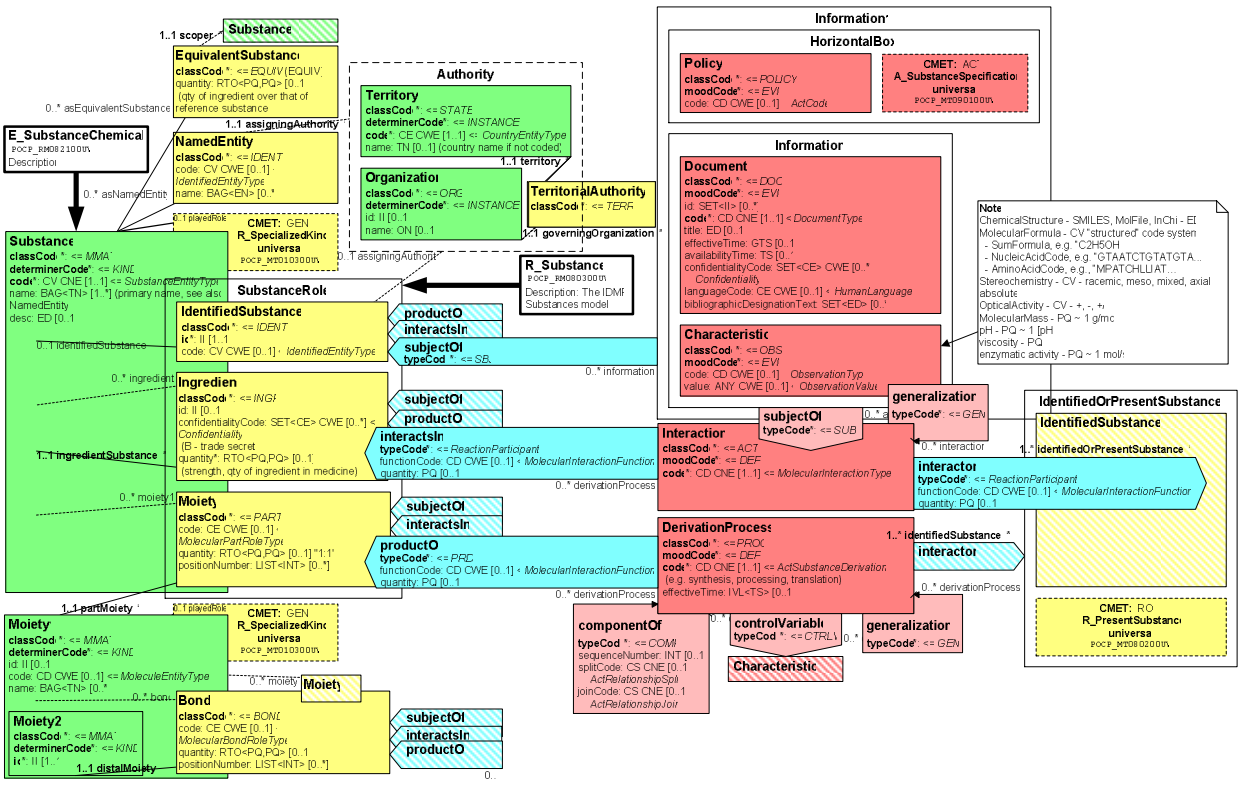
This document is a preliminary working draft for the development of the HL7 v3 chemical substance model. The motivation and use case requirements for this model in general have been specified in the ISO IDMP process and are not further elaborated here at this time.

Overview of the Substance Model Draft

The substance model shown below is the result of a 2-step process. First the outline of the requirements specification that divides molecules into the categories, simple chemical, polymer, protein, nucleic acid, and other complex described substance was turned into a first draft design in which these categories were immediately visible.



In a second step the similar nature of the elements of each of these categories was realized, and also further analysis of the stated requirements revealed many common facets that were re-used across the different categories. Therefore a simpler slightly more generic model was created that still contains the same features and expressiveness.



The details of this model as well as examples are being exposed in direct relationships with the requirements analysis in the rest of this document.

Note that while this model addresses all the requirements in principle, there are some details which are still under development and corrections and completions are made to this model as the specific examples to the requirements are formulated.

Brief Walkthrough

Entry Points: The usual entry point for a substance submission is the IdentifiedSubstance role. The entry point on the Substance entity allows the detailed substance structure to be used in combination with other entities (e.g., directly connected to formulations.)

The Substance Entity class represents the substance (as a universal). Upon hearing the word “substance”, one may think of (1) a larger amount of this substance, such as a pail full of a liquid, as sizeable crystal, or of (2) a single smallest quantum of the substance, such as a single molecule or a single complex. Which intuition we choose determines how we consider the notions of quantity (measured vs. counted) or part (a cup from the pail, chip from the crystal vs. a molecular substructure, or moiety, from the molecule). Whether we have a single molecule or a large amount, however, is only a difference in quantity (many singular entities make up the amount and there is no amount of the substance smaller than its singular molecule. The differences in the conception of part along this quantity axis are in no way special even macroscopic objects can be subdivided into parts in different ways. E.g., a loaf of bread can be divided into (1) crust outside and the crumb inside or (2) it may be divided into slices; in whole-grain bread, one may (3) discern on elongated bread, such as French Baguette, the ends from the shaft; and one may (4) dissect individual kernels of grain from the more homogenous polymerized starch mass. Thus it is with substances also: they may be dissected into portions, regional parts, and molecular parts. The molecular parts are called moieties.

The Substance is identified by a **code**, which is the primary code, a primitive without further post-coordination and no translations. Any use case where the Substance needs to be related to other conceptualizations of the same (or sufficiently equivalent) Substance, shall use the EquivalentSubstance role. The SpecializedKind structure can be used, as always, to declare the Substance a specialization of any number of Substance classifications.

The IdentifiedSubstance role, also one of the entry points to the Substance model, declares that this substance, identified by that **id**, is so identified by the scoping Organization or for the scoping region ("Territory"). The substance id (extension and root) and the playing Substance Entity's code (code and codeSystem) must be the same. This will allow references to Substances as Entities and in a generic IdentifiedSubstance role to resolve to one unique object.

Moiety is a molecular part, or, more generally, a part of the smallest amount (quantum) of the substance. Moiety is a structural part, meaning that removal of such parts changes the nature of the substance. Conversely, removal of a portion from a crystal does not change the substance. The Moiety class in the model is a partitive Role class with a code attribute, allowing the further distinction of the kinds of molecular subdivisions.

Moieties may be constructed of smaller moieties (e.g., functional groups). For example, if one wishes to say that one of the hydrogen atoms on a benzene ring is substituted with a hydroxyl group (to form phenol), one may say that the base moiety is benzene but a Modification exists connecting the benzene with a hydroxyl group. Note, this detail of small molecules may be (and often would be) conveyed in encapsulated structure specifications elsewhere. However, for larger molecules, such as proteins, the sub-moieties would specify the existence of, for example, a glycan or phosphate residue substituted on the base moiety.

When substances are being defined based on moieties, these base moieties are often stated by their basic Substance code that stands for the unmodified substance. However, this would create a problem, since the code in the Substance would mean that the substance is unmodified, when in fact it is being modified by the connection with the Modification role. To prevent ambiguity, one must in these cases provide also a unique identifier on the base substance. This identifier makes the modified Substance special and different from the unmodified Substance referenced in the code.

Moiety1 (Entity) is the substance that is the part of the larger substance. Rather than repeating all the descriptive elements on the Moiety (Entity), the Moiety (Entity)'s code contains the code of another defined Substance so its description may be looked up (or can be sent with the same data package).

Bond: A bond is a connection between two specified moieties. Normally, moiety structures can be specified simply by breaking down the molecule into its part moieties and breaking those moieties down further. However, when multiple bonds exists between moieties, those may be specified. This can also be used to specify all the bonds in case where the moieties are provided as a parts list. For instance, in the most general case, a substance is specified by all its elements and all its bonds (this is, for example, how the MDL/MOLFILE format works.) Specific examples of the use of the bond role is to specify a protein based on its sub-units (chains) as moieties and the disulfide bridges specified as bonds connecting the subunits.

The Moiety and Modification roles have quantity and positionNumber attributes. The **quantity** specifies how many (numerator) of the specified Moieties or Modifications exist in the whole molecule (denominator). The **positionNumber** allows indicating where the moiety or modification attaches to the base molecule. The detail of how positionNumber is used and interpreted must be specified in the definition of the Moiety.code or Modification.code.

DerivationProcess: specifies how the substance is being made. This information is not relevant for substances that are structurally completely defined and should not be used in those cases. The derivation information may be necessary, however, if the substance is not completely defined by exact structure. For example, the exact number and placement of substitutions of additional moieties to a base substance may not be constant, the length of chains of a polymer may not be specified exactly, and the length of chains of a tryptic peptide mix may not be exactly defined. When all that can be said of a substance is what the source material and processes are, then this is specified using the DerivationProcess and referencing the source material as an IdentifiedSubstance.

Interaction allows specifying the counter-part of a molecule which is essentially defined by its interaction. For example, antibodies are defined by their antigen interaction. Likewise, receptors are defined by the target (e.g. hormone) and a complex peptide-hormone may be defined by the receptor that it reacts with. Note, the Interaction structure even allows to specify detailed molecular reactions which form pathways, while this is a logical step with many important applications in biomedical research computing, it is not the main purpose for including the Interaction structure into the Substance model. Here the Interaction structure is intended for use if this Interaction contributes essentially to the very definition of the Substance.

Characteristic is a recurring theme in any specification of material. These Characteristics are observable properties that can be examined (measured) on the substance and their specification contributes to the definition of the substance or provides essential and useful information about the substance. For example, molecular mass can be ascribed to the substance by means of a Characteristic. Coded Characteristics may specify chiral properties of the substance. And Characteristics with encapsulated data can provide a structure specification using a standard external to HL7, such as MOLFILE, SMILES or InChi.

Documents may be references to other substance monographs or journal articles that contribute to the very definition of the substance.

Data Element Requirements Traceability

After the brief walk through the substance model, the following sections provide a detailed description of the use of this substance model to address the stated requirements of the Substance project. To do this, the requirements are being listed, described, and analyzed in several levels of details.

The following data element requirements were taken from ISO committee drafts and the result of further IDMP task group activity. The data elements below were received in the form of an XML structure. This structure was transformed into the first draft of the itemized lists presented in this section. Then the descriptions were carefully edited to be in the style appropriate for data element definitions. The nesting and ordering of itemize list was edited to be systematic and to avoid unnecessary nesting.

The resulting lists will be shown below in 3 levels of detail, first as a complete overview, then each logical module separately. Finally common grouping of data elements which occur in several logical module are summarized.

Brief Data Elements Overview

The substance standard conveys definitions, and definitions have 2 parts: (1) the id and/or name to be defined, and (2) the definition itself. The definition should is a formal one, hence any alias names, translations, ids, and reference literature are only supplemental information given in the id/name part, whereas the definition contains actual content.

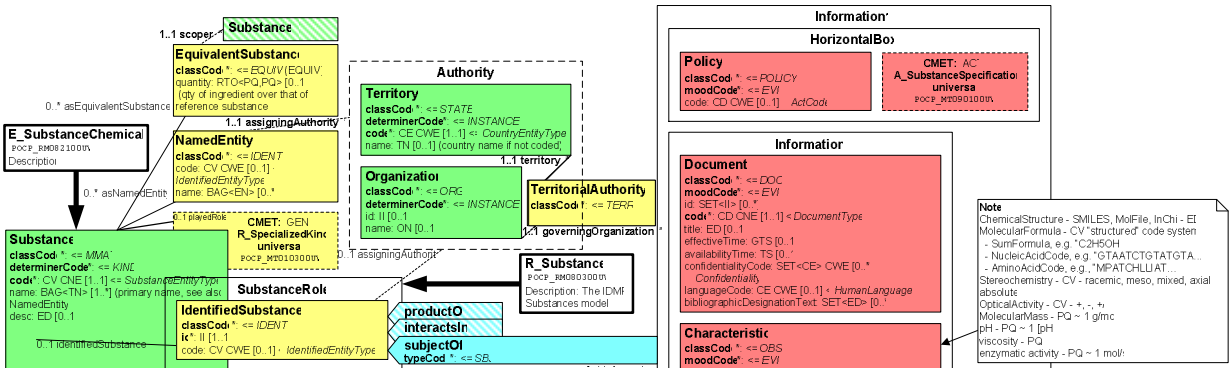
1. **Substance to be defined:**
   - 1.1. **Substance Id:** a single unique identifier being defined.
   - 1.2. **Substance Name(s):** names, identifiers and codes given to the substance in different languages, jurisdictions, regions, and fields. Includes, name, code, brand name, established name, primary name, and other names. Established name is a non-proprietary name that is assigned to a substance by either a regulatory agency or by a scientific organization. A primary name is a name in common use when no established name exists. Also includes references to published literature or regulatory documents.
2. **Definition of the substance:**
   - 2.1. **Chemical Substance:** substance defined by structures, using such chemical structure representations as MOLFILE; InChi; SMILES, etc.
   - 2.2. **Polymer Substance:** naturally occurring or synthetic linear, branched, crosslinked, 2d-network; 3d-network; multi-branched; circular structures consisting of monomer units in repeated patterns. Excludes proteins and nucleic acid structures which are describes separately.
   - 2.3. **Protein Substance:** a protein is defined as a subunit or combination of subunits that are either covalently linked or have a defined invariant stoichiometric relationship. Each subunit will be described separately by its amino acid sequence.
   - 2.4. **Nucleic Acid:** DNA and RNA as oligonucleotides, genes, and any nucleic acid aptomers with a length greater than three nucleotides.
   - 2.5. **Structurally Diverse Substance:** substances where the ingredients are not fully defined but instead that are defined by origin and processing.
   - 2.6. **Mixtures** are multiple substances that are isolated or synthesized together and where each component can be specified by a substance id, not including mixtures (no mixtures of mixtures.) Racemic mixtures, substances containing unknown or mixed stereochemistry, impurities or degradents are not described as mixtures. Substances present in trace amounts will not be listed in a mixture unless they are known to have a specific effect. Mixtures are also used when substance ambiguity exists in authoritative sources (aloe).

Substance Id and Name Data Elements

Data Elements Overview

1. **Substance Id:** single unique identifier (code, e.g., UNII code).
2. **Substance Name(s):** names, identifiers and codes given to the substance in different languages, jurisdictions, regions, and fields. Includes, name, code, brand name, established name, primary name, and other names. Established name is a non-proprietary name that is assigned to a substance by either a regulatory agency or by a scientific organization. A primary name is a name in common use when no established name exists. Also includes references to published literature or regulatory documents.
   - 2.1. **Name and Name Type:** code, brand name, established name, primary name, other name.
   - 2.2. **Language Name and Code**
   - 2.3. **Country Name and Code**
   - 2.4. **Domain** e.g., food, drug, biologic, devices, excipients, it is the domain of use of a name which determines which names must be used. For example, the same substance would be called “FD&C Yellow #8” as an excipient vs. “flourescein” for drug active ingredient.
   - 2.5. **Established Name Source:** organization that is the authority for the established name (USAN, INN, JAN, BAN, USP, EP, Tropicos, INCI, JECFA.)
   - 2.6. **Reference:**
     - 2.6.1. **Reference Type:** public literature, regulatory submission
     - 2.6.2. **Reference Code:** e.g. NDA, IND, BLA and other ids.
     - 2.6.3. **Reference Citation:** literature reference or specific reference to a regulatory document

The following is the fragment of the RMIM diagram that shows the design representing the above listed data requirements.



This RMIM provides for the following XML data structure (entering from R_Substance):

```
<identifiedSubstance classCode="IDENT">
  <id extension="P88XT4IS4D" root="2.16.840.1.113883.4.9"/>
  <identifiedSubstance classCode="MMAT" determinerCode="KIND">
    <code code="P88XT4IS4D" codeSystem="2.16.840.1.113883.4.9"/>
    <name>paclitaxel</name>

    <asNamedEntity>
      <code code="C??????" codeSystem="2.16.840.1.113883.3.26.1.1"
            displayName="established name"/>
      <name xml:lang="fr_FR">paclitaxèle</name>
      <assigningOrganization>
        <!-- an authority organization present -> established name -->
        <id extension="EP" root="2.16.840.1.113883.9.9.9"/>
        <name>European Pharmacopoeia</name>
      </assigningOrganization>
```

```
      <subjectOf>
        <namePolicy>
      </subjectOf>
    </asNamedEntity>

    <asNamedEntity>
      <code code="C??????" codeSystem="2.16.840.1.113883.3.26.1.1"
            displayName="primary (non-established) name"/>
      <name xml:lang="ru_RU">Паклйтакснлы</name>
      <assigningTerritory>
        <code code="RUS" codeSystem="..." displayName="Russia"/>
      </assigningTerritory>
    </asNamedEntity>

    <asEquivalentSubstance>
      <definingMaterialKind>
        <code code="CHEBI:123456" codeSystem="1.2.3.99.1"/>
      </definingMaterialKind>
    </asEquivalentSubstance>

  </identifiedSubstance>

  <subjectOf>
    <document>
      <id extension="PMID:123445553" codeSystem="1.2.3.99.2"/>
      <bibliographicDesignationText>Johnson JJ, Gipson GG. Compounds used in drug
eluting stents. 2009, J Am Pharm Assoc, 123(5), p1203-
5.</bibliographicDesignationText>
    </document>
  </subjectOf>

</identifiedSubstance>
```

NOTE: For the regulatory reference (NDA, IND, BLA number, etc.) we could repeat here the approval structure as we have it for products.

Substance Description

Overview

1. **Chemical Substance:** other reference information fields may be developed
   1.1. **Structural Representation:** representation of structure according to the MOLFILE, InChi, or SMILES format.
   1.2. **Non stoichiometric** definition for chemical substances that do not have defined stochiometry, described by moieties.
2. **Polymer Substance:** naturally occurring or synthetic linear, branched, crosslinked, 2d-network; 3d-network; multi-branched; circular structures consisting of monomer units in repeated patterns. Excludes proteins and nucleic acid structures which are described under their special category.
   2.1. **Structural Representation:** representation of the repeating units according to the MOLFILE, InChi, or SMILES format.
   2.2. **Polymer Type:** type (e.g. homopolymer, copolymer), geometry (e.g. linear, branched, crosslinked, 2d-network; 3d-network; multi-branched; circular), copolymer sequence (e.g., block, random, graft.)
   2.3. **Monomers:** specifies and quantifies the monomers used for the synthesis of the polymer. Applicable to synthetic polymers.
   2.4. **Repeat Unit:** specifies and quantifies the repeated units and their configuration.
   2.5. **Source Material:** the material from which the final substance was originated (e.g., biological, organism, mineral)
   2.6. **Modification:** irreversible modifications to a polymer when derived from natural polymers. Synthetic variations such as r-group modifications will be described structurally and as specific substituents.

3. **Protein Substance:** a protein is defined as a subunit or combination of subunits that are either covalently linked or have a defined invariant stoichiometric relationship. Each subunit will be described separately by its amino acid sequence.
    3.1. **Structural Representation:** representation of structure according to the MOLFILE, InChi, or SMILES format. While the amino-acid sequence is also provided, the structure format allows specifying any modifications in great detail.
    3.2. **Amino Acid Sequence:** either, complete or partial when such a sequence is available or derivable from a nucleic acid sequence.
    3.3. **Subunits:** complete linear sequence of amino acids linked through peptide bonds. Subunits are used when the finished protein is a complex of multiple sequences; subunits are not used to delineate domains within a single sequence. Subunits are listed in order of decreasing length; sequences of the same length will be ordered by molecular weight; subunits that have identical sequences will be repeated multiple times.
    3.4. **Disulfide Bonds:** position of the disulfide bonds in the protein listed in increasing order of subunit number and position within subunit.
    3.5. **Glycosylation:** site and type of glycosylation. The type of glycosylation is described by indicating the host system from which the protein was isolated (e.g., human, mammalian, avain, insect, plant, fungal, etc.) but not the detailed structure and branching pattern.
    3.6. **Other Modifications:** irreversible modifications to a protein or peptide (physical, chemical, enzymatic) other than disulfide bonds or glycosylation, e.g., PEGylation, phosphorylation.
4. **Nucleic Acid:** DNA and RNA as oligonucleotides, genes, and any nucleic acid aptomers with a length greater than three nucleotides.
    4.1. **Sequence Type:** DNA, RNA, mixed
    4.2. **Subunits:** the strands that make up the nucleic acid structure; in most gene examples only the transcribed strand will be described. Sometimes, both of the strands need to be described. Such cases arise for mismatched double-stranded oligonucleotides (e.g. aptomers), or single overhangs from restriction enzyme cleavages.
    4.3. **Component:** genetic elements within a given gene; enhancer; promoter; coding sequence; termination signal; silencer etc; or molecules linked to nucleic acid. A functional gene would have multiple components.
5. **Structurally Diverse Substance:** substances where the ingredients are not fully defined but instead that are defined by origin and processing.
    5.1. **Source Material:** describes the material fom which the final substance was originated from; cv: biological, organism, mineral
    5.2. **Modification:** irreversible modifications to a substance
    5.3. **Properties:** physical, chemical or biological characteristic of the substance (e.g. viscosity, density, ph, enzymatic activity etc)
6. **Mixture:** Multiple substances can be mixtures if they are isolated or synthesized together. Racemic mixtures or substances containing unknown or mixed stereochemistry will not be defined as mixtures. Substances that contain impurities or degradents will not be described as mixtures. To avoid confusion and database problems, mixtures of mixtures will not be allowed. Each component of a mixture should be listed. Substances present in trace amounts will not be listed in a mixture unless they are known to have a specific effect. mixtures are also used when substance ambiguity exists in authoritative sources (aloe)
    6.1. **Mixture type:** one of, all of, any of
    6.2. **Component:** specifies the presence of each component (ingredient) in a mixture

Common Features

Many features are not specific to any one of the substance categories, but apply to several or all of them. Among them are:

1.1. **Structural Representation:** representation of structure according to the MOLFILE, InChi, or SMILES format.
1.2. **Moieties,** sub-units, components of the substance, fully described or referenced by code. This is also used to describe substituents groups which are structural modifications of a base substance skeleton.
1.3. **Source Material:** the material from which the final substance was originated (e.g., biological, organism, mineral)

1.4. **Modification:** irreversible modifications to a substance specified in terms of processes that the source substance undergoes (e.g., hydrolysis). Structural variations are usually described by the structure representation of by moieties.
1.5. **Molecular Mass** of the substance following modifications where applicable
1.6. **Other Properties:** physical, chemical or biological characteristic of the polymer (e.g. viscosity, density, ph, enzymatic activity etc)
1.7. **Isotopes:** specifies the presence of a radionuclide or a non-natural isotopic ratio (e g. C-13 enriched material). All radionuclide and non-natural isotope will also be represented in the structure field.
1.8. **Comments:** to be used rarely (i.e., substances should be defined formally, and comments should be reserved for very rare cases.

Specific Differentiating Features

We can subtract these common facets that are shared by many of the substance categories and gain a view on those features which are unique to the different categories:

1. **Chemical Substance:** other reference information fields may be developed
   1.1. **Stoichiometric / Non stoichiometric** definition for chemical substances that do not have defined stoichiometry, described by moieties.
2. **Polymer Substance:** naturally occurring or synthetic linear, branched, crosslinked, 2d-network; 3d-network; multi-branched; circular structures consisting of monomer units in repeated patterns. Excludes proteins and nucleic acid structures which are describes separately.
   2.1. **Polymer Type:** type (e.g. homopolymer, copolymer), geometry (e.g. linear, branched, crosslinked, 2d-network; 3d-network; multi-branched; circular), copolymer sequence (e.g., block, random, graft.)
   2.2. **Monomers:** specifies and quantifies the monomers used for the synthesis of the polymer. Applicable to synthetic polymers.
   2.3. **Repeat Unit:** specifies and quantifies the repeated units and their configuration.
3. **Protein Substance:** a protein is defined as a subunit or combination of subunits that are either covalently linked or have a defined invariant stoichiometric relationship. Each subunit will be described separately by its amino acid sequence. …
   3.1. **Amino Acid Sequence:** either, complete or partial when such a sequence is available or derivable from a nucleic acid sequence.
   3.2. **Disulfide Bonds:** position of the disulfide bonds in the protein listed in increasing order of subunit number and position within subunit.
   3.3. **Glycosylation:** site and type of glycosylation. The type of glycosylation is described by indicating the host system from which the protein was isolated (e.g., human, mammalian, avain, insect, plant, fungal, etc.) but not the detailed structure and branching pattern.
   3.4. **Other structurally Defined Post-Translational Modifications:** e.g., phosphorylation.
4. **Nucleic Acid:** DNA and RNA as oligonucleotides, genes, and any nucleic acid aptomers with a length greater than three nucleotides.
   4.1. **Sequence Type:** DNA, RNA, mixed
   4.2. **Subunits:** the number of strands that make up the nucleic acid structure; in most gene examples only the transcribed strand will be described
   4.3. **Component:** genetic elements within a given gene; enhancer; promoter; coding sequence; termination signal; silencer etc; or molecules linked to nucleic acid. A functional gene would have multiple components.
5. **Structurally Diverse Substance:** substances where the ingredients are not fully defined but instead that are defined by origin and processing.
6. **Mixture:** Multiple substances can be mixtures if they are isolated or synthesized together. Racemic mixtures or substances containing unknown or mixed stereochemistry will not be defined as mixtures. Substances that contain impurities or degradents will not be described as mixtures. To avoid confusion and database problems, mixtures of mixtures will not be allowed. Each component of a mixture should be listed. Substances present in trace amounts will not be listed in a mixture unless they are known to have a specific effect. mixtures are also used when substance ambiguity exists in authoritative sources (aloe)
   6.1. **Mixture type:** one of, all of, any of
   6.2. **Component:** used to indicate the presence of each component in a mixture

While there appear to be marked differences between the 3 polymers (polymer in general, protein and nucleic acid) these are really specializations of the generic moiety / sub-unit theme. There should remain no reason to describe the primary linear structure of proteins and nucleic acids any different, i.e., apart from the sequence one might identify linear sub-units and label them with specific types, such as promoter, coding sequence, termination signal, for DNA and, signaling peptide for proteins. For proteins, the range of post-translational modifications (glycosylation, phosphorylation) is richer than for nucleic acids, but not principally different. The exception might be disulfide bonds within or between protein chains which are not as simple to describe as a substituents moiety.

Detail Data Elements

Chemical Substance (Small Molecule)

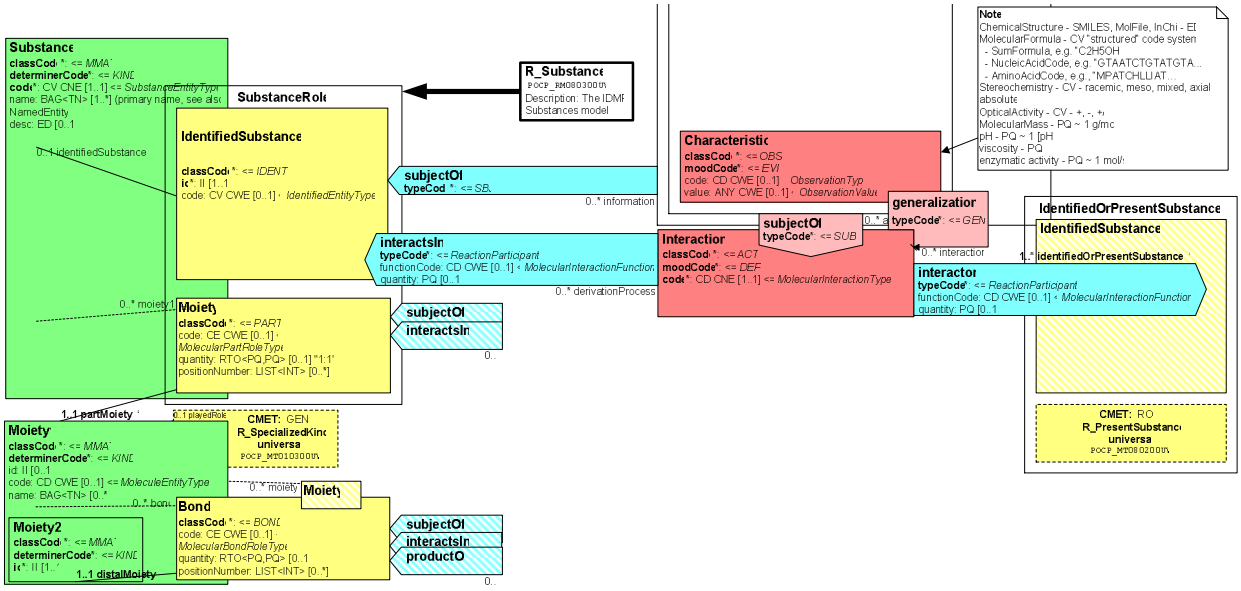
All our substances are of course in some way chemical substances. This specific category of substances, however, is reserved for substances that are chiefly described by low-level chemical structure representations, defining the constituent elements and their bonds. This low-level specification provides the greatest flexibility and accuracy; however, it is practical only for molecules of low and intermediate molecular mass. Polymers can be more efficiently described by enumerating the higher-order sub-units and their configuration.

1. **Structure:**
   - 1.1. **Structure Type:** MOLFILE; InChi; SMILES
   - 1.2. **Structural Representation:** representation of structure according to the structure type
   - 1.3. **Stereochemistry:** racemic; meso; mixed; unknown; axial, absolute; achiral (cis/trans or geometric isomerism indicated in structure)
   - 1.4. **Optical Activity:** +;-;+/-
   - 1.5. **Molecular Formula:** derived from structural representation for defined stoichiometry. Specified according to the Hill system, i.e., first C, then H, then alphabetical.
   - 1.6. **Molecular Mass:** calculated from structure
2. **Stoichiometric:** yes/no if no use non-stoichimetric schema
3. **Non Stoichiometric:** applicable to chemical substances that do not have defined stochiometry
   - 3.1. **Moieties:** an entity within a substance that has a complete and continuous molecular structure absent of counter-ions or solvate entities.
     - 3.1.1. **Moiety Id:** a label (e.g. “A”, “B”, “C”) that refers to labeled brackets in the structural representation
     - 3.1.2. **Moiety Amount Type:** cv (mole ratio; weight percent)
     - 3.1.3. **Amount Group:**
       - 3.1.3.1. **Average:** typically given in a monograph or product description of a substance; if limits are only given the arithmetic mean would be the average.
       - 3.1.3.2. **Low Limit:**
       - 3.1.3.3. **Unit:** if relative amounts are not expressed as a ratio
       - 3.1.3.4. **High Limit:** typically given in a monograph
   - 3.2. **Properties:** physical or chemical characteristics necessary to distinguish similar substances. (e.g. pH of an magnesium aluminometasilicate)
     - 3.2.1. **Property Name:** pH, viscosity, density, enzymatic activity, etc
     - 3.2.2. **Non-numeric Value:** qualitative, coded property value (e.g. yes/no; +/-) **OR**
     - 3.2.3. **Amount:** (same as above)
4. **Isotopes:** applicable for substances that contain a radionuclide or a non-natural isotopic ratio (i.e c-13 enriched material) all radionuclide and non-natural isotope will also be represented in the structure.
   - 4.1. **Nuclide Name:** example 13C
   - 4.2. **Nuclide ID:** UNII code for each non-natural or radionuclide isotope
   - 4.3. **Substitution Type:** specific (site of attachment/substitution indicated in structure); non-specific (nuclide distributed throughout molecule or substance); unknown (site unknown); extent of substitution not captured at substance level
5. **Reference Information:**
   - 5.1. **Target:** target of a given active chemical, e.g., for a receptor agent the receptor that is inhibited or activated, for an enzyme the substrate, for antibodies the antigen.
     - 5.1.1. **Target Organism:** The organism type for which the active substance is targeted (e.g., human, bacterial, viral, etc.)

5.1.2. **Target Gene Id:** id (code) for the target's gene of origin.
5.1.3. **Target Name:** The receptor or enzyme for which the active substance binds to or inhibits.
5.1.4. **Target Reference Source:** The source which indicated the target of the drug (primary or reputable secondary sources would suffice).

6. **Comment:** to be used infrequently.

The following diagram shows the elements of the substances RMIM dealing with the requirements above.



Example: Albuterol Sulfate

The following is one example, albuterol sulfate, and it's encoding.

This substance is given the primary id "021SEF3731".

```
<identifiedSubstance classCode="IDENT">
  <id extension="021SEF3731" root="2.16.840.1.113883.4.9"/>
  <identifiedSubstance classCode="MMAT" determinerCode="KIND">
    <code code="021SEF3731" codeSystem="2.16.840.1.113883.4.9"/>
    <name>albuterol sulfate</name>
```

The name "albuterol sulfate" and it is an English established name, established by the authority of the United States Adopted Names (USAN) Council for the US. This is testified to by the "USP DICTIONARY 2009", which is cited as a reference:

```
    <asNamedEntity>
      <code code="C??????" codeSystem="2.16.840.1.113883.3.26.1.1"
            displayName="established name"/>
      <name xml:lang="en_US">albuterol sulfate</name>
      <assigningOrganization>
        <id extension="USAN" root="2.16.840.1.113883.9.9.9"/>
        <name>United States Adopted Name Council</name>
        <territorialAuthority>
          <territory>
            <code code="USA" root="1.0.3166.2.2.3"/>
          </territory>
        <territorialAuthority>
      </assigningOrganization>
      <subjectOf>
        <document>
          <title>USP DICTIONARY 2009</title>
        </document>
      </subjectOf>
    </asNamedEntity>
```

Another established name is given as “Salbutamol Sulfate” in English language and for the countries of Japan as well as for the European Union. As established name source is given the JAN (presumably for Japan, not for Europe). And again, the reference “USP DICTIONARY 2009” is given.

```
    <asNamedEntity>
      <code code="C??????" codeSystem="2.16.840.1.113883.3.26.1.1"
            displayName="established name"/>
      <name xml:lang="en_UK">Salbutamol Sulfate</name>
      <assigningOrganization>
        <id extension="JAN" root="2.16.840.1.113883.9.9.9"/>
        <name>Japanese Approved Names</name>
        <territorialAuthority>
          <territory>
            <code code="JPN" root="1.0.3166.2.2.3"/>
          </territory>
        <territorialAuthority>
      </assigningOrganization>
      <subjectOf>
        <document>
          <title>USP DICTIONARY 2009</title>
        </document>
      </subjectOf>
    </asNamedEntity>
```

Another name is provided “(+/-)-2-tert-Butylamino-1-(4-hydroxy-3-hydroxymethylphenyl)ethanol SULFATE SALT (2:1)” as a chemical systematic name (rather than common or established name), this too is in English. But a source of this name has not been provided:

```
    <asNamedEntity>
      <code code="C??????" codeSystem="2.16.840.1.113883.3.26.1.1"
            displayName="systematic (non-established) name"/>
      <name xml:lang="en_UK">(+/-)-2-tert-Butylamino-1-(4-hydroxy-3-
hydroxymethylphenyl)ethanol SULFATE SALT (2:1)</name>
    </asNamedEntity>
```

The substance is also known as “51022-70-9” in the Chemical Abstract Service (CAS) (as the USP DICTIONARY 2009 claims.) These sorts of references to other databases of substances are provided as an equivalence (similarity) relation between the entities, not as an IdentifiedEntity role, because these external references may refer to slightly different molecular conceptualizations.

```
    <asEquivalentEntity>
      <definingMaterialKind>
        <code code="51022-70-9" codeSystem="1.3.6.1.4.1.5193"/>
      </definingMaterialKind>
    </asEquivalentEntity>
```

The European Bioinformatics Institute (EBI)'s Chemical Entities of Biological Interest (ChEBI) ontology gives the code 2549 to albuterol (but has not yet assigned a code to albuterol sulfate, even though it has many other organic sulfates). So, presently one would provide a similar reference as to the CAS number (for CAS, because the database is not free, most people can't even check if the CAS reference is for the sulfate or the albuterol moiety.)

```
    <asEquivalentEntity>
      <definingMaterialKind>
        <code code="2549" codeSystem="1.2.3.99.9.4"/>
      </definingMaterialKind>
    </asEquivalentEntity>
```

The Kyoto Encyclopedia of Genes and Genomes (KEGG) gives in its Anatomical Therapeutic Chemical (ATC) Classification, the code "R03CC02 - Salbutamol" under the class "R03CC – Adrenergics for Systemic Use, Selective beta-2-adrenoreceptor agonists", and "R03AC02 - Salbutamol" under the class "R03AC – Adrenergics, Inhalant, Selective beta-2-adrenoreceptor agonists" and is unlikely to assign a code for albuterol sulfate. The purpose of providing the KEGG term is therefore clearly in order to provide the classification information that comes with the ATC classification. This is similar as with MeSH concept id references. Such classifications are then represented as:

```
    <asSpecializedKind>
      <definingMaterialKind>
        <code code="R03CC02" codeSystem="1.2.3.99.9.2"/>
      </definingMaterialKind>
    </asEquivalentSubstance>

    <asSpecializedKind>
      <definingMaterialKind>
        <code code="R03AC02" codeSystem="1.2.3.99.9.3"/>
      </definingMaterialKind>
    </asEquivalentSubstance>
```

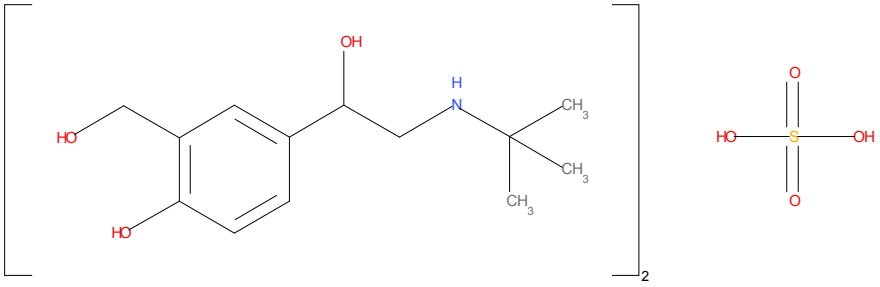
Note that one could, to the same effect, cite only the ATC classes R03AC and R03CC respectively, and would then have an even clearer representation:

```
    <asSpecializedKind>
      <definingMaterialKind>
        <code code="R03AC" codeSystem="1.2.3.99.9.3"
              displayName="Adrenergics, Inhalants, Selective beta-2-adrenoreceptor
agonists"/>
      </definingMaterialKind>
    </asEquivalentSubstance>

    <asSpecializedKind>
      <definingMaterialKind>
        <code code="R03CC" codeSystem="1.2.3.99.9.2"
              displayName="Adrenergics for Systemic Use, Selective beta-2-
adrenoreceptor agonists"/>
      </definingMaterialKind>
    </asEquivalentSubstance>
```

On the other hand, on closer look, the value of providing the KEGG ATC classification (which really is similar to many other ad-hoc pharmaceutical classifications, including MeSH), seems somewhat arbitrary and of limited value.

After dealing with naming and ids, the structure is specified in a molecular structure representation that can be drawn as:



but for the example, the MOLFILE representation would be encoded as:

```
<identifiedSubstance classCode="IDENT">
  <id extension="021SEF3731" root="2.16.840.1.113883.4.9"/>
  <identifiedSubstance classCode="MMAT" determinerCode="KIND">
    <code code="021SEF3731" codeSystem="2.16.840.1.113883.4.9"/>
    <name>albuterol sulfate</name>

  </identifiedSubstance>

  <subjectOf>
    <characteristic>
      <code code="9999-9" codeSystem="2.16.840.1.113883.6.1"
            displayName="Chemical Structure"/>
      <value xsi:type="ED" mediaType="application/x-molfile">
1001
38 38  0     1  0  0  0  0  0999 V2000
    4.2690    1.2220    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
    5.1350   -3.2780    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
    2.5369   -1.7780    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
    6.0010    2.2220    0.0000 N   0  0  0  0  0  0  0  0  0  0  0  0
    6.8671    2.7220    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0

    6.0010    1.2220    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    5.1350    0.7220    0.0000 C   0  0  3  0  0  0  0  0  0  0  0  0
  ...
  1  7  1  0  0  0  0
  1 33  1  0  0  0  0
  2 15  1  0  0  0  0
  2 37  1  0  0  0  0
  3 17  1  0  0  0  0
  ...
M  END
      </value>
    </characteristic>
  </subjectOf>
```

Another structure is provided in the IUPAC InChi format as “1S/2C13H21NO3.H2O4S/c2*1-13(2,3)14-7-12(17)9-4-5-11(16)10(6-9)8-15;1-5(2,3)4/h2*4-6,12,14-17H,7-8H2,1-3H3;(H2,1,2,3,4)”.

```
  <subjectOf>
    <characteristic>
      <code code="9999-9" codeSystem="2.16.840.1.113883.6.1"
            displayName="Chemical Structure"/>
      <value xsi:type="ED" mediaType="application/x-inchi">
1S/2C13H21NO3.H2O4S/c2*1-13(2,3)14-7-12(17)9-4-5-11(16)10(6-9)8-15;1-5(2,3)4/h2*4-
6,12,14-17H,7-8H2,1-3H3;(H2,1,2,3,4)
      </value>
    </characteristic>
  </subjectOf>
```

The sum formula is $[C_{13}H_{21}NO_3]_2 \cdot H_2O_4S$ which may be written in the formalism "2C13H21NO3.H2O4S" (which really is a code system).

```
<subjectOf>
  <characteristic>
    <code code="9998-9" codeSystem="2.16.840.1.113883.6.1"
          displayName="Chemical Sum Formula"/>
    <value xsi:type="CV" code="2C13H21NO3.H2O4S" codeSystem="1.2.3.99.9.8"/>
  </characteristic>
</subjectOf>
```

From this, one may calculate the molecular mass 576.70 g/mol.

```
<subjectOf>
  <characteristic>
    <code code="9997-9" codeSystem="2.16.840.1.113883.6.1"
          displayName="Molecular Mass"/>
    <value xsi:type="PQ" value="576.70" unit="g/mol"/>
  </characteristic>
</subjectOf>
```

This completes the example of a complete stoichiometric specification (i.e., all the elements and their quantities are known).

Stereochemistry properties are RACEMIC and optical activity "+/-".

```
<subjectOf>
  <characteristic>
    <code code="9996-9" codeSystem="2.16.840.1.113883.6.1"
          displayName="Stereochemistry Type"/>
    <value xsi:type="CV" code="C00001" codeSystem="2.16.840.1.113883.3.26.1.1"
           displayName="RACEMIC"/>
  </characteristic>
</subjectOf>

<subjectOf>
  <characteristic>
    <code code="9996-9" codeSystem="2.16.840.1.113883.6.1"
          displayName="Optical Activity"/>
    <value xsi:type="CV" code="C00002" codeSystem="2.16.840.1.113883.3.26.1.1"
           displayName="+/-"/>
  </characteristic>
</subjectOf>
```

Data Elements Represented as Characteristics

The key element of this model is the Characteristic element. It is used for giving the structure (as encapsulated data), and any properties such as stereochemistry, optical activity, molecular formula (sum formula), molecular mass and other properties such as pH, density, viscosity, biological activity, etc. These are all specified using the name-value-pair structure, for which we show the example below:

```
<identifiedSubstance classCode="IDENT">
  <id extension="P88XT4IS4D" root="2.16.840.1.113883.4.9"/>
  <identifiedSubstance classCode="MMAT" determinerCode="KIND">
    <code code="P88XT4IS4D" codeSystem="2.16.840.1.113883.4.9"/>
    <name>paclitaxel</name>

    <subjectOf>
      <characteristic>
        <code code="9997-9" codeSystem="1.2.3.99.2"
              displayName="molecular structure specification"/>
        <value xsi:type="ED" mediaType="application/molfile">
```

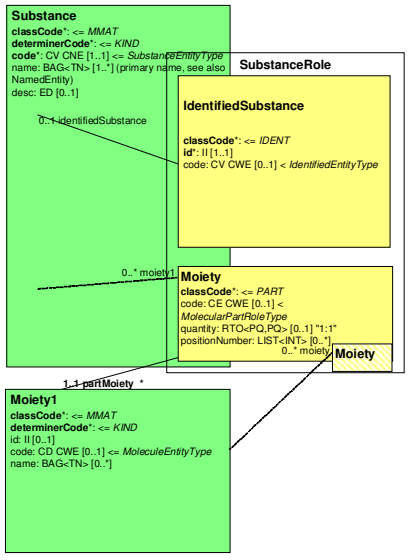
```
          ...
        </value>
      </characteristic>
    </subjectOf>

    <subjectOf>
      <characteristic>
        <code code="9999-9" codeSystem="1.2.3.99.2"
              displayName="optical activity"/>
        <value xsi:type="CV" code="+" codeSystem="1.2.3.99.3"/>
      </characteristic>
    </subjectOf>

    <subjectOf>
      <characteristic>
        <code code="9998-9" codeSystem="1.2.3.99.2"
              displayName="molecular mass"/>
        <value xsi:type="PQ" value="543" unit="g/mol"/>
      </characteristic>
    </subjectOf>

    <subjectOf>
      <characteristic>
        <code code="" codeSystem="" displayName=""/>
        <value xsi:type="URG_PQ">
          <low value="" unit=""/>
          <high value="" unit=""/>
        </value>
      </characteristic>
    </subjectOf>
```

Data Elements Specified as Moieties



Substances that cannot be represented by fixed number of moieties. The non-stoichiometric form specifies the molecule by way of is moieties and providing an approximate amount for each.

The following is an example:

```
<identifiedSubstance classCode="IDENT">
   ...
  <identifiedSubstance classCode="MMAT" determinerCode="KIND">
    ...
```

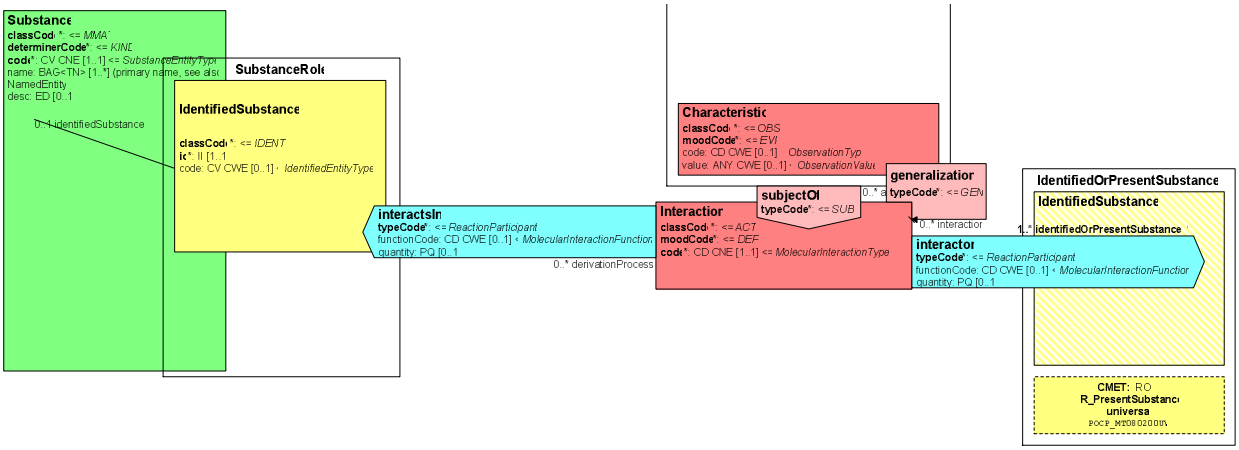
```
    <moiety>
      <quantity>
        <numerator value="2" unit="1"/>
        <denominator value="1" unit="1"/>
      </quantity>
      <partMoiety>
        <code code="ABC123XYZ9" codeSystem="2.16.840.1.113883.4.9"/>
        <name>...</name>
      </partMoiety>
    </moiety>

    <moiety>
      <quantity>
        <numerator value="10" unit="g"/>
        <denominator value="100" unit="g"/>
      </quantity>
      <partMoiety>
        <code code="ZYX987CBA9" codeSystem="2.16.840.1.113883.4.9"/>
        <name>...</name>
      </partMoiety>
    </moiety>
```

The amount can be by number of these moieties per complete molecule (unit="1"), or, which is the same, by amount of substance (unit="mol"), or finally by mass (unit="g").

Molecular Interactions

The target information is about a molecular interaction (receptor binding) so it is represented as a molecular Interaction class with another molecule, which is the Target:



The interaction is described as a code using a to be determined code system (candidate might be an IUPAC nomenclature or OBO PSI-MI, where neither one might be a perfect fit.) The concepts required here are few and simple for the time being:

- receptor binding, possibly with more specific distinction between:
    - receptor effector
    - competitive receptor inhibitor (but consider intrinsic effector activity)
    - allosteric receptor inhibitor
    - allosteric receptor modulator
- antibody-antigen binding
- enzymatic cleavage

- other binding? marking?

```
<subjectOf>
  <interaction>
    <subject>
      <identifiedSubstance>
        <id extension="L9M8N7ABC9" root="2.16.840.1.113883.4.9"/>
        <name>alpha-abcin receptor</name>
```

Polymer Substance

Naturally occurring or synthetic linear, branched, crosslinked, 2d-network; 3d-network; multi-branched; circular structures consisting of monomer units in repeated patterns. Excludes proteins and nucleic acid structures which are described under their special category.

The representation of structure has already been specified for general chemical substances.

1. **Structure:** (see above) the structural representation is to show the structure of the repeating units.
2. **Polymer Type:** (e.g., homopolymer, copolymer)
   - 2.1. **Polymer Geometry:** (e.g., linear, branched, crosslinked, 2d-network; 3d-network; multi-branched; circular)
   - 2.2. **Co-Polymer Sequence:** (e.g., block, random, graft?)
3. **Monomers Description:** specifies and quantifies the monomers used for the synthesis of the polymer. Applicable to synthetic polymers.
   - 3.1. **Monomer Number:** number of monomers used to synthesize the polymer
   - 3.2. **Monomer Amount Type:** (e.g., mole ratio; weight percent)
   - 3.3. **Monomer:** identifies and quantifies the monomer(s) used in the synthesis of the polymer
     - 3.3.1. **Monomer Name:** displayed name
     - 3.3.2. **Monomer Id:** UNII code
     - 3.3.3. **Amount:** (same as above)
4. **Repeat Unit:** specifies and quantifies the repeated units and their configuration.
   - 4.1. **Orientation of Polymerization:** (e.g., head-tail vs. random)
   - 4.2. **Repeat Unit Number:** number of repeated units represented in the structure of the polymer
   - 4.3. **Repeat Unit Amount Type:** (e.g., mole ratio; weight percent)
   - 4.4. **Repeat Unit:** identifies and quantifies the repeated units represented in the structure of the polymer
     - 4.4.1. **Repeat Unit Subscript:** (A, B, C,…) relates back to the repeating units described in the structure in order of decreasing molecular weight;
     - 4.4.2. **Amount:** (same as above)
     - 4.4.3. **Substituent:** apply when incomplete or partial site substitution is present(position or extent of substitution is not completely known)
       - 4.4.3.1. **Substituent Number:** number of substituent
       - 4.4.3.2. **Amount:** (same as above)
   - 4.5. **Degree Polymerization:** applies to homopolymer and block co-polymers where the degree of polymerization within a block can be described.
     - 4.5.1. **Amount:** (same as above)
5. **Molecular Mass:** molecular mass of the polymer following modifications where applicable
   - 5.1. **Molecular Mass Type:** the method by which the molecular mass was determined. E.g., sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE), calculated from formula, light scattering viscosity, gel permeation, etc.
     - 5.1.1. **Amount:** (same as above)
6. **Other Properties:** physical, chemical or biological characteristic of the polymer, e.g. stereochemistry, tacticity (isotactic, sydiotactic, atactic), viscosity, density, pH, enzymatic activity, etc. – as specified above.
7. **Isotopes:** applicable for substances that contain a radionuclide or a non-natural isotopic ratio (i.e c-13 enriched material) all radionuclide and non-natural isotope will also be represented in the structure field
   - 7.1. **Nuclide Name:** example 13C
   - 7.2. **Nuclide ID:** UNII code for each non-natural or radionuclide isotope

7.3. **Substitution Type:** specific (site of attachment/substitution indicated in structure); non-specific (nuclide distributed throughout molecule or substance); unknown (site unknown); extent of substitution not captured at substance level

8. **Source Material:** the material from which the final substance was originated (e.g., biological, organism, mineral)
   8.1. **Source Material Class:** general classification of the source material; e.g., bacterium, human, fungus, virus, plantae; for vaccines this is the class of infectious agents;
   8.2. **Source Material Type:** for tissue derived substance whether it is autologous, allogeneic, or xenogeneic
   8.3. **Source Material State:** live, inactivated, attenuated, conjugated, live attenuated; for inactivated vaccines, the inactivation method and agent is to be specified in the modification group;
   8.4. **Organism Id:** identifier associated with the organism name as applicable and as assigned within the organism taxonomy
   8.5. **Organism Name:** display name
   8.6. **Organism Part:** the part of the organism used to produce the substance;
      8.6.1. **Part Location:** the detail anatomic location when the part can be extracted from different anatomical location of the organism; (e.g. for cartilage: knee, elbow) (multiple alternative locations may apply.)
      8.6.2. **Developmental Stage:** stage of life for animals, plants, insects and microorganisms, e.g., fetal, juvenile, adult, larvae, sporon. Will only be captured when the substance is significantly different in these stages (e.g., fetal bovine serum).
   8.7. **Component:**
      8.7.1. **Component Class:** general classification of the component derived from the source material (e.g. cell, for plasma derived product, blood is the part and plasma or serum is the component); for herbals this refers to any component or form derived from plants/animals/minerals and not processed (e.g. oil, juice). For conjugated vaccines linker and carrier applies. For vaccines this is may describe, if applicable, the antigen characterization (e.g. whole cell, split virion, surface antigen).
         8.7.1.1. **Component Type:** the specific type of the material constituting the component (e.g. plasmid, extra-chromosomal, chondrocyte, lipase, triglycerides.
         8.7.1.2. **Component Id:** UNII code
         8.7.1.3. **Component Name:** displayed name; primarily used for gene therapy where the component is described within the nucleic acid description (e.g. P-LLO-E7; HPV-16); for cell therapy this expresses the protein which is expressed within the relevant cell (e.g. IL12); for conjugated vaccines this describe the organism strain of the carrier; for vaccines, this may identify the antigen used in the vaccine.

9. **Modification:** irreversible modifications to a polymer when derived from natural polymers. Synthetic variations such as r-group modifications will be described structurally and as specific substituents.
   9.1. **Modification Type:** general classification of modification (cv: physical, chemical, enzymatic)
      9.1.1. **Modification Description:** for each modification the specific modification will be described within a single group. each different modification will have a separate modification description group;
         9.1.1.1. **Description:** the specific modification (cv: PEGylation, phosphorylation, hydrolysis etc)
         9.1.1.2. **Modification Specificity:** nonspecific, specific, unknown
         9.1.1.3. **Modification Extent Type:** specifies how a modification is quantified or the extent of physical treatment. Applicable for nonspecific modifications.
            9.1.1.3.1. **Modification Extent Reference:** time, temperature
            9.1.1.3.2. **Amount:** needed to express amount or extent of treatment (per molecule; time and temperature; ph and time) – detail as other "amounts" above.
         9.1.1.4. **Modification Substance:**
            9.1.1.4.1. **Modification Substance Role:** agent (chemical that brings about nonspecific modifications) or moiety (specific moiety added to a polymer)
            9.1.1.4.2. **Substance Name:** displayed name
            9.1.1.4.3. **Substance Id:** UNII code

10. **Reference Information:**
    10.1. **Target:** (as specified above.)
11. **Comments:** to be used sparingly

Data Elements Represented as Characteristics

The following data elements are specified as Characteristics:

2. **Polymer Type:** (e.g., homopolymer, copolymer)
   - 2.1. **Polymer Geometry:** (e.g., linear, branched, crosslinked, 2d-network; 3d-network; multi-branched; circular)
   - 2.2. **Co-Polymer Sequence:** (e.g., block, random, graft?)

5. **Molecular Mass:** molecular mass of the polymer following modifications where applicable
   - 5.1. **Molecular Mass Type:** the method by which the molecular mass was determined. E.g., sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE), calculated from formula, light scattering viscosity, gel permeation, etc.
     - 5.1.1. **Amount:** (same as above)
6. **Other Properties:** physical, chemical or biological characteristic of the polymer (e.g. viscosity, density, pH, enzymatic activity etc) – as specified above.

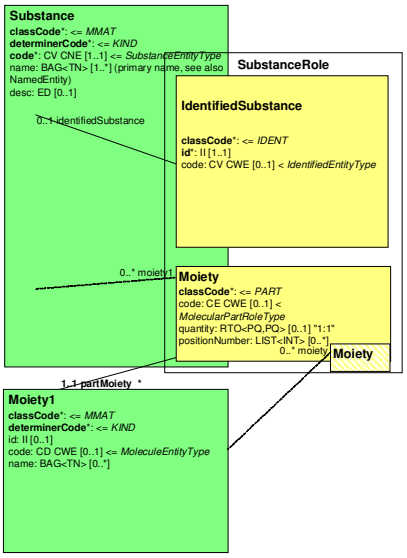
11. **Comments:** to be used sparingly

Data Elements Represented as Moieties

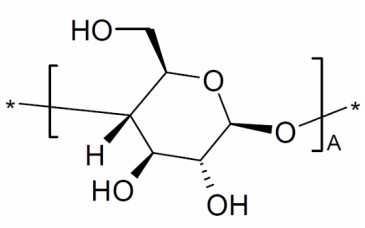
The following data elements are specified as Moieties, i.e., as molecular parts.

4. **Repeat Unit:** specifies and quantifies the repeated units and their configuration.
   - 4.1. **Orientation of Polymerization:** (e.g., head-tail)
   - 4.2. **Repeat Unit Number:** number of repeated units represented in the structure of the polymer
   - 4.3. **Repeat Unit Amount Type:** (e.g., mole ratio; weight percent)
   - 4.4. **Repeat Unit:** identifies and quantifies the repeated units represented in the structure of the polymer
     - 4.4.1. **Repeat Unit Subscript:** (A, B, C,…) relates back to structure in order of decreasing molecular weight;
     - 4.4.2. **Amount:** (same as above) – the repeat unit amount differs from the degree of polymorphism by TBD?
     - 4.4.3. **Substituent:** apply when incomplete or partial site substitution is present (position or extent of substitution is not completely known)
       - 4.4.3.1. **Substituent Number:** number of substituent
       - 4.4.3.2. **Amount:** (same as above)
   - 4.5. **Degree Polymerization:** applies to homopolymer and block co-polymers where the degree of polymerization within a block can be described. The degree of polymerization differs from the amount of the repeat unit by TBD?
     - 4.5.1. **Amount:** (same as above)

7. **Isotopes:** applicable for substances that contain a radionuclide or a non-natural isotopic ratio (i.e c-13 enriched material) all radionuclide and non-natural isotope will also be represented in the structure field
   - 7.1. **Nuclide Name:** example 13C
   - 7.2. **Nuclide ID:** UNII code for each non-natural or radionuclide isotope
   - 7.3. **Substitution Type:** specific (site of attachment/substitution indicated in structure); non-specific (nuclide distributed throughout molecule or substance); unknown (site unknown); extent of substitution not captured at substance level



The moiety structure connects an identified moiety entity as part of the to-be-defined substance as in the following example of a cellulose polymer:



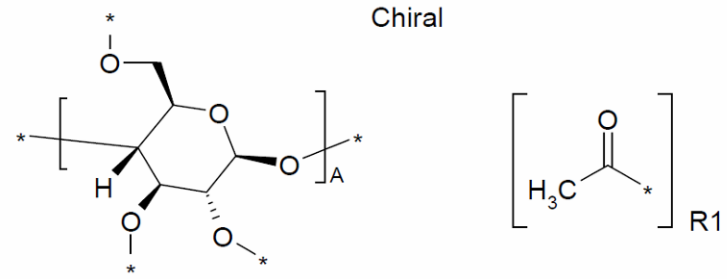
With degree of polymerization more than 440 and less than 2250:

```
<identifiedSubstance classCode="IDENT">
  <id extension="..." root="2.16.840.1.113883.4.9"/>
  <identifiedSubstance classCode="MMAT" determinerCode="KIND">
    <code code="..." codeSystem="2.16.840.1.113883.4.9"/>
    <name>cellulose</name>

    <moiety>
      <quantity>
        <numerator xsi:type="URG_PQ">
          <low value="440" unit="1"/>
          <high value="2250" unit="1"/>
        </numerator>
        <denominator value="1" unit="1"/>
      </quantity>
      <partMoiety>
        <id extension="A" root="..."/>
        <code code="{UNII}" codeSystem="2.16.840.1.113883.4.9"/>
        <name>...</name>
      </partMoiety>
    </moiety>
```

There may be two ways of referring to the repeated sub-unit, by local reference into the molecular structure data or, simpler and more obviously interoperable, by reference to the repeated sub-unit as an identified substance.

For an example of substituents, consider cellulose acetate, whereby, according to USP criteria, the weight percent of acetate is not less than 29.0% and not greater than 44.8%. Positions of potential substitution are indicated by an asterisk for both the polymer and the substituent.



Such substitutions are represented with the Modification role linking two MoietyEntities.

```
<identifiedSubstance classCode="IDENT">
  <id extension="{UNII}" root="2.16.840.1.113883.4.9"/>
  <identifiedSubstance classCode="MMAT" determinerCode="KIND">
    <code code="{UNII}" codeSystem="2.16.840.1.113883.4.9"/>
    <name>cellulose acetate</name>

    <moiety>
      <quantity>...</quantity>
      <partMoiety>
        <id extension="A" root="{internal root}"/>
        <code code="{UNII}" codeSystem="2.16.840.1.113883.4.9"/>
        <name>...</name>

        <moiety>
          <quantity>
            <numerator xsi:type="URG_PQ">
              <low value="29.0" unit="g"/>
              <high value="44.8" unit="g"/>
            </numerator>
            <denominator value="100" unit="g"/>
          </quantity>
          <moiety>
            <id extension="R1" root="{internal root}"/>
            <code code="{UNII}" codeSystem="2.16.840.1.113883.4.9"/>
          </moiety>
        </moiety>
      </partMoiety>
    </moiety>
```

Derivation Process (Modification) and Source Material

When the formula and the repeated moieties can not be specified, a less precise method is to specify the source material and the process descriptions that create the substance to be defined.

8. **Monomers Description:** specifies and quantifies the monomers used for the synthesis of the polymer. Applicable to synthetic polymers.
    - 8.1. **Monomer Number:** number of monomers used to synthesize the polymer
    - 8.2. **Monomer Amount Type:** (e.g., mole ratio; weight percent)
    - 8.3. **Monomer:** identifies and quantifies the monomer(s) used in the synthesis of the polymer
        - 8.3.1. **Monomer Name:** displayed name
        - 8.3.2. **Monomer Id:** UNII code
        - 8.3.3. **Amount:** (same as above)
9. **Source Material:** the material from which the final substance was originated (e.g., biological, organism, mineral)
    - 9.1. **Source Material Class:** general classification of the source material; e.g., bacterium, human, fungus, virus, plantae; for vaccines this is the class of infectious agents;
    - 9.2. **Source Material Type:** for tissue derived substance whether it is autologous, allogeneic, or xenogeneic
    - 9.3. **Source Material State:** live, inactivated, attenuated, conjugated, live attenuated; for inactivated vaccines, the inactivation method and agent is to be specified in the modification group;
    - 9.4. **Organism Id:** identifier associated with the organism name as applicable and as assigned within the organism taxonomy
    - 9.5. **Organism Name:** display name
    - 9.6. **Organism Part:** the part of the organism used to produce the substance;
        - 9.6.1. **Part Location:** the detail anatomic location when the part can be extracted from different anatomical location of the organism; (e.g. for cartilage: knee, elbow) (multiple alternative locations may apply.)
        - 9.6.2. **Developmental Stage:** stage of life for animals, plants, insects and microorganisms, e.g., fetal, juvenile, adult, larvae, sporon. Will only be captured when the substance is significantly different in these stages (e.g., fetal bovine serum).

9.7. **Component:**

9.7.1. **Component Class:** general classification of the component derived from the source material (e.g. cell, for plasma derived product, blood is the part and plasma or serum is the component); for herbals this refers to any component or form derived from plants/animals/minerals and not processed (e.g. oil, juice). For conjugated vaccines linker and carrier applies. For vaccines this is may describe, if applicable, the antigen characterization (e.g. whole cell, split virion, surface antigen).

9.7.1.1. **Component Type:** the specific type of the material constituting the component (e.g. plasmid, extra-chromosomal, chondrocyte, lipase, triglycerides.

9.7.1.2. **Component Id:** UNII code

9.7.1.3. **Component Name:** displayed name; primarily used for gene therapy where the component is described within the nucleic acid description (e.g. P-LLO-E7; HPV-16); for cell therapy this expresses the protein which is expressed within the relevant cell (e.g. IL12); for conjugated vaccines this describe the organism strain of the carrier; for vaccines, this may identify the antigen used in the vaccine.

10. **Modification:** irreversible modifications to a polymer when derived from natural polymers. Synthetic variations such as r-group modifications will be described structurally and as specific substituents.

10.1. **Modification Type:** general classification of modification (cv: physical, chemical, enzymatic)

10.1.1. **Modification Description:** for each modification the specific modification will be described within a single group. each different modification will have a separate modification description group;

10.1.1.1. **Description:** the specific modification (cv: PEGylation, phosphorylation, hydrolysis etc)

10.1.1.2. **Modification Specificity:** nonspecific, specific, unknown

10.1.1.3. **Modification Extent Type:** specifies how a modification is quantified or the extent of physical treatment. Applicable for nonspecific modifications.

10.1.1.3.1. **Modification Extent Reference:** time, temperature

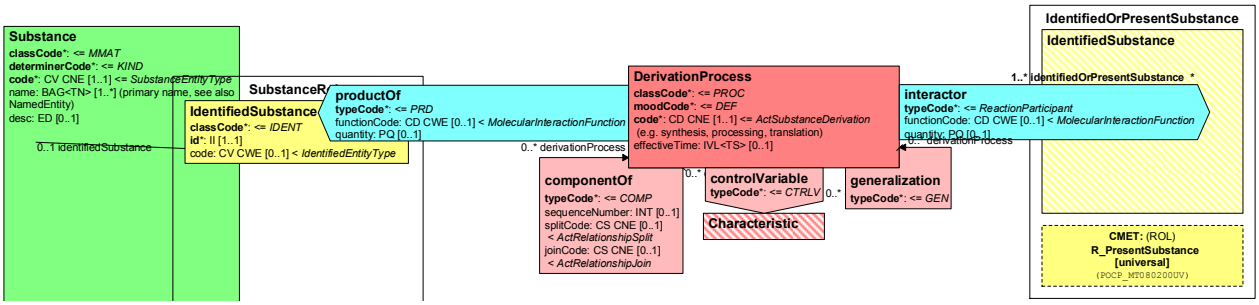
10.1.1.3.2. **Amount:** needed to express amount or extent of treatment (per molecule; time and temperature; ph and time) – detail as other "amounts" above.

10.1.1.4. **Modification Substance:**

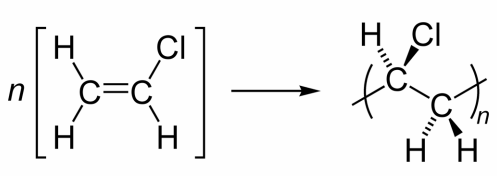
10.1.1.4.1. **Modification Substance Role:** agent (chemical that brings about nonspecific modifications) or moiety (specific moiety added to a polymer)

10.1.1.4.2. **Substance Name:** displayed name

10.1.1.4.3. **Substance Id:** UNII code



For a simple example of the monomer description take the polyvinylchloride polymerization reaction:

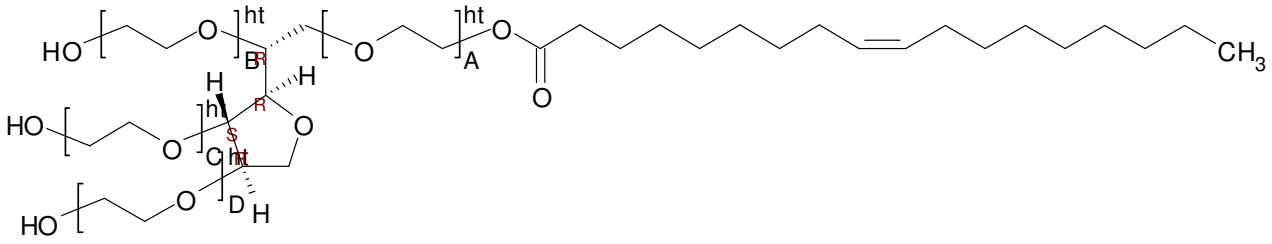


```
<identifiedSubstance classCode="IDENT">
  <id extension="{UNII}" root="2.16.840.1.113883.4.9"/>
  <identifiedSubstance classCode="MMAT" determinerCode="KIND">
    <code code="{UNII}" codeSystem="2.16.840.1.113883.4.9"/>
    <name>polivinylchloride</name>

    <productOf>
      <derivationProcess>
        <code code="{UNII}" codeSystem="2.16.840.1.113883.4.9"/>
        <subject>
          <identifiedSubstance>
            <identifiedSubstance>
              <code code="{UNII}" codeSystem="2.16.840.1.113883.4.9"/>
```

Example: Polysorbate 80

The substance 6OZP39ZG8H to be defined is Polysorbate 80 with the following structure:



"Polysorbate 80" is an English established name defined for the US by authority of the USP National Formulary (NF) according to the information in the USP Dictionary 2009. The substance is also known as "9005-65-6" in the Chemical Abstract Service (CAS) (as the USP DICTIONARY 2009 claims.)

```
<identifiedSubstance classCode="IDENT">
  <id extension="6OZP39ZG8H" root="2.16.840.1.113883.4.9"/>
  <identifiedSubstance classCode="MMAT" determinerCode="KIND">
    <code code="6OZP39ZG8H" codeSystem="2.16.840.1.113883.4.9"/>
    <name>polysorbate 80</name>

    <asNamedEntity>
      <code code="C??????" codeSystem="2.16.840.1.113883.3.26.1.1"
            displayName="established name"/>
      <name xml:lang="en_US">polysorbate 80</name>
      <assigningOrganization>
        <id extension="NF" root="2.16.840.1.113883.9.9.9"/>
        <name>USP National Formulary</name>
        <territorialAuthority>
          <territory>
            <code code="USA" root="1.0.3166.2.2.3"/>
          </territory>
        <territorialAuthority>
      </assigningOrganization>
      <subjectOf>
        <document>
          <title>USP DICTIONARY 2009</title>
        </document>
      </subjectOf>
    </asNamedEntity>

    <asEquivalentEntity>
      <definingMaterialKind>
        <code code="51022-70-9" codeSystem="1.3.6.1.4.1.5193"/>
      </definingMaterialKind>
    </asEquivalentEntity>
```

Now the structure is provided in SMILES (with an extension to indicate repeating units A, B, C, and D in the polymer structure (using the curly braces with the label, e.g., "{A:CCO}" for the repeating unit A):

```
  <subjectOf>
    <characteristic>
      <code code="9999-9" codeSystem="2.16.840.1.113883.6.1"
            displayName="Chemical Structure"/>
      <value xsi:type="ED" mediaType="application/x-smiles">
         CCCCCCCC/C=C\CCCCCCCC(=O)O{A:CCO}CC(C1C(C(CO1){D:OCC}O){C:OCC}O){B:OCC}O
      </value>
    </characteristic>
  </subjectOf>
```

According to the structure this polymer can also be characterized as a *branched homopolymer* and we can also know the molecular mass as 1274 g/mol, which is – as all molecular mass specifications always are – an average.

```
  <subjectOf>
    <characteristic>
      <code code="9994-9" codeSystem="2.16.840.1.113883.6.1"
            displayName="Polymer Type"/>
      <value xsi:type="CV" code="C00006" codeSystem="2.16.840.1.113883.3.26.1.1"
             displayName="homopolymer"/>
    </characteristic>
  </subjectOf>

  <subjectOf>
    <characteristic>
      <code code="9993-9" codeSystem="2.16.840.1.113883.6.1"
            displayName="Polymer Geometry Type"/>
      <value xsi:type="CV" code="C00007" codeSystem="2.16.840.1.113883.3.26.1.1"
             displayName="branched polymer"/>
    </characteristic>
  </subjectOf>

  <subjectOf>
    <characteristic>
      <code code="9997-9" codeSystem="2.16.840.1.113883.6.1"
            displayName="Molecular Mass"/>
      <value xsi:type="PQ" value="1274" unit="g/mol"/>
    </characteristic>
  </subjectOf>
```

The repeat unit description will describe the 4 unit A, B, C, and D all of which are being quantified in the amount of substance ratio ("mole ratio"), where the amount of A is 20, B is 20, C is 20 and D is 20. The degree polymerization value also is 20, 20, 20, and 20.

The repeating units are described as identified molecular parts (moieties):

```
<identifiedSubstance classCode="IDENT">
  <id extension="6OZP39ZG8H" root="2.16.840.1.113883.4.9"/>
  <identifiedSubstance classCode="MMAT" determinerCode="KIND">
    <code code="6OZP39ZG8H" codeSystem="2.16.840.1.113883.4.9"/>
    <name>polysorbate 80</name>

    <moiety>
      <code code="C00008" codeSystem="2.16.840.1.113883.3.26.1.1"
            displayName="polymer repeat unit"/>
      <quantity>
         <numerator value="20" unit="mol"/>
         <denominator value="1" unit="mol"/>
      </quantity>
```

```
      <partMoiety>
         <id extension="A" root="000000000000-0000-0000-0000-00000000"/>
      </partMoiety>
    </moiety>
```

This has identified the repeating unit “A”. The UUID can be the UUID of the document or message UUID, which makes “A” an identifier unique within this document. The molecular part relationship is coded specifically to be a repeat unit distinguishing it from other moieties that may also be described. The quantity of the repeating unit is specified as 20 of those moieties (numerator) in 1 complete molecule. The other moieties B, C, and D are specified accordingly:

```
    <moiety>
      <quantity>
         <numerator value="20" unit="mol"/>
         <denominator value="1" unit="mol"/>
      </quantity>

      <partMoiety>
         <id extension="C" root="000000000000-0000-0000-0000-00000000"/>
      </partMoiety>
    </moiety>

    <moiety>
      <quantity>
         <numerator value="20" unit="mol"/>
         <denominator value="1" unit="mol"/>
      </quantity>

      <partMoiety>
         <id extension="A" root="000000000000-0000-0000-0000-00000000"/>
      </partMoiety>
    </moiety>

    <moiety>
      <quantity>
         <numerator value="20" unit="mol"/>
         <denominator value="1" unit="mol"/>
      </quantity>

      <partMoiety>
         <id extension="D" root="000000000000-0000-0000-0000-00000000"/>
      </partMoiety>
    </moiety>
```

We also learn about this polymer that it has been synthesized using ethylene oxide (JJH7GNN18P, SMILES: C1CO1) and its amount is specified as 20 (again as a mole ratio). Other monomer substances are also used in the synthesis, but are not described.

```
  <productOf>
    <quantity value="1" unit="mol"/>

    <derivationProcess>
      <code code="C00008" codeSystem="2.16.840.1.113883.3.26.1.1"
            displayName="Synthesis"/>
      <value xsi:type="PQ" value="1274" unit="g/mol"/>

      <directTarget classCode="CSM">
        <quantity value="20" unit="mol"/>
        <identifiedSubstance>
          <id extension="JJH7GNN18P" root="2.16.840.1.113883.4.9"/>
```

Protein Substance

A protein is defined as a single unit of a linear amino acid sequence, or a combination of sub-units that are either covalently linked or have a defined invariant stoichiometric relationship. This includes all synthetic, recombinant

and purified proteins of defined sequence whether the use is therapeutic or prophylactic. Examples are albumins, coagulation factors, cytokines, growth factors, peptide/protein hormones, enzymes, toxins, toxoids, recombinant vaccines, and immunomodulators. This will be used to describe peptides and proteins greater than three amino acids in length; peptides that contain three or less amino acids will be described by molecular structure alone. Peptides greater than fifteen amino acids will not contain a molecular structure representation but will be described by the following descriptive elements. Peptides between four and fifteen amino acid residues in length will contain both structural and descriptive elements.

1. **Structure:** (as above) – The structure will only be used for short peptides that frequently have non-natural amino-acids. When proteins are fully specified by the Amino Acid letter sequence with disulfide-bonds and modifications, the structure file is not used.
2. **Sequence Type:** can be either, complete, partial; the protein descriptive elements will only be used when a conmplete or partial amino acid sequence is available or derivable from a nucleic acid sequence
3. **Subunit Number:** number of subunits present in a protein;;
4. **Subunit:** a linear sequence of amino acids linked through peptide bonds
   - 4.1. **Subunit Index:** (e.g., 1, 2, 3, …) relates back to subunit in order of decreasing length; sequences of the same length will be ordered by molecular weight; subunits that have identical sequences will be repeated and have sequential subscripts.
   - 4.2. **Sequence:** amino-acids enumerated from N- to C-terminal end using standard single-letter amino acid codes. Transcribed proteins will always be described using the translated sequence; for synthetic peptide containing amino acids that are not represented with a single letter code an X will be used within the sequence.
   - 4.3. **N-terminal Moiety Id and Name:** UNII code and name for a moiety at the N-terminal end, (e.g., acetyl, formyl, etc)
   - 4.4. **C-Terminal moiety id:** UNII code and name for a moiety at the C-terminal end (e.g., amide, ethyl ester etc)
5. **Disulfide Bond:** link between two cysteine residues either on the same sub-unit or on two different sub-units. (position of the disulfide bonds in the protein listed in increasing order of subunit number and position within subunit).
6. **Glycosylation:** applicable to entire protein micro-heterogeneity due to glycosylation is not described
   - 6.1. **Glycosylation type:** refers to the host system from which the protein was isolated (e.g. human, mammalian, avian, insect, plant, fungal, etc)
   - 6.2. **N-Glycosylation:** site of n-glycosylation (asparagine); n-glycosylation is listed according to the protein sequence. The extent and type of modification at a given site is not captured.
   - 6.3. **O-Glycosylation:** site of o-glycosylation (serine, threonine, etc)
   - 6.4. **C-Glycosylation:** site of c-glycosylation (tryptophan)
7. **Other Modifications:** used to describe irreversible modifications to a protein or peptide (e.g., PEGylation, phosphorylation, etc). The modifications may be physical, chemical, enzymatic, etc. Modifications may be described by their structural result (substitution of moieties to residues, etc.) or only by the process, reagents, processing time, etc.
   - 7.1. **Modification Type:** general classification of modification (e.g. physical, chemical, enzymatic)
     - 7.1.1. **Modification Description:** the specific modification described. Each different modification will have a separate modification description.
       - 7.1.1.1. **Description:** the specific modification
       - 7.1.1.2. **Modification Specificity:** cv: nonspecific, specific, unknown
       - 7.1.1.3. **Modified Residue:** every different amino acid modified is a separate modified residue
         - 7.1.1.3.1. **Residue Modified:** e.g., 20 amino acids plus N-terminal or C-terminal)
         - 7.1.1.3.2. **Residue Site:** position of specific modifications (i.e. the 10^{th} residue on the 1^{st} subunit)
       - 7.1.1.4. **Modification Extent:** primarily used for nonspecific modification, specifies how a modification is quantified or extent of physical treatment
         - 7.1.1.4.1. **Modification Extent Type:** (e.g. molecule, time, temperature)
         - 7.1.1.4.2. **Amount:** (as above)
       - 7.1.1.5. **Modification Substance:**
         - 7.1.1.5.1. **Modification Substance Role:** for proteins, agent (a chemical that results in nonspecific modifications of a protein) or moeity (a specific moeity added to a protein molecule, e.g., bis-mono-methoxy branched polethylene glycol 40000).

7.1.1.5.2. **Substance Id and Name:** UNII code and displayed name

8. **Molecular Mass** following modifications where applicable
   8.1. **Molecular Mass Determination Method:** the method by which the molecular weight was determined, e.g., SDS-PAGE, calculated, light scattering viscosity, gel permeation, etc.
      8.1.1. **Amount:** (as above)
9. **Isotope:** applicable for substances that contain a radionuclide or a non-natural isotopic ratio (i.e c-13 enriched material) all radionuclide and non-natural isotope will also be represented in the structure field
   9.1. **Nuclide Name and Id:** example 13C, UNII code for each non-natural or radionuclide isotope.
   9.2. **Substitution Site:** specific site (site of attachment/substitution indicated in structure); non-specific (nuclide distributed throughout molecule or substance); unknown (site unknown); extent of substitution not captured at substance level
10. **Reference Information:**
   10.1. **Protein Type:** general classification of the protein (e.g., enzyme, hormone)
      10.1.1. **Protein Subtype:** specific classification of protein (e.g., IGG1)
   10.2. **Target:** target of a given protein (antibody) or enzyme or receptor inhibited or activated by a different substance or substrate upon which an enzyme acts; the human gene id and name is captured
   10.3. **Gene:** specifies the origin of the protein
      10.3.1. **Genome:** organism from which the final sequence originated.
      10.3.2. **Gene Id:** code for the gene of origin, e.g., NCBI EntrezGene
      10.3.3. **Gene Name/Symbol:** name given for a gene
      10.3.4. **Gene Reference Source:** source from which the gene sequence was derived (e.g., GenBank)
11. **Comments:** to be used infrequently

Example: Calcitonin Salmon

"Calcitonin Salmon" is an English established name used in the US by authority of the United States Pharmacopoeia (USP) and as referenced in the publication "USP 32". The substance id so defined is 7SFC6U2VI5. It has a CAS number 47931-85-1 reported by some database whose name is abbreviated as "STN". It is also classified as a peptide hormone:

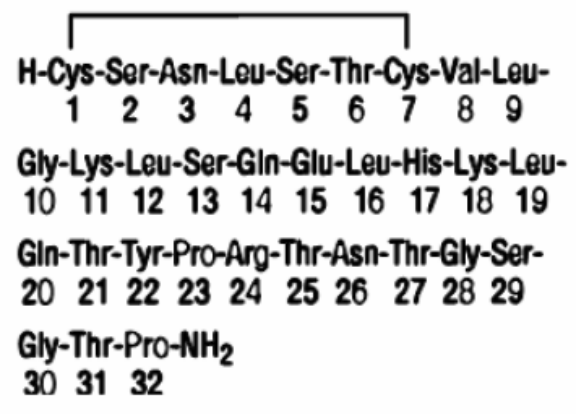
```
<identifiedSubstance classCode="IDENT">
  <id extension="7SFC6U2VI5" root="2.16.840.1.113883.4.9"/>
  <identifiedSubstance classCode="MMAT" determinerCode="KIND">
    <code code="7SFC6U2VI5" codeSystem="2.16.840.1.113883.4.9"/>
    <name>Calcitoning Salmon</name>

    <asNamedEntity>
      <code code="C??????" codeSystem="2.16.840.1.113883.3.26.1.1"
            displayName="established name"/>
      <name xml:lang="en_US">Calcitonin Salmon</name>
      <assigningOrganization>
        <id extension="USP" root="2.16.840.1.113883.9.9.9"/>
        <name>United States Pharmacopoeia</name>
        <territorialAuthority>
          <territory>
            <code code="USA" root="1.0.3166.2.2.3"/>
          </territory>
        <territorialAuthority>
      </assigningOrganization>
      <subjectOf>
        <document>
          <title>USP 32</title>
        </document>
      </subjectOf>
    </asNamedEntity>

    <asEquivalentEntity>
      <definingMaterialKind>
        <code code="47931-85-1" codeSystem="1.3.6.1.4.1.5193"/>
      </definingMaterialKind>
    </asEquivalentEntity>
```

```
    <asSpecializedKind>
      <definingMaterialKind>
        <code code="C00009" codeSystem="2.16.840.1.113883.3.26.1.1"
              displayName="peptide hormone"/>
      </definingMaterialKind>
    </asEquivalentEntity>
```

The calcitonin structure is shown in the following diagram:



The peptide hormone consists of a single base sequence of 32 amino-acids written in Dr. Margaret Dayhoff's single letter notation as "CSNLSTCVLGKLSQELHKLQTYPRTNTGSGTP" .

```
    <moiety>
      <quantity>
        <numerator value="1" unit="1"/>
        <denominator value="1" unit="1"/>
      </quantity>
      <partMoiety> </partMoiety>
      <subjectOf>
        <characteristic>
          <code code="9999-9" codeSystem="2.16.840.1.113883.6.1"
                displayName="Amino Acid Sequence"/>
          <value xsi:type="ED" mediaType="application/x-aa-seq">
            CSNLSTCVLGKLSQELHKLQTYPRTNTGSGTP
          </value>
        </characteristic>
      </subjectOf>
    </moiety>
```

There are **a few important considerations** regarding the design of the standard and the logical analysis of the situation at hand. Notice that in the above example, the partMoiety has no information in it. This entity is fully defined by the amino-acid sequence. The sequence is provided as a characteristic (property) of the moiety similar to the MOLFILE example above. This is not the only option considered. We could have considered the amino-acid sequence as a code and could have written:

```
      <partMoiety>
        <code code="CSNLSTCVLGKLSQELHKLQTYPRTNTGSGTP"
              codeSystem="2.16.840.1.113883.6.9999"
              displayName="Amino Acid Sequence"/>
      </partMoiety>
```

This makes sense because this code is fully defining of the entity that it represents. The fact that the code is also meaningful is no less of an issue than any other post-coordinated code notation that can be communicated with the HL7 code value (CV) data type. The Unified Code for Units of Measure (UCUM) is an analogous example. Still it is better not to place amino-acid sequences in the code because in many practical applications, system designers are forced by their database technology to place stringent length constraints on the codes, and amino-acid sequence can have a length of over thousand.

Another option would be to carry the amino acid sequence in the Entity.description attribute:

```
<partMoiety>
    <descr mediaType="application/x-aa-seq">
      CSNLSTCVLGKLSQELHKLQTYPRTNTGSGTP"
    </descr>
</partMoiety>
```

however, this convention would occupy the one and only descr attribute, which is usually considered as something that would be shown to humans as text. Hence, using the same "characteristic" approach provides the best flexibility and aligns well with other alternative chemical structure representations. For instance, the SMILES or MOLFILE notation could all be included as well at this point.

After these technical discussions, we must also consider a logical point. Wishing to not leave the partMoiety completely empty, we could at least give a code for this sequence. There are plenty of codes one might give to it from various sequence databases. However, one must be careful, because Genebank Ids (GI) or Uniprot Ids may or may not have exactly the same sequence. Many sequences are reported as the original mRNA transcription sequence, which contains signaling peptide and other pro-peptide constituents in the sequence. We need to make sure that if we give a code to a chemical entity then it should be really that entity, not one that is derived from it though irreversible modifications, regardless how trivial they may be (it is acceptable to ignore tautomerism and protein folding isomerism, but only because they are readily reversible, while post-translational modifications of proteins are not.)

To assign a code and still hold on to the principle of a precisely defined moiety, we could use the UniProt feature id "PRO_0000004078" for the sequence:

```
<partMoiety>
  <code code="PRO_0000004078" codeSystem="2.16.840.1.113883.6.999"/>
  <name>...</name>
</partMoiety>
```

While it is acceptable to ignore tautomerism and conformational isomerism (protein folding) when defining a precise meaning for structurally defined substances and their moieties, post-translational modifications of proteins give rise to different substances or moieties. Hence using the UniProt feature identifier as a code for the calcitonin unit, would still be violating this rule because UniProt contains sequences not finished proteins.

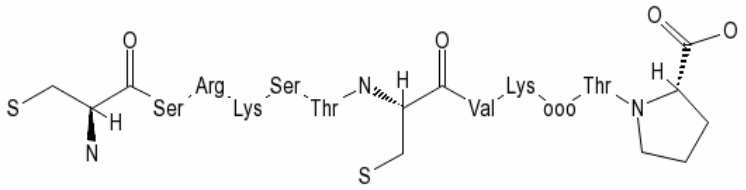
If we are to build interoperable systems, and refer to a chemical entity in some way, we must be sure that we mean what we say, and that we refer to precisely the entity that we represent, and not some derivative of it. However, in chemistry, substances are frequently classified by derivation and similarity. For example, Amoxicillin is called Penicillin, even though it is different. Outside of chemistry, scientifically correct (hence necessarily pedantic) naming conventions might have permitted calling Amoxicillin a "penicillinoid", or "penicillin-like compound", but not just "penicillin" for amoxicillin is a *penicillin derivative*, but amoxicillin *is not penicillin* (it has even clinically important different properties!). But such rigorous language is not used in chemistry. The word "penicillin derivative" is acceptable and often heard, but not usually required. So too, we tend to conflate the sequence and the sequence with the post-translational modifications.

Therefore, to signify that the moiety that we define may be similar but not identical to the unmodified amino acid sequence, we give to this moiety entity a unique id which does not reference anything but this very precise same modified moiety:

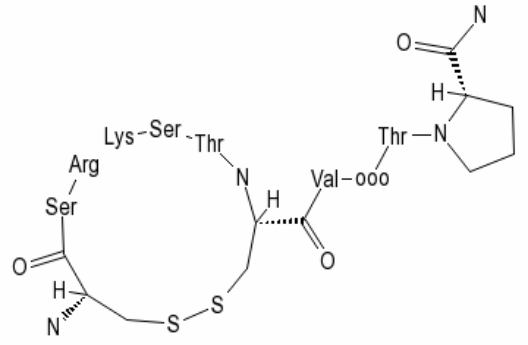
```
<partMoiety>
  <id extension="1" root="000000000000-0000-0000-0000-00000000"/>
  <code code="PRO_0000004078" codeSystem="2.16.840.1.113883.6.999"/>
</partMoiety>
```

This id can be formed by placing a UUID into the root and using short index labels (1, 2, 3, or A, B, C, etc.) for extensions. Other ways of creating such ids are possible, such as using a constant root OID and constructing extensions by the UNII code with an index, e.g., "7SFC6U2VI5-1", "7SFC6U2VI5-2", etc.

The post-translational modifications of our Calcitonin sequence after cleavage of the signaling peptide and pro-peptide moieties at the start and end of the sequence, a disulfide bond between the two Cysteines on the N-terminal position 1 and position 7 joining them to Cystine, and the C-terminal -OH group is substituted with $-NH_2$ to form an amide, so from this raw amino acid sequence:



the initial sequence with the two cysteines and the terminal final proline drawn out and other amino acids abbreviated is shown above, the impact of the disulfide bond and the amide substitution is shown below:



We can describe these modifications by specifying that the protein has one such amide group $-CONH_2$ at the 32nd, i.e., C-terminal position:

```
<moiety>
  <code code="C00008" codeSystem="2.16.840.1.113883.3.26.1.1"
        displayName="C-terminal substitution"/>
  <quantity>
    <numerator value="1" unit="1"/>
    <denominator value="1" unit="1"/>
  </quantity>
  <positionNumber value="32"/>
  <partMoiety>
    <code code="..." codeSystem="2.16.840.1.113883.6.999.."/>
    <name>amide group</name>
  </partMoiety>
</moiety>
```

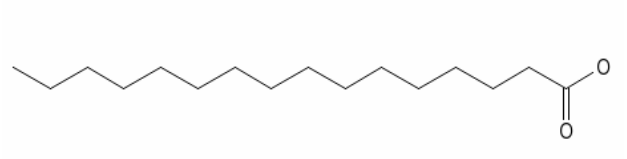
Note we provide not only a position number 32, but we also provide a specific moiety code, which indicates how the amide group –C(=O)N is not just *added* onto the existing carboxyl group as if the –C(=O)O terminus was turned into to –C(O(=ONH2))O; no, the carboxyl group is actually *replaced* by the amide group. So, to be clear about the manner in which the moieties are assembled, we use the moiety code.

The moiety part type codes which we could recognize are:

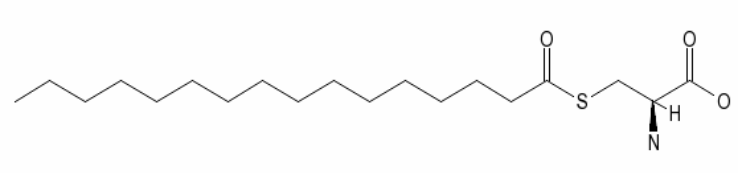
Name	Code	Description
C-terminal substitution of carboxyl group	C??????	The substitution of the C-terminal carboxyl group –COO with something else. E.g., –C(=O)OH being replaced by and amide group $-C(=O)NH_2$.
C-terminal substitution of hydroxyl group	C??????	The substitution of the C-terminal hydroxyl group –OH with something else. E.g., -OH being replaced by $-NH_2$ to form $-C(=O)NH_2$.

C-terminal substitution of hydrogen	C??????	The substitution of the C-terminal hydroxyl's hydrogen, e.g., –C(=O)OH to –C(=O)O-R.
N-terminal substitution of the amino group	C??????	The substitution of the N-terminal amino group with something else, i.e., in the way that a transaminase operates.
N-terminal substitution of hydrogen	C??????	The substitution of a hydrogen atom on the N-terminal amino group by a residue, i.e., change N-Prot with R-N-Prot.

Note that there are modifications that are commonly referred to as "N-terminal modifications" which are actually modifications of the specific amino acid residue of the amino acid in position 1, e.g., N-terminal S-palmitoylation. S-palmitoylation is a post-translational modifications of a different sort, that could possibly occur on any Cystein or Methionin residue, the fact that it may be positioned at the N-cerminal does not make it a modification of the N-terminus.

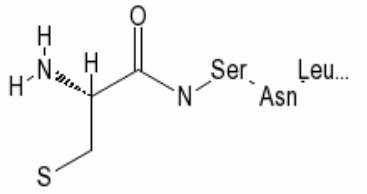


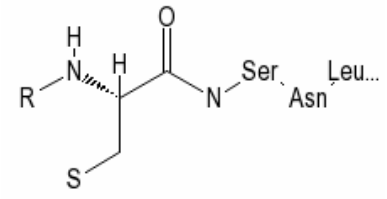
Palmytoylate + Cysteine



S-Palmitoyl Cysteine

However, a true N-Terminal substitution of hydrogen H_2N-… by a new residue R to form R-HN-… would be:





H-Cys-Ser-Asn-Leu-… R-Cys-Ser-Asn-Leu-…

So, aside from the true terminal modifications, we have other post-translational modifications which we need to represent. Such common modifications are listed below:

Name	Code	Description
S-linked residual substitution	C??????	substitution of the –SH group of cysteine or the –SCH3 group of methionine with –S-R.
O-linked residual substitution	C??????	substitution of the –OH group of serine or threonine or thyrosine with with –O–R.
N-linked residual substitution	C??????	addition of a group at the end of glutamine, asparagin, lysine, or arginine residues (includes methylation, acetylation, acylation, carboxylation, glutamylation, N-linked glycosylation, etc.)
C-linked residual substitution	C??????	Can occur anywhere, e.g. additions on aromatic residues of tryptophane, tyrosine, phenylalanine or substitutions on aliphatic residues (anywhere else.) These modifications require a specific identification of the modification site.
amino acid residue substitution	C??????	replacement of an entire amino acid residue (i.e., everything connected to the alpha C atom.
amino acid substitution	C??????	replacement of an entire amino acid including the alpha C by another amino acid.

There is a large number of modifications, most of which can be specified by identifying the moiety that is added to the site. Structural changes of other kinds would be identified by substituting entire amino acids.

So, there is clearly a difference between the peptide CSNLSTCVLGKLSQELHKLQTYPRTNTGSGTP and the so modified peptide, and now that we know how to specify base protein chains and modifications each individually, we still need to put them together, this is shown in the following complete example:

```
<identifiedSubstance classCode="IDENT">
  <id extension="7SFC6U2VI5" root="2.16.840.1.113883.4.9"/>
  <identifiedSubstance classCode="MMAT" determinerCode="KIND">
    <code code="7SFC6U2VI5" codeSystem="2.16.840.1.113883.4.9"/>
    <name>Calcitoning Salmon</name>

    <moiety>
      <code code="C??????" codeSystem="2.16.840.1.113883.3.26.1.1"
            displayName="protein sub-unit"/>
      <quantity>
        <numerator value="1" unit="1"/>
        <denominator value="1" unit="1"/>
      </quantity>

      <partMoiety>
        <id extension="1" root="000000000000-0000-0000-0000-00000000"/>
        <code code="PRO_0000004078" codeSystem="2.16.840.1.113883.6.999"/>

        <moiety classCode="PART" >
          <code code="C??????" codeSystem="2.16.840.1.113883.3.26.1.1"
                displayName="C-terminal substitution"/>
          <positionNumber value="32"/>
          <partMoiety>
            <code code="..." codeSystem="2.16.840.1.113883.6.999.."/>
            <name>amide group</name>
          </partMoiety>
        </moiety>
```

Note that we connect the modification by means of the modification element, it has a classCode which is either “PART” (for moiety) or “BOND” (see below). There is a code defining the detail of the modification.

The positionNumber is understood depending on that code. There may be 1, 2, or 3 positionNumber elements. The first one locates the amino acid from N-terminal to C-terminal end beginning with 1. The second positionNumber may locate a particular C atom (1 = alpha, 2 = beta, 3 = gamma, …). For ring structures, the C atroms are counted in clockwise order.

The other modification class is “BOND”, which can represent the disulfide bond:

```
        <bond>
          <code code="C??????" codeSystem="2.16.840.1.113883.3.26.1.1"
                displayName="disulfide bond"/>
          <positionNumber value="1"/>
          <positionNumber value="7"/>
          <distalMoiety>
            <id extension="1" root="000000000000-0000-0000-0000-00000000"/>
          </distalMoiety>
        </bond>
```

The disulfide bond has 2 positionNumbers, the first one identifying the cystein position at the scoping subunit (i.e., here the one represented by the XML element within which the modification is nested). The second positionNumber indicates the cystein position of the moiety entity referenced. The moiety entity referenced by the bond is referenced by id. Since this example is a disulfide bond within the same sub-unit, the id is the id of the nested sub-unit.

If there were 2 or more subunits and the disulfide bonds connect the sub-units, only the XML element representing the first sub-unit requires the modification element of classCode BOND. The other sub-unit will be referenced by id only.

This concludes the cystein sub-unit specification:

```
    </partMoiety>

    <subjectOf>
      <characteristic>
        <code code="9999-9" codeSystem="2.16.840.1.113883.6.1"
              displayName="Amino Acid Sequence"/>
        <value xsi:type="ED" mediaType="application/x-aa-seq">
          CSNLSTCVLGKLSQELHKLQTYPRTNTGSGTP
        </value>
      </characteristic>
    </subjectOf>
  </moiety>
</identifiedSubstance>
```

From this core structure with all modifications specified above, follows a chemical sum formula of $C_{145}H_{240}N_{44}O_{48}S_2$ and an average molecular mass of 3431.8530 g/mol, which may just as well be rounded off to 3432 g/mol.

```
<subjectOf>
  <characteristic>
    <code code="9998-9" codeSystem="2.16.840.1.113883.6.1"
          displayName="Chemical Sum Formula"/>
    <value xsi:type="CV" code="C145H240N44O48S2" codeSystem="1.2.3.99.9.8"/>
  </characteristic>
</subjectOf>

<subjectOf>
  <characteristic>
    <code code="9997-9" codeSystem="2.16.840.1.113883.6.1"
          displayName="molecular mass, computed"/>
    <value xsi:type="PQ" value="3431" unit="g/mol"/>
  </characteristic>
</subjectOf>
```

The drug-target for human use of this substance is the human calcitonin receptor, which is referred to by NCBI Entrez Gene Id 799 and Entrez taxonomy id 9606 for homo sapiens.

```
<directTargetOf classCode="SBR">
  <functionCode code="C00012" codeSystem="2.16.840.1.113883.3.26.1.1"
                displayName="receptor agent"/>
  <interaction>
    <code code="C00011" codeSystem="2.16.840.1.113883.3.26.1.1"
          displayName="effective receptor binding reaction"/>
    <directTarget classCode="SBR">
      <functionCode code="C00013" codeSystem="2.16.840.1.113883.3.26.1.1"
                    displayName="receptor"/>
      <presentSubstance>
        <presentSubstance>
          <code code="799" codeSystem="2.16.840.1.99.9999"/>
          <name>calcitonin receptor</name>
        </presentSubstance>
        <scoper>
          <code code="9606" codeSystem="2.16.840.1.113883.4.999999"
                displayName="homo sapiens"/>
        </scoper>
      </presentSubstance>
    </directTarget>
  </interaction>
</directTargetOf>
```

Example: Yttrium 90 Clivatuzumab Tetraxetan

Yttrium 90 clivatuzumab tetraxetan (id "2L271110ED"), is a protein based substance, but it has many complex aspects to demonstrate this standard's capability. The name "yttrium y-90 clivatuzumab tetraxetan" is an English name established for use in the USA by the authority of the United States Adopted Names Council (USAN), and the USAN Council Records of 2009 are named as reference testifying to this fact. The CAS number is 943976-23-6.

```
<identifiedSubstance classCode="IDENT">
  <id extension="2L271110ED" root="2.16.840.1.113883.4.9"/>
  <identifiedSubstance classCode="MMAT" determinerCode="KIND">
    <code code="2L271110ED" codeSystem="2.16.840.1.113883.4.9"/>
    <name>yttrium y-90 clivatuzumab tetraxetan</name>

    <asNamedEntity>
      <code code="C??????" codeSystem="2.16.840.1.113883.3.26.1.1"
            displayName="established name"/>
      <name xml:lang="en_US">yttrium y-90 clivatuzumab tetraxetan</name>
      <assigningOrganization>
        <id extension="USP" root="2.16.840.1.113883.9.9.9"/>
        <name>United States Pharmacopoeia</name>
        <territorialAuthority>
          <territory>
            <code code="USA" root="1.0.3166.2.2.3"/>
          </territory>
        <territorialAuthority>
      </assigningOrganization>
      <subjectOf>
        <document>
          <bibliographicDesignationText>
            United States Adopted Names Council 2009
          </bibliographicDesignationText>
        </document>
      </subjectOf>
    </asNamedEntity>

    <asEquivalentEntity>
      <definingMaterialKind>
        <code code="943976-23-6" codeSystem="1.3.6.1.4.1.5193"/>
      </definingMaterialKind>
    </asEquivalentEntity>

    <asSpecializedKind>
      <definingMaterialKind>
        <code code="C00019" codeSystem="2.16.840.1.113883.3.26.1.1"
              displayName="monoclonal antibody"/>
      </definingMaterialKind>
    </asEquivalentEntity>
```

Yttrium 90 clivatuzumab tetraxetan is a combination of an antibody (clivatuzumab) with a chelate metal ion carrier (tetraxetan) and the yttrium 90 ($^{90}Y^{3+}$) metal ion. The antibody is directed against mucin, an antigen expressed in most pancreatic cancers, but not in pancreatitis, normal pancreas or most other normal tissues. The radioactive yttrium 90 ($^{90}Y^{3+}$) has an initial half-life of 64 hours and decays with a β^- radiation at 2.28 MeV to zirconium 90 (possibly though intermediary energy stages and γ radiation?) where it ends up a stable isotope.

Being a gamma globulin antibody, clivatuzumab has 2 pairs of subunits, a light chain connected to a heavy chain and then the two heavy chains connected together. Although the two pairs are identical, we must represent all 4 subunits as separate entities order to be able to link them properly

```
    <moiety>
      <quantity>
        <numerator value="1" unit="1"/>
        <denominator value="1" unit="1"/>
      </quantity>
      <partMoiety>
        <id extension="1" root="000000000000-0000-0000-0000-00000000"/>
        <name>clivatuzumab heavy chain</name>
```

What follows are numerous disulfide bounds, first the two internal to the heavy chain:

```
        <bond classCode="BOND" >
          <code code="C??????" codeSystem="2.16.840.1.113883.3.26.1.1"
                displayName="disulfide bond"/>
          <positionNumber value="22"/>
          <positionNumber value="96"/>
          <distalMoiety>
            <id extension="1" root="000000000000-0000-0000-0000-00000000"/>
          </distalMoiety>
        </bond>

        <bond>
          <code code="C??????" codeSystem="2.16.840.1.113883.3.26.1.1"
                displayName="disulfide bond"/>
          <positionNumber value="146"/>
          <positionNumber value="202"/>
          <distalMoiety>
            <id extension="1" root="000000000000-0000-0000-0000-00000000"/>
          </distalMoiety>
        </bond>

        <bond>
          <code code="C??????" codeSystem="2.16.840.1.113883.3.26.1.1"
                displayName="disulfide bond"/>
          <positionNumber value="263"/>
          <positionNumber value="323"/>
          <distalMoiety>
            <id extension="1" root="000000000000-0000-0000-0000-00000000"/>
          </distalMoiety>
        </bond>

        <bond>
          <code code="C??????" codeSystem="2.16.840.1.113883.3.26.1.1"
                displayName="disulfide bond"/>
          <positionNumber value="369"/>
          <positionNumber value="427"/>
          <distalMoiety>
            <id extension="1" root="000000000000-0000-0000-0000-00000000"/>
          </distalMoiety>
        </bond>
```

Now the disulfide bond to the light chain:

```
        <bond>
          <code code="C??????" codeSystem="2.16.840.1.113883.3.26.1.1"
                displayName="disulfide bond"/>
          <positionNumber value="222"/>
          <positionNumber value="215"/>
          <distalMoiety>
            <id extension="3" root="000000000000-0000-0000-0000-00000000"/>
          </distalMoiety>
        </bond>
```

And two connecting the 2 heavy chains to each other:

```
        <bond>
          <code code="C??????" codeSystem="2.16.840.1.113883.3.26.1.1"
                displayName="disulfide bond"/>
          <positionNumber value="228"/>
          <positionNumber value="228"/>
          <distalMoiety>
            <id extension="2" root="000000000000-0000-0000-0000-00000000"/>
          </distalMoiety>
        </bond>

        <bond>
          <code code="C??????" codeSystem="2.16.840.1.113883.3.26.1.1"
                displayName="disulfide bond"/>
          <positionNumber value="231"/>
          <positionNumber value="231"/>
          <distalMoiety>
            <id extension="2" root="000000000000-0000-0000-0000-00000000"/>
          </distalMoiety>
        </bond>
```

There is also a specific O-linked glycosylation at the threonine at position 299, which

```
        <bond>
          <code code="C??????" codeSystem="2.16.840.1.113883.3.26.1.1"
                displayName="O-linked residual substitution, glycosylation"/>
          <positionNumber value="299"/>

          <distalMoiety>
            <code code="..." codeSystem="2.16.840.1.113883.4.9"/>
            <name>mammalian glycan residue</name>
          </distalMoiety>
        </bond>
```

This concludes the chains and what follows is the sequence, we give a unique id to that sequence, so we can reuse it:

```
      </partMoiety>

      <subjectOf>
        <characteristic>
          <id root="100000000000-0000-0000-0000-00000001"/>
          <code code="9999-9" codeSystem="2.16.840.1.113883.6.1"
                displayName="Amino Acid Sequence"/>
          <value xsi:type="ED" mediaType="application/x-aa-
seq">QVQLQQSGAEVKKFGASVKVSCEASGYTFPSYVLHWVKQAPGQGLEWIGYINPYNDGTQYNKKFKGKATLTRDTSIN
TAYMELSRLRSDDTAVYYCARGFGGSYGFAYWGQGTLVIVSSASTKGPSVFPLAPSSKSTSGGTAALGCLVKDYFPEPVTVS
WNSGALTSGVHTFPAVLQSSGLYSLSSVVTVPSSSLGTQTYICNVNHKPSNTKVDKRVEPKSCDKTHTCPPCPAPELLGGPS
VFLFPPKPKDTLMISRTPEVTCVVVDVSHEDPEVKFNWYVDGVEVHNAKTKPRESQYNSTYRVVSVLTVLHQDWLNGKEYKC
KVSNEALPAPIEKTISKAKGQPREPQVYTLPPSREEMTKNQVSLTCLVKGFYPSDIAVEWESNGQPENNYKTTPPVLDSDGS
FFLYSKLTVDKSRWQQGNVFSCSVNHEALHNHYTQKSLSLSPGK</value>
        </characteristic>
      </subjectOf>
    </moiety>
```

Now comes the second heavy chain:

```
    <moiety>
      <quantity>
        <numerator value="1" unit="1"/>
        <denominator value="1" unit="1"/>
      </quantity>

      <partMoiety>
        <id extension="2" root="000000000000-0000-0000-0000-00000000"/>
        <name>clivatuzumab heavy chain</name>
```

```
        <bond>
          <code code="C??????" codeSystem="2.16.840.1.113883.3.26.1.1"
                displayName="disulfide bond"/>
          <positionNumber value="22"/>
          <positionNumber value="96"/>
          <distalMoiety>
            <id extension="2" root="000000000000-0000-0000-0000-00000000"/>
          </distalMoiety>
        </bond>

        <bond>
          <code code="C??????" codeSystem="2.16.840.1.113883.3.26.1.1"
                displayName="disulfide bond"/>
          <positionNumber value="146"/>
          <positionNumber value="202"/>
          <distalMoiety>
            <id extension="2" root="000000000000-0000-0000-0000-00000000"/>
          </distalMoiety>
        </bond>

        <bond>
          <code code="C??????" codeSystem="2.16.840.1.113883.3.26.1.1"
                displayName="disulfide bond"/>
          <positionNumber value="263"/>
          <positionNumber value="323"/>
          <distalMoiety>
            <id extension="2" root="000000000000-0000-0000-0000-00000000"/>
          </distalMoiety>
        </bond>

        <bond>
          <code code="C??????" codeSystem="2.16.840.1.113883.3.26.1.1"
                displayName="disulfide bond"/>
          <positionNumber value="369"/>
          <positionNumber value="427"/>
          <distalMoiety>
            <id extension="2" root="000000000000-0000-0000-0000-00000000"/>
          </distalMoiety>
        </bond>

        <bond>
          <code code="C??????" codeSystem="2.16.840.1.113883.3.26.1.1"
                displayName="disulfide bond"/>
          <positionNumber value="222"/>
          <positionNumber value="215"/>
          <distalMoiety>
            <id extension="4" root="000000000000-0000-0000-0000-00000000"/>
          </distalMoiety>
        </bond>

        <moiety>
          <code code="C??????" codeSystem="2.16.840.1.113883.3.26.1.1"
                displayName="O-linked residual substitution, glycosylation"/>
          <positionNumber value="299"/>

          <partMoiety>
            <code code="..." codeSystem="2.16.840.1.113883.4.9"/>
            <name>mammalian glycosyl residue</name>
```

```
            <moiety>
              <code code="C??????" codeSystem="2.16.840.1.113883.3.26.1.1"
                    displayName="chelate-bound metal ion"/>
              <positionNumber value="1"/>
              <partMoiety>
                <code code="433ME2ATHW" codeSystem="2.16.840.1.113883.4.9"/>
                <name>yttrium 90 cation</name>
              </partMoiety>
            </moiety>
          </partMoiety>
        </moiety>
      </partMoiety>
```

The bonds to the first heavy chain do not need to be restated. Also, the sequence is identical to the first heavy chain, so we can just reuse it by stating the unique id that we had given above:

```
      <subjectOf>
        <characteristic>
          <id root="100000000000-0000-0000-0000-00000001"/>
        </characteristic>
      </subjectOf>
    </moiety>
```

Now the first light chain with 2 internal disulfide bonds, the bonds to the heavy chain need not be restated:

```
    <moiety>
      <quantity>
        <numerator value="1" unit="1"/>
        <denominator value="1" unit="1"/>
      </quantity>
      <partMoiety>
        <id extension="3" root="000000000000-0000-0000-0000-00000000"/>
        <name>clivatuzumab light chain</name>


        <bond>
          <code code="C??????" codeSystem="2.16.840.1.113883.3.26.1.1"
                displayName="disulfide bond"/>
          <positionNumber value="23"/>
          <positionNumber value="89"/>
          <distalMoiety>
            <id extension="3" root="000000000000-0000-0000-0000-00000000"/>
          </distalMoiety>
        </bond>

        <bond>
          <code code="C??????" codeSystem="2.16.840.1.113883.3.26.1.1"
                displayName="disulfide bond"/>
          <positionNumber value="135"/>
          <positionNumber value="195"/>
          <distalMoiety>
            <id extension="3" root="000000000000-0000-0000-0000-00000000"/>
          </distalMoiety>
        </bond>
```

```
      <subjectOf>
        <characteristic>
          <id root="200000000000-0000-0000-0000-00000002"/>
          <code code="9999-9" codeSystem="2.16.840.1.113883.6.1"
                displayName="Amino Acid Sequence"/>
          <value xsi:type="ED" mediaType="application/x-aa-
seq">DIQLTQSPSSLSASVGDRVTMTCSASSSVSSSYLYWYQQKPGKAPKLWIYSTSNLASGVPARFSGSGSGTDFTLTIS
SLQPEDSASYFCHQWNRYPYTFGGGTRLEIKRTVAAPSVFIFPPSDEQLKSGTASVVCLLNNFYEAKVQWKVDNALQSGNSQ
ESVTEQDSKDSTYSLSSTLTLSPRKADYEKHKVYACEVTHQGLSSPVTKSFNRGEC</value>
        </characteristic>
      </subjectOf>
    </moiety>
```

And finally the second light chain with those same 2 internal disulfide bonds, again we reuse the sequence:

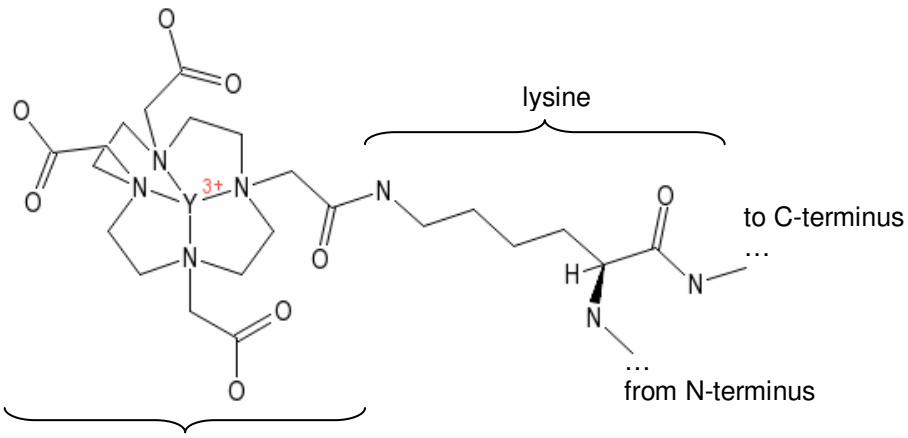
```
    <moiety>
      <quantity>
        <numerator value="1" unit="1"/>
        <denominator value="1" unit="1"/>
      </quantity>
      <partMoiety>
        <id extension="4" root="000000000000-0000-0000-0000-00000000"/>
        <name>clivatuzumab light chain</name>

        <bond>
          <code code="C??????" codeSystem="2.16.840.1.113883.3.26.1.1"
                displayName="disulfide bond"/>
          <positionNumber value="23"/>
          <positionNumber value="89"/>
          <distalMoiety>
            <id extension="4" root="000000000000-0000-0000-0000-00000000"/>
          </distalMoiety>
        </bond>

        <bond>
          <code code="C??????" codeSystem="2.16.840.1.113883.3.26.1.1"
                displayName="disulfide bond"/>
          <positionNumber value="135"/>
          <positionNumber value="195"/>
          <distalMoiety>
            <id extension="4" root="000000000000-0000-0000-0000-00000000"/>
          </distalMoiety>
        </bond>
      </partMoiety>

      <subjectOf>
        <characteristic>
          <id root="200000000000-0000-0000-0000-00000002"/>
        </characteristic>
      </subjectOf>
    </moiety>
```

This concludes the list of all the moieties. Now the antibody as a whole is also modified with those tetraxetan structures. The tetraxetan with the chelate bound yttrium (III) is connected to the lysine residues in the clivatuzumab as shown in the following diagram



Approximately 2-5 such lysine residues of the clivatuzumab antibody are connected to the marked amino-group of tetraxetan. This yttrium 90 tetraxetan is described in the following modification:

```
<moiety>
  <code code="C??????" codeSystem="2.16.840.1.113883.3.26.1.1"
        displayName="N-linked residual substitution, lysine, non-specific"/>
  <quantity>
    <numerator xsi:type="URG_PQ">
      <low value="2" unit="mol"/>
      <high value="5" unit="mol"/>
    </numerator>
    <denominator value="1" unit="mol"/>
  </quantity>

  <partMoiety>
    <id extension="5" root="000000000000-0000-0000-0000-00000000"/>
    <code code="1HTE449DGZ" codeSystem="2.16.840.1.113883.4.9"/>
    <name>tetraxetan</name>

    <moiety>
      <code code="C??????" codeSystem="2.16.840.1.113883.3.26.1.1"
            displayName="chelate-bound metal ion"/>
      <positionNumber value="1"/>
      <partMoiety>
        <code code="433ME2ATHW" codeSystem="2.16.840.1.113883.4.9"/>
        <name>yttrium 90 cation</name>
      </partMoiety>
    </moiety>

    <asSpecializedKind>
      <generalizedMaterialKind>
        <code code="C??????" codeSystem="2.16.840.1.113883.3.26.1.1"
              displayName="chemical chelator"
      </generalizedMaterialKind>
    </asSpecializedKind>
  </partMoiety>
```

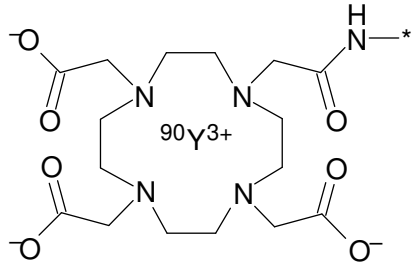
Note also that the tetraxetan moiety above was labeled with a functional class “chemical chelator”. Other such functional classes could be:

Name	Code	Description
chemical linker	C??????	a small moiety whose function is to
polymer linker	C??????	a polymer moiety whose function is to
chemical chelator	C??????	a small moiety whose function is to hold a metal ion in a chelate bond

polymer chelator	C??????	a polymer moiety whose function is to hold a metal ion in a chelate bond
protein conjugate	C?????	a protein moiety that has an intended pharmacological effect
toxoid conjugate	C?????	a modified toxin moiety whose function is to induce immune response
polymer conjugate	C?????	a polymer moiety that has the function of masking a main moiety from stimulating immune response or protecting the protein from breakdown (e.g. PEGylation)

In cases where the sequential order of modification matters (e.g., first a PEGylation then an acetylation), one will have to include a reference to

We describe this moiety with the MOLFILE format that encodes the following structure drawing:




```
      <subjectOf>
        <characteristic>
          <code code="9999-9" codeSystem="2.16.840.1.113883.6.1"
                displayName="Chemical Structure"/>
          <value xsi:type="ED" mediaType="application/x-molfile">30 29 0 0 0 0 0 0
0 0999 V2000 9.9563 -7.3055 0.0000 Y 1 1 0 0 0 0 0 0 0 0 0 0 15.0355 -4.8847
0.0000 * 0 0 0 0 0 0 0 0 0 0 0 0 13.3609 -8.0134 0.0000 O 0 0 0 0 0 0 0 0 0 0 0 0
13.8867 -9.9869 0.0000 O 0 5 0 0 0 0 0 0 0 0 0 0 0 6.4178 -6.8678 0.0000 O 0 0 0 0 0
0 0 0 0 0 0 0 5.8872 -4.8955 0.0000 O 0 5 0 0 0 0 0 0 0 0 0 0 6.7218 -5.7285
0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 13.0541 -9.1519 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
13.3408 -6.8634 0.0000 O 0 0 0 0 0 0 0 0 0 0 0 0 13.8599 -4.8881 0.0000 N 0 0 0 0
0 0 0 0 0 0 0 0 13.0301 -5.7260 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 5.9099 -9.9441
0.0000 O 0 5 0 0 0 0 0 0 0 0 0 0 6.4492 -7.9743 0.0000 O 0 0 0 0 0 0 0 0 0 0 0 0
6.7482 -9.1149 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 7.8605 -5.4221 0.0000 C 0 0 0 0 0
0 0 0 0 0 0 0 11.8897 -5.4263 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 11.9147 -9.4555
0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 7.8855 -9.4263 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
7.6897 -8.0305 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 7.6897 -6.8513 0.0000 C 0 0 0 0 0
0 0 0 0 0 0 0 8.7018 -6.2618 0.0000 N 0 0 0 0 0 0 0 0 0 0 0 0 9.2908 -5.2506
0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 10.4700 -5.2524 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
11.0577 -6.2664 0.0000 N 0 0 0 0 0 0 0 0 0 0 0 0 12.0761 -6.8427 0.0000 C 0 0 0 0
0 0 0 0 0 0 0 0 12.0891 -8.0218 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 8.7257 -8.5952
0.0000 N 0 0 0 0 0 0 0 0 0 0 0 0 11.0839 -8.6223 0.0000 N 0 0 0 0 0 0 0 0 0 0 0 0
10.4848 -9.6275 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 9.3057 -9.6139 0.0000 C 0 0 0 0 0
0 0 0 0 0 0 0 10 2 1 0 0 0 0 8 3 2 0 0 0 0 25 24 1 0 0 0 0 8 4 1 0 0 0 0 27 18 1 0
0 0 0 7 5 2 0 0 0 0 26 28 1 0 0 0 0 7 6 1 0 0 0 0 19 27 1 0 0 0 0 15 7 1 0 0 0 0
20 21 1 0 0 0 0 17 8 1 0 0 0 0 30 27 1 0 0 0 0 11 9 2 0 0 0 0 30 29 1 0 0 0 0 11
10 1 0 0 0 0 20 19 1 0 0 0 0 16 11 1 0 0 0 0 22 21 1 0 0 0 0 14 12 1 0 0 0 0 23 24
1 0 0 0 0 14 13 2 0 0 0 0 18 14 1 0 0 0 0 26 25 1 0 0 0 0 21 15 1 0 0 0 0 29 28 1
0 0 0 0 24 16 1 0 0 0 0 23 22 1 0 0 0 0 28 17 1 0 0 0 0 M CHG 4 1 3 4 -1 6 -1 12 -
1 M ISO 1 1 90 M END</value>
        </characteristic>
      </subjectOf>
    </modification>
  </identifiedSubstance>
```

From this core structure with all modifications specified above (except the glycosylation), follows a chemical sum formula of $C_{6496}H_{9952}N_{1716}O_{2014}S_{44} \cdot [C_{16}H_{23}N_4O_{790}Y]_n$.

```
  <subjectOf>
    <characteristic>
      <code code="9998-9" codeSystem="2.16.840.1.113883.6.1"
            displayName="Chemical Sum Formula"/>
      <value xsi:type="CV" code=" C6496H9952N1716O2014S44.(C16H23N4O790Y)n"
             codeSystem="1.2.3.99.9.8"/>
    </characteristic>
  </subjectOf>
```

The drug-target of this antibody is the human mucin 1 (MUC1) cell surface associated antigen:

```
  <directTargetOf classCode="SBR">
    <functionCode code="C00018" codeSystem="2.16.840.1.113883.3.26.1.1"
                  displayName="antibody"/>
    <interaction>
      <code code="C00011" codeSystem="2.16.840.1.113883.3.26.1.1"
            displayName="antigen-antibody binding"/>

      <directTarget classCode="SBR">
        <functionCode code="C00013" codeSystem="2.16.840.1.113883.3.26.1.1"
                      displayName="antigen"/>

        <presentSubstance classCode="LOCE">
```

this is specified completely using the SwissProt code P15941, identifying the genome as human so we can even specify the socper of the presentSubstance as the tumor cell (NCI Thesaurus has only "malignant cell" which it links to "pancreatic malignant neoplasm", but not "pancreatic malignant cell".

```
          <presentSubstance>
            <code code="P15941" codeSystem="2.16.840.1.99.9999"/>
            <name>Mucin-1</name>
          </presentSubstance>
          <scoper>
            <code code="C12917" codeSystem="2.16.840.1.113883.3.26.1.1"
                  displayName="malignant cell"/>
          </scoper>
        </presentSubstance>
      </directTarget>
    </interaction>
  </directTargetOf>
</identifiedSubstance>
```

This concludes the example.

Nucleic Acid Sequence

This category will only be used to described nucleic acids that have a length greater than three bases or base pairs; elements to be used for oligonucleotides, genes used in gene therapy and any nucleic acid aptomers

1. **Sequence Type:** DNA, RNA, or mixed.
2. **Subunit Number:** the number of strands that up the nucleic acid; in most gene examples only the transcribed strand will be described
3. **Subunit Group:**
   - 3.1. **Subunit Id:** numbered in order of decreased molecular weight
   - 3.2. **Length:** length of the sequence
   - 3.3. **Sequence:** actual nucleotide sequence notation from 5' to 3' end using standard single letter codes (G,C,T,A,U).
4. **Modification:** used to describe irreversible modifications to a nucleic acid
   - 4.1. **Modification:** general classification of modification (cv: physical, chemical, enzymatic)
     - 4.1.1. **Modification description group:** for each modification the specific modification will be described within a single group. each different modification will have a separate modification description group
       - 4.1.1.1. **description:** the specific modification (cv: PEGylation, phosphorothioate, arabino, etc)

4.1.1.2. **modification specificity:** cv: nonspecific, specific, unknown
4.1.1.3. **residue modification group:** for each modification has a single group
4.1.1.3.1. **residue modified:** specific nucleic acid residue modified or a modified nucleic acid residue within a chain
4.1.1.3.2. **residue site:** position of specific modifications (ie. 1_10 refers to the 10th residue on the 1st subunit)
4.1.1.4. **modification extent type group:** primarily used for nonspecific modification refers to how a modification is quantified or extent of physical treatment; will be used sparingly for nucleic acids
4.1.1.4.1. **modification extent reference:** cv: molecule, time, temperature
4.1.1.4.2. **amount group:** needed to express amount or extent of treatment (per molecule; time and temperature; ph and time)
4.1.1.4.2.1. **average:** average, if only limits are given, use the arithmetic mean
4.1.1.4.2.2. **low limit:**
4.1.1.4.2.3. **high limit:**
4.1.1.4.2.4. **unit:** associated with extent reference
4.1.1.5. **modification substance group:**
4.1.1.5.1. **modification substance role:** agent, refers to a chemical that results in nonspecific modifications of a nucleic acid or moeity, refers to a specific moeity added to a nucleic acid or that replaces a common nucleoside or nucleotide
4.1.1.5.2. **substance name:** displayed name
4.1.1.5.3. **substance id:** unii code

5. **component group:** used to describe genetic elements within a given gene; enhancer; promoter; coding sequence; termination signal; silencer etc; or molecules linked to nucleic acid; a functional gene would have multiple components
   5.1. **component class:** gene element; chemical moiety
   5.1.1. **component type:** enhancer; promoter etc.
   5.1.2. **component id:** unii code gene element
   5.1.3. **component name:** displayed name;
6. **isotope group:** applicable for substances that contain a radionuclide or a non-natural isotopic ratio (i.e c-13 enriched material) all radionuclide and non-natural isotope will also be represented in the structure field
   6.1. **nuclide name:** example 13c
   6.2. **nuclide id:** unii code for each non-natural or radionuclide isotope
   6.3. **substitution type:** specific (site of attachment/substitution indicated in structure); non-specific (nuclide distributed throughout molecule or substance); unknown (site unknown); extent of substitution not captured at substance level
7. **reference information:**
   7.1. **Target:** (as above) for antisense oligonucletides and aptomers the gene or protein targeted by the nucleic acid
   7.2. **gene group:** gene group is used to define the origin of the nucleic acid sequence
   7.2.1. **gene sequence origin:** the common name for the organism from which the final sequence originated.
   7.2.2. **gene id:** a numeric or alpha numeric id associated with the gene of origin
   7.2.3. **gene name:** complete name given for a gene
   7.2.4. **gene reference source:** the source from which the gene sequence was derived (genbank)
8. **Comment:** to be used infrequently

Structurally Diverse Substance

These are substances that are complex mixtures that cannot be fully described by enumerating and defining all their constituents. The majority of these substances are derived from a biological organism but they could be other complex natural materials such as coal tar or mineral oil.

For organism-based substances, the parent organism is essential defining information. Herbals are typically described by parent organism genus, species, and part or parts (e.g. flower, leaf, and stem) and an indication of the organism's life cycle (flowering, larvae, etc.). Some organisms require identification of subspecies, variety, strain, serovar, type, or cultivar group to accurately describe them and distinguish them from related substances. Time and place of harvest, type of soil and fertilizer, amount of daylight and water, and degree of plant maturity are also not

captured. Taxonomic information from NCBI (http://www.ncbi.nlm.nih.gov/Taxonomy/) or ITIS taxonomic identification numbers (http://www.itis.gov/) are helpful in parent organism identification. Catalog of Life (http://www.catalogueoflife.org) and Tropicos (http://www.tropicos.org/) are also valuable taxonomic resources.

Purified blood products (distinct clotting factors, human serum albumin) and monoclonal immunoglobulins are described as proteins. But fresh frozen plasma or polyclonal immunoglobulins are described as structurally diverse materials and require identification of the immunoglobulin type and targeted antigen. Cells and tissues are also described as structurally diverse substances. Information on individual donors or extent of pooling is not captured at the substance level.

Many natural substances are modified chemically, physically, or biologically. The process may be specified with a code (more or less specific) with a duration (effectiveTime, e.g., incubate for 40 hours), and with other parameters ("control variables") such as for temperature. For polysaccharide conjugate vaccine, components are used to describe the protein carrier, the linker if present and the component antigen along with the part of the organism the antigen was derived from. The type of conjugation chemistry and the identity of the conjugated chemical entity or entities are needed to describe the resulting structurally diverse conjugate.

Genetically-modified organism or cells will capture the inserted gene and resultant expressed protein as components of the parent cell or organism. The type of gene modification (transfection, transduction) will also be captured. Polyclonal immunosera will describe the targeted or immunized antigens as immunological modifications of normal sera.

1. **Source Material:** the material from which the final substance was originated from; (e.g., biological, organism, mineral)
    - 1.1. **Source Material Class:** the generally classification of the source material; (e.g., bacterium, human, fungus, virus, plantae; for vaccines this is the class of infectious agents) – this is implied by the taxonomy code.
    - 1.2. **Source Material Type:** for substance derived from the same species as the recipient (e.g., human source for human use) and then specifies whether the material is autologous or allogeneic. Xenogeneic is implied if the organism is specified as non human (for a product for human use.)
    - 1.3. **Source Material State:** e.g., live, inactivated, attenuated, conjugated, live attenuated; for inactivated vaccines, the inactivation method and agent is to be specified as a modification;
    - 1.4. **Organism Name and Code:** the name and code for the organism from an organism nomenclature / taxonomy.
    - 1.5. **Part:** the anatomical part of the organism used to produce the substance;
        - 1.5.1. **Part Location:** applies when the part can be extracted from different anatomical location of the organism; (e.g. for cartilage: knee, elbow)- repeatable
    - 1.6. **Developmental Stage:** stage of life for animals, plants, insects and microorganisms, e.g., fetal, juvenile, adult, larvae, sporon. Will only be captured when the substance is significantly different in these stages (e.g., fetal bovine serum).
    - 1.7. **Component:**
        - 1.7.1. **Component Class:** general classification of the component derived from the source material (e.g. cell, for plasma derived product, blood is the part and plasma or sera is the component); for herbals this refers to any component or form derived from plants/animals/minerlas and not processed (e.g. oil, juice). for conjugated vaccines linker and carrier applies.for vaccines this is may describe if applicable the antigen characterisation (e.g. whole cell, split virion, surface antigen).cv.
        - 1.7.2. **Component Type:** refers to the specific type of the material constituting the component(e.g. plasmid, extrachromosomal, chondrocyte, lipase, triglycerides; cv.
        - 1.7.3. **Component id:** UNII code
        - 1.7.4. **Component name:** display name; primarily used for gene therapy where the component is described within the nucleic acid description (e.g. p-llo-e7; hpv-16); for cell therapy this expresses the protein which is expressed within the relevant cell (e.g. il12); for conjugated vaccines this describe the organism strain of the carrier; for vaccines, this may identify the antigen used in the vaccine.
    - 1.8. **Modification Extent Type:** primarily used for nonspecific modification refers to how a modification is quantified or extent of physical treatment
        - 1.8.1. **Modification Extent Reference:** e.g., molecule, time, temperature

1.8.2. **Amount:** (as above)

2. **Modification:** specifies irreversible modifications to a substance. One special example of this is a plasmid that was added to a cell or a vector virus, either as a free-floating plasmid or one integrated into the chromosome.
   2.1. **Description:** a code for the specific modification (e.g. PEGylation, phosphorylation, transfaction/transduction of a plasmid, etc)
   2.2. **Modification Specificity:** nonspecific, specific, unknown
   2.3. **Modification Substance:**
      2.3.1. **Modification Substance Role:** for proteins, agent (refers to a chemical that results in nonspecific modifications of a protein)
      2.3.2. **Substance Name:** displayed name
      2.3.3. **Substance Id:** UNII code
3. **Properties:** (as above)
4. **Isotope Group:** (as above)
5. **Reference Information:** (as above)
6. **Comment:** to be used infrequently

Example: Celvapan (H1N1 Influenza Vaccine)

This example shows a pandemic influenza vaccine for the H1N1 strain using a whole inactivated virus (vero-cell derived inactivated whole virion). This substance has no established name, but an English primary name is given as "A/CALIFORNIA/7/2009 (H1N1) V-LIKE STRAIN (X-179A) (WHOLE VIRION, PROPIOLACTONE, INACTIVATED, MAMMALIAN EXPRESSED)" by the US FDA Substance Registration System (SRS).

```
<identifiedSubstance classCode="IDENT">
  <id extension="..." root="2.16.840.1.113883.4.9"/>
  <identifiedSubstance classCode="MMAT" determinerCode="KIND">
    <code code="..." codeSystem="2.16.840.1.113883.4.9"/>
    <name>A/CALIFORNIA/7/2009 (H1N1) V-LIKE STRAIN (X-179A) (WHOLE VIRION,
PROPIOLACTONE, INACTIVATED, MAMMALIAN EXPRESSED)</name>

    <asNamedEntity>
      <code code="C??????" codeSystem="2.16.840.1.113883.3.26.1.1"
            displayName="primary (non-established) name"/>
      <name xml:lang="en_US">A/CALIFORNIA/7/2009 (H1N1) V-LIKE STRAIN (X-179A)
(WHOLE VIRION, PROPIOLACTONE, INACTIVATED, MAMMALIAN EXPRESSED)</name>
      <assigningOrganization>
        <id extension="FDA-SRS" root="2.16.840.1.113883.9.9.9"/>
        <name>Food and Drug Administration Substance Registration System</name>
        <territorialAuthority>
          <territory>
            <code code="USA" root="1.0.3166.2.2.3"/>
          </territory>
        <territorialAuthority>
      </assigningOrganization>
    </asNamedEntity>
  <identifiedSubstance>

  <productOf>
    <derivationProcess>
      <code code="C?????" codeSystem="2.16.840.1.113883.3.26.1.1"
            displayName="Propiolactone Inactivates"/>
      <effectiveTime>
        <width value="40" unit="h"/>
      </effectiveTime>
```

```
      <directTarget classCode="CSM">
        <presentSubstance>
          <substance>
            <code code="C12954" codeSystem="2.16.840.1.113883.3.26.1.1"
                  displayName="peripheral blood mononuclear blood cell"/>
          </substance>
          <scoper>
            <code code="..." codeSystem="2.16.840.1.113883.4.999999"
                  displayName="INFLUENZA A/CALIFORNIA/7/2009(H1N1)-LIKE"/>
          </scoper>

          <subjectOf>
            <characteristic>
              <code code="9992-9" codeSystem="2.16.840.1.113883.6.1"
                    displayName="Organism Vital Status"/>
              <value xsi:type="CV" code="C00009" displayName="living"
                     codeSystem="2.16.840.1.113883.3.26.1.1"/>
            </characteristic>
          </subjectOf>

          <subjectOf>
            <characteristic>
              <code code="9991-9" codeSystem="2.16.840.1.113883.6.1"
                    displayName="Tissue Source Type"/>
              <value xsi:type="CV" code="C28000" displayName="autologous"
                     codeSystem="2.16.840.1.113883.3.26.1.1"/>
            </characteristic>
          </subjectOf>

        </presentSubstance>
      </directTarget>
```

This process also involves a protein known as the "Prostatic acid phosphatase (PAP) granulocyte-macrophage colony-stimulating factor (GM-CSF) fusion protein" which is its own identified substance N5E5Q8249O

```
      <directTarget classCode="CSM">
        <identifiedSubstance>
          <id extension="N5E5Q8249O" root="2.16.840.1.113883.4.9"/>
        </identifiedSubstance>
      </directTarget>
    </derivationProcess>
  </productOf>
```

Example: Sipuleucel-T

"Sipuleucel-T" is an English established name defined for the USA by the authority of USAN according to the USP Dictionary 2009. It is a proprietary substance and method to which the company gives its own code "APC8015" and which is described in US Patent 6194152

```
<identifiedSubstance classCode="IDENT">
  <id extension="..." root="2.16.840.1.113883.4.9"/>
  <identifiedSubstance classCode="MMAT" determinerCode="KIND">
    <code code="..." codeSystem="2.16.840.1.113883.4.9"/>
    <name>Sipuleucel-T</name>
```

```
    <asNamedEntity>
      <code code="C??????" codeSystem="2.16.840.1.113883.3.26.1.1"
            displayName="established name"/>
      <name xml:lang="en_US">Sipuleucel-T</name>
      <assigningOrganization>
        <id extension="USAN" root="2.16.840.1.113883.9.9.9"/>
        <name>United States Adopted Names Council</name>
        <territorialAuthority>
          <territory>
            <code code="USA" root="1.0.3166.2.2.3"/>
          </territory>
        <territorialAuthority>
      </assigningOrganization>
      <subjectOf>
        <document>
          <title>USP DICTIONARY 2009</title>
        </document>
      </subjectOf>
    </asNamedEntity>

    <asEquivalentEntity>
      <definingMaterialKind>
        <code code="APC8015" codeSystem="1.3.6.1.4.1.9999999.99"/>
      </definingMaterialKind>
    </asEquivalentEntity>
  <identifiedSubstance>

  <subjectOf>
    <document>
      <id extension="6194152" codeSystem="1.2.3.9988.2"/>
      <bibliographicDesignationText>Laus, et al. Prostate tumor polynucleotide
compositions and methods of detection thereof [patent]. February 27, 2001,
USPTO</bibliographicDesignationText>
    </document>
  </subjectOf>
```

This is a biologic cancer therapy method using antibodies created from autologous human blood cells exposed to a tumor antigen. This antibody is created through a *cultured antigen loading* process for 40 hours. The source material is live autologous mononuclear cells from the blood of a human (the patient himself, hence autologous.) Human organism is specified using the Integrated Taxonomic Information System (ITIS) Taxonomic Serial Number (TSN) 180092. The concepts for "peripheral blood mononuclear cell" is taken from the NCI Thesaurus.

```
  <productOf>
    <quantity value="1" unit="mol"/>

    <derivationProcess>
      <code code="C00009" codeSystem="2.16.840.1.113883.3.26.1.1"
            displayName="Cultured Antigen Loading"/>
      <effectiveTime>
        <width value="40" unit="h"/>
      </effectiveTime>

      <directTarget classCode="CSM">
        <presentSubstance>
          <substance>
            <code code="C12954" codeSystem="2.16.840.1.113883.3.26.1.1"
                  displayName="peripheral blood mononuclear blood cell"/>
          </substance>
          <scoper>
            <code code="180092" codeSystem="2.16.840.1.113883.4.999999"
                  displayName="homo sapiens"/>
          </scoper>
```

```
          <subjectOf>
            <characteristic>
              <code code="9992-9" codeSystem="2.16.840.1.113883.6.1"
                    displayName="Organism Vital Status"/>
              <value xsi:type="CV" code="C00009" displayName="living"
                     codeSystem="2.16.840.1.113883.3.26.1.1"/>
            </characteristic>
          </subjectOf>

          <subjectOf>
            <characteristic>
              <code code="9991-9" codeSystem="2.16.840.1.113883.6.1"
                    displayName="Tissue Source Type"/>
              <value xsi:type="CV" code="C28000" displayName="autologous"
                     codeSystem="2.16.840.1.113883.3.26.1.1"/>
            </characteristic>
          </subjectOf>

        </presentSubstance>
      </directTarget>
```

This process also involves a protein known as the "Prostatic acid phosphatase (PAP) granulocyte-macrophage colony-stimulating factor (GM-CSF) fusion protein" which is its own identified substance N5E5Q8249O

```
      <directTarget classCode="CSM">
        <identifiedSubstance>
          <id extension="N5E5Q8249O" root="2.16.840.1.113883.4.9"/>
        </identifiedSubstance>
      </directTarget>
    </derivationProcess>
  </productOf>
```

The resulting substance is defined as undergoing antigen-antibody binding reaction with human prostate tumor cells which is described in a publication in Expert Rev. Anticancer Ther. 6:1163-1167 (2006)

```
  <directTargetOf classCode="SBR">
    <interaction>
      <code code="C00010" codeSystem="2.16.840.1.113883.3.26.1.1"
            displayName="antibody-antigen binding reaction"/>
      <directTarget classCode="SBR">
        <presentSubstance>
          <presentSubstance>
            <name>PROSTATE TUMOR CELLS</name>
          </presentSubstance>
          <scoper>
            <code code="180092" codeSystem="2.16.840.1.113883.4.999999"
                  displayName="homo sapiens"/>
          </scoper>
        </presentSubstance>
      </directTarget>

      <subjectOf>
        <document>
          <id extension="6194152" codeSystem="1.2.3.9988.2"/>
          <bibliographicDesignationText>Expert Rev. Anticancer Ther. 6:1163-1167
(2006)</bibliographicDesignationText>
        </document>
      </subjectOf>
    </interaction>
  </directTargetOf>
```

Mixtures

Multiple substances can be mixtures if they are isolated or synthesized together. Racemic mixtures or substances containing unknown or mixed stereochemistry will not be defined as mixtures. Substances that contain impurities or degradents will not be described as mixtures. To avoid confusion and database problems, mixtures of mixtures will not be allowed. Each component of a mixture should be listed. Substances present in trace amounts will not be listed in a mixture unless they are known to have a specific effect. mixtures are also used when substance ambiguity exists in authoritative sources (aloe)

1. **Mixture type:** one of, all of, any of
2. **Ingredient group:** specifies the presence of each component (ingredient) in a mixture
   2.1. **Ingredient Name:** display name of substance
   2.2. **Ingredient id:** UNII for substance
   2.3. **Required Component:** a flag indicating if this component is required.
3. **Target:** (as above) for antisense oligonucletides and aptomers the gene or protein targeted by the nucleic acid

Specified Substance

The Specified Substance describe further characteristics of a single substance or multiple substances. One way to conceptualize Specified Substance is as an "intermediate product", as a product which is used as a substance. Another way of conceptualizing specified substance is as a specification added to a substance when it is used as an ingredient in a product.

1. **Specified Substance Id:** code
2. **Specified substance Name:** SS display name
3. **Physical Form:** detailed form of the final specified substance (e.g. crystalline amorphous, tinticure, dry extract), the state of matter, slightly more detailed form of the final specified substance (e.g., solid, liquid, gas, emulsion).Biphasic insulin is one example of a crystal in which part is dissolved and part is in solution. There are also instances of substances being partially crystalline and partially amorphous which would be two different form types or different polymorphic crystalline type
4. **Constituent:** the substance(s) or specified substances that are mixed together to form the specified substance
   4.1. **Role:** refers to the role of the substance in the specified substance (e.g. adjuvant, impurity, degradant, additive)
   4.2. **Amount:**
   4.3. **Substance (or Specified Substance):** refers to the elements describing the substance
      4.3.1. **Substance Id:** UNII code for the substance or specified substance id
      4.3.2. **Substance Name:** display name of substance or specified substance
      4.3.3. **Physical Form:** detailed form of the constituent substance, as above. Biphasic insulin is one example of a crystal in which part is dissolved and part is in solution. There are also instances of substances being partially crystalline and partially amorphous which would be two different form types or different polymorphic crystalline type
5. **Manufacturing:** the manufacturing information of the substance/specified substance
   5.1. **Manufacturer Name:** display name of the organisation that produce the final specified substance(s) – if the specified substance is given in the abstract
   5.2. **Manufacturer:**
   5.3. **Manufacturing Type:** e.g., recombinant, synthetic, extracted
   5.4. **Critical Process Step -** this intends to drill down into the details of the manufacturing process in an organized manner.
      5.4.1. **Process Code:** a code specifying the nature of the process step (e.g., cell culture, purification, extraction, synthesis)
      5.4.2. **Critical Process Organization and Facility: physical plant where process is performed**
      5.4.3. **Process Starting Materials**
         5.4.3.1. Starting Material Substances or Specified Substances
            5.4.3.1.1. **Starting Material Specifications**
            5.4.3.1.2. **Starting Material Amounts**

5.4.4. Critical Process Processing Materials – materials with which the substance comes in contact during that process, e.g. extraction solvents, spray solvent, growth media, chromatographic matrix. These are products by themselves (e.g., again they have manufacturers, ingredients).

5.4.5. Critical Process Equipment – devices used for the production process step, e.g. glass tubing, stainless steal reactor, polyethylene. These devices in turn are products by themselves.

5.4.5.1. Equipment Manufacturer

5.4.5.2. Equipment Model

5.4.6. Critical Process Parameter

5.4.6.1. **Parameter Code** (Extraction-concentration, time, temperature; chromatography-loading solvent, eluting solvents)

5.4.6.2. Parameter Value (code or amount/range)

5.4.7. Final Material of this Process Step

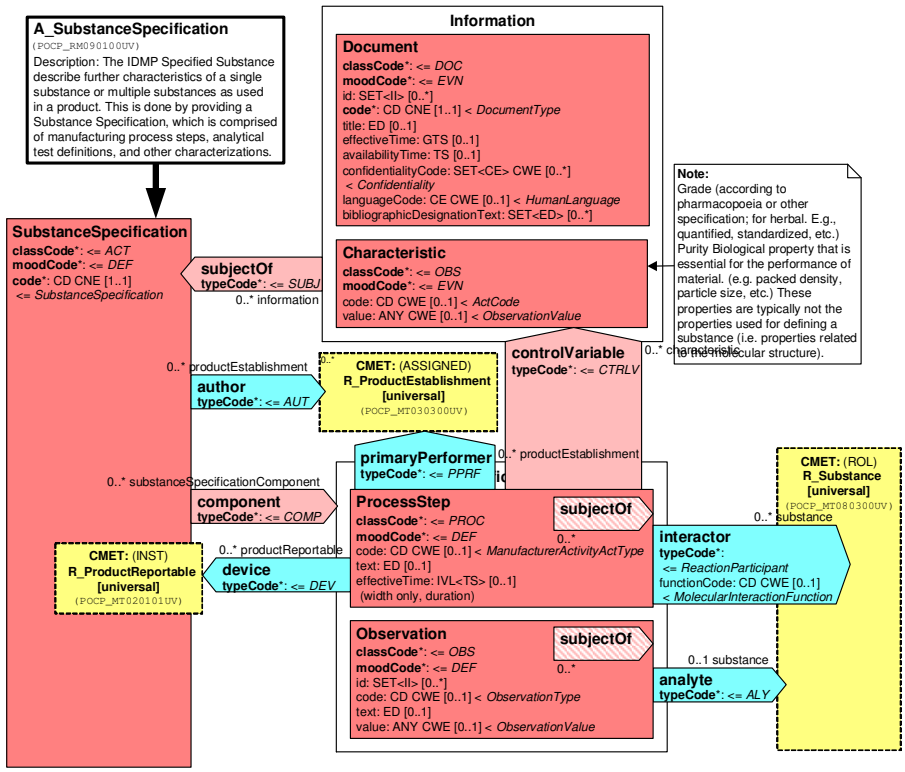
5.4.7.1. Final Material Substances or Specified Substances

5.4.7.1.1. **Final Material Specifications**

6. **Analytical Data**
   - 6.1. **Analytical Data Role:** e.g. for determining Potency, Identity, Purity, and Limit test.
   - 6.2. **Analytical Method Type:** Spectroscopic, Chromatographic, Microbiological etc.
     - 6.2.1. **Analytical Method Subtype:** NMR, LC-MS, Elisa
     - **6.2.2. Analytical Method Description**
       - **6.2.2.1. Analytical Method Reference Source (e.g., USP, EP, Company)**
         - **6.2.2.1.1. Analytical Method Reference Data (Validation, Qualification Data)**
     - **6.2.3. Reference Standard - Specified substance**
   - **6.3. Analytical Method Result Specifications;** gives the tolerance ranges.
7. **Grade:**
   - 7.1. **Grade Type:** pharmacopoeia specification type or other specification type; for herbal. E.g., quantified, standardized, etc.
   - 7.2. **Grade Reference Source:** refers to the relevant reference source/monograph (e.g., USP, EP, ACS, technical)
8. **Properties:**
   - 8.1. **Property:** physical property or biological property that is essential for the performance of material. (e.g. packed density, particle size, etc.) These properties are typically not the properties used for defining a substance (i.e. properties related to the molecular structure).
   - 8.2. **Non-numeric Value:** (e.g., plus or minus)
   - 8.3. **Amount:** (as above)

(The original has here a Units of Measure data element at the end, but that does not make sense to have one unit of measure element that somehow applies for all of the annotations.)

Model



The specified substance CMET annotates Ingredients in Products with a detailed Specification. The Specification describes characteristic properties, monograph documents, details of manufacturing specification as well as the definition of analytic observations.