The pathology markup language (PathoML)

Language Specification for Level 1

Chen Li	*Xi'an Jiaotong University, China*
Peiliang Lou	*Xi'an Jiaotong University, China*
Chunbao Wang	*The First Affiliated Hospital of XJTU, China*
Caixia Ding	*Shaanxi Provincial Tumor Hospital, China*
Ruifeng Guo	*Mayo Clinic, USA*
Lixia Yao	*Temple University, USA*
Guanjun Zhang	*The First Affiliated Hospital of XJTU, China*
Jun Yang	*The Second Affiliated Hospital of XJTU, China*
Yong Yuan	*Shaanxi Provincial Tumor Hospital, China*

20 October 2022

Contents

1 Introduction

This document defines the PathoML level 1 ontology. It provides definitions of the PathoML ontology classes, object properties and data properties. As for how to use PathoML, we provide some example representations and best practice recommendations. PathoML level 1 OWL file could be downloaded from https://PathoML.com/.

1.1 Document Convention

We use the following typographical conventions to distinguish classes, properties from other entities:

a) **`Class`**: Class names start with an uppercase character and are highlighted in bold in the text of this document;

b) *`property`*: Property names start with a lowercase character. *`Object properties`* are italicized while `datatype properties` are not;

c) The structure “object:ClassName” refers to an individual of a class, used to illustrate the range of an object property.

1.2 UML notation

In this document, we use UML, the Unified Modeling Language, as a basis for defining PathoML classes and properties.

1.2.1 PathoML Class Diagram

PathoML uses UML class notation to define PathoML ontology classes. Classes in UML class notation are drawn as simple tripartite boxes, as UML allows for operators

as well as data attributes to be defined. But PathoML only uses data attributes (i.e. datatype properties), so all PathoML class diagrams use only the top two portions of a UML class box as Figure 1-1 shows.

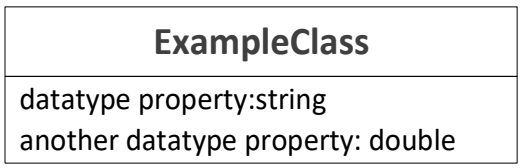

Figure 1-1 Example of a PathoML class diagram.

1.2.2 PathoML Object Property

PathoML uses UML Composition to represent a PathoML object property which links a PathoML class to another one. In PathoML, UML Composition represents different object properties in addition to the composition relationship. Figure 1-2 gives an example.

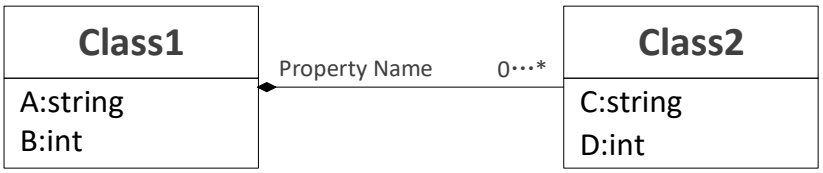

Figure 1-2 Example illustrating a PathoML object property

The line with the black diamond indicates a PathoML object property. If this property links Class1 to Class2, the diamond would locate on the Class1 side. The name of this property is on the line. Numbers are placed above the line near the Class2 side to indicate how many instances can be contained. The common cases in PathoML are the following: [0..*] signifies a list containing zero or more; [1..*] signifies a list containing at least one; and [0..1] signifies exactly zero or one. The absence of a numerical label means “exactly 1”.

1.2.3 Inheritance

PathoML classes can inherit properties from other classes, and inheritance in PathoML involves object and datatype properties from a parent class being inherited by the child classes. Inheritance is indicated by a line between two classes, with an open triangle next to the parent class. Figure 1-3 gives an example.

Figure 1-3 Inheritance

2 PathoML Ontology Class Structure

In this section, we define each class of PathoML Level 1. Text definitions of the classes are provided along with the synonyms, comments and examples to help readers understand the definition and intended use of each class. The most specific class available should be used.

PathoML Level 1 ontology has three root classes which are **Entity**, **Utility** and **Data**. Figure 2-1 shows a high-level view of **Entity** and its subclasses, Figure 2-2 shows a high-level view of **Utility** and its subclasses and Figure 2-3 shows a high-level view of **Data**. PathoML classes are shown as boxes and the arrows represent subclass relationships.

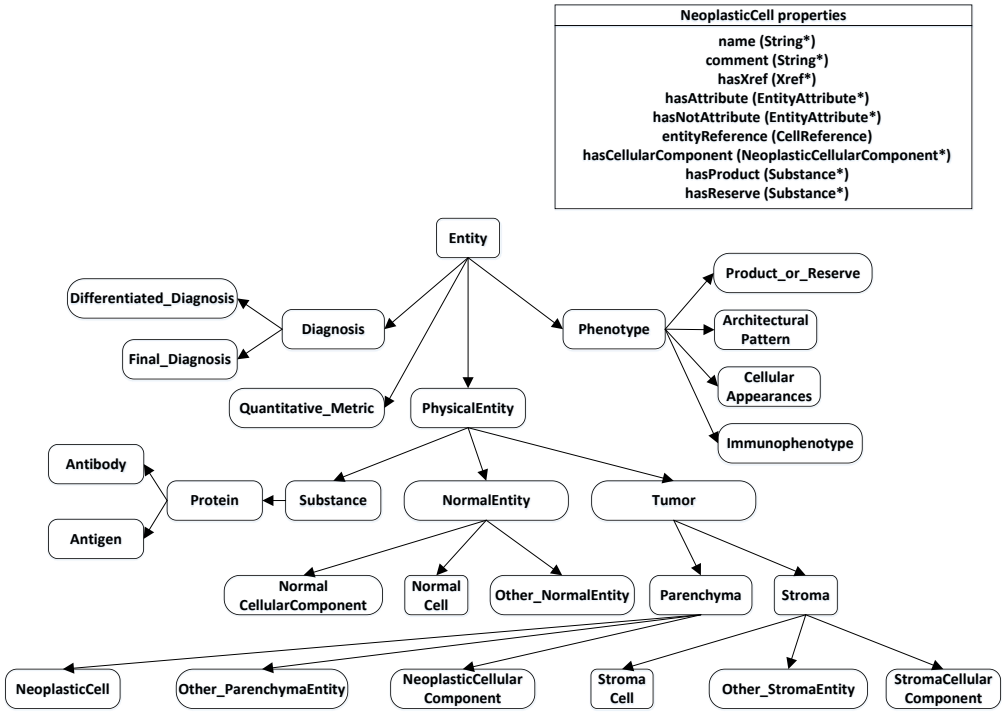

Figure 2-1 High-level view of **Entity** and its subclasses

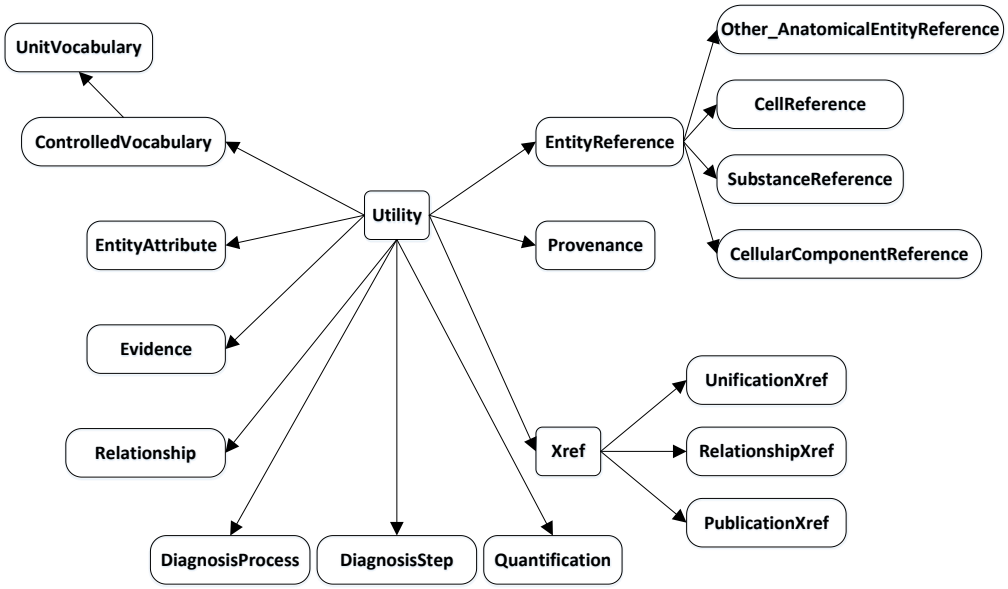

Figure 2-2 High-level view of **Utility** and its subclasses

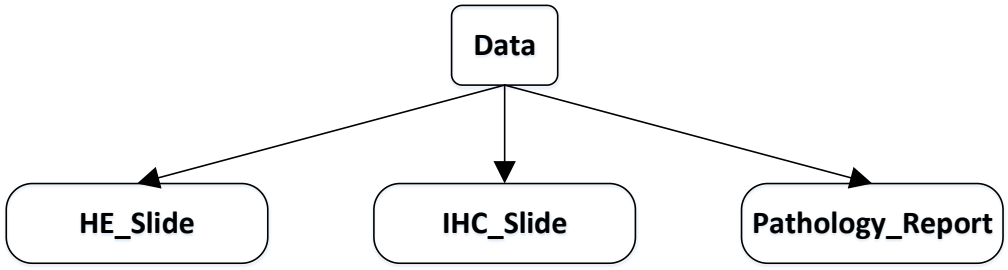

Figure 2-3 High-level view of **Data** and its subclasses

2.1 Top level entity class

The PathoML ontology defines four basic classes: the root level **Entity** class and its three subclasses: **Phenotype**, **PhysicalEntity**, **Diagnosis**. A phenotype refers to a set of physical entities or one entity with some specific traits. Different types of phenotypes are defined as children of the **Phenotype** class. Involving phenotypes as evidences, individual diagnoses are defined using **Diagnosis**.

2.2 Entity

Entity is the root class of PathoML. Its definition is shown in Figure 2-4.

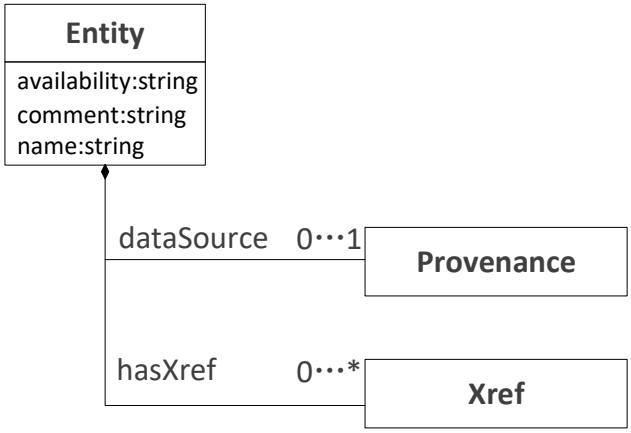

Figure 2-4

availability: (xsd:string) Describes the availability of this data (e.g. a copyright statement).

comment: (xsd:string) Comment on the data in the container class.

name: (xsd:string) One or more names of this entity. This will automatically include values of the displayName and standardName properties, as they are child properties of the name. displayName values are short names suitable for display in a graphic. standardName values are names that follow a standard nomenclature.

datasource: (0 or 1 object:Provenance) A description of the source of this data, e.g. a database or person name.

hasXref: (0 or more object:Xref) Values of this property define external cross-references from this entity to entities in external databases (e.g. controlled vocabulary).

2.3 PhysicalEntity

PhysicalEntity class is used to represent microscopically observable entities in

a haematoxylin and eosin (HE) slide and an immunohistochemistry (IHC) slide. According to their different types, **PhysicalEntity** has several subclasses covering cells, cellular components, substances (e.g. product or reserve of a cell, antigen), tissues, parts of organs (e.g. the glomerulus of the kidney) and immaterial morphological structures(e.g. a cavity or duct). It's further categorized into 'NormalEntity' and 'Tumor' based on whether an entity is neoplastic or induced by tumor or not (e.g. blood vessels induced by tumor belong to 'Tumor' instead of 'NormalEntity').

To represent a phenotype, a physical entity in diverse states should be used. For example, in a HE slide, a tumor cell is a generic physical entity while there are different forms of it such as a multinucleate giant tumor cell, a tumor cell having cytoplasm with a cyst, a tumor cell having evenly distributed chromatin. Under one slide, it is likely for you to observe more than one of those entities. To support representing physical entities in diverse states, a generic physical entity is represented using the **EntityReference** class which stores constant attributes of the entity while its different forms are represented using **PhysicalEntity** class which stores the variable features. The **EntityReference** is referenced from the **PhysicalEntity**. This design makes it easier to create different forms of an entity while not duplicating information common to all forms and explicitly linking all forms of an entity together through the shared **EntityReference**. Section 2.8.5 provides more information about **EntityReference** and principles for the use of this class. The definition of **PhysicalEntity** is shown in Figure 2-5.

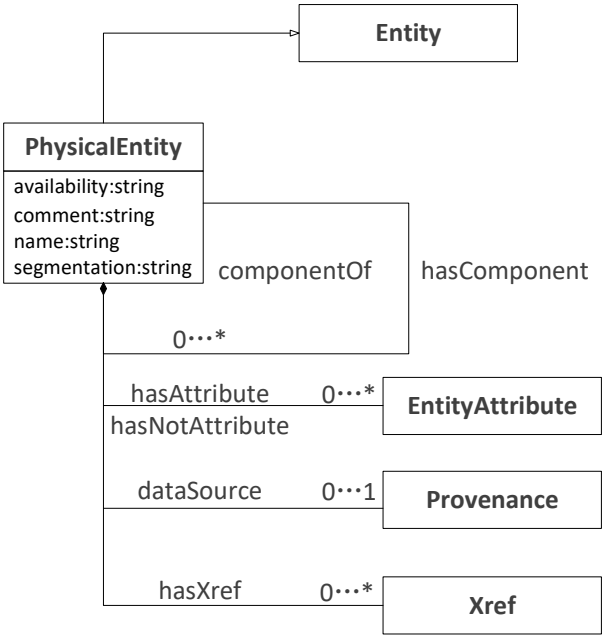

Figure 2-5

hasComponent: (0 or more object:PhysicalEntity) Used to define the components of a physical entity. Take a cell as an example, a cell consists of cytoplasm, membrane, nucleus, nucleolus and chromatin etc. *hasComponent* has three child properties which are *hasCell*, *hasCellularComponent* and *hasAnatomicalEntity*.

componentOf: (0 or more object:PhysicalEntity) Inverse property of *hasComponent*. Similarly, *componentOf* has three child properties which are *cellOf*, *cellularComponentOf* and *anatomicalEntityOf*. Since *componentOf* is inverse property of *hasComponent*, when representing an entity using PathoML, either *hasComponent* or *componentOf* is specified, the other could be inferred.

hasAttribute: (0 or more object:EntityAttribute) Attributes of the owner physical entity. For example, shape, size or other chemical attributes such as eosinophilic. A set of attributes helps define the state of an entity.

hasNotAttribute: (0 or more object: EntityAttribute) Attributes of this physical entity which are known to be lacking. Attributes not specified are not known to be absent and

only Attributes known to be lacking should be specified using this property.

`segmentation`: (xsd:string) The identifier of the segmentation or the annotation mask of this entity in the digital slide.

hasXref: (0 or more object:Xref) Values of this property define external cross-references from this entity to entities in external databases (e.g. controlled vocabulary). PathoML recommends using "Normal anatomical Entity" as well as its sub-classes from Tumor Pathology Ontology.

2.4 Phenotype

Phenotype class in PathoML is used to represent histopathological phenotypes in HE slides that prompt pathologists to look more closely, and immunophenotypes in IHC slides. **Phenotype** class has five subclasses, **Cellular_Appearances**, **Product_or_Reserve**, **Architecturel_Pattern** and **Immunophenotype**, **Quantitative_Metric** covering all different levels of phenotypes in a HE slide and IHC slide. **Cellular_Appearances** and **Product_or_Reserve** are used to describe a physical change of a cell or a subcellular structure observed microscopically at high power. **Architecturel_Pattern** is used to describe histologic patterns of cell populations and tumor behaviors (e.g. extension, invasion) which are usually observed at medium and low power. **Immunophenotype** describes the expression patterns of antigens in the context of cells and tissues. **Quantitative_Metric** describes quantitative measurements of phenotypes such as tumor-stroma ratio for characterizing intra-tumoral stroma and the semi-quantitative scoring for assessing HER2 expression. In PathoML, a phenotype is represented in which each of the individual components is described, including their properties (e.g. size, length, the extent and intensity of protein expression), relationships (e.g. membership) and behaviors (e.g. invasion). More information about these PathoML classes is given in Section 2.6. The definition of

Phenotype is shown in Figure 2-6.

hasXref: (0 or more object:Xref) Values of this property define external cross-references from this entity to entities in external databases (e.g. controlled vocabulary). PathoML recommends using "Microscopic Finding" as well as its sub-classes from Tumor Pathology Ontology.

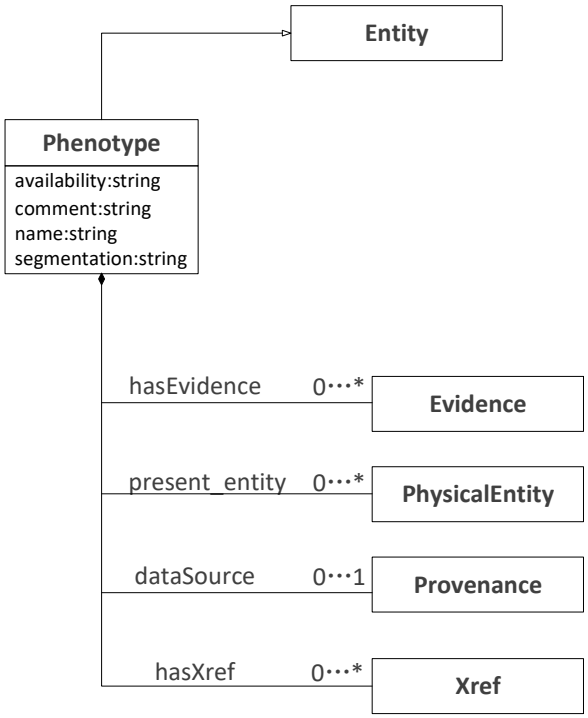

Figure 2-6

2.5 PhysicalEntity Subclass

The structure of **PhysicalEntity** is shown in Figure 2-7. It has three child classes which are **Tumor**, **NormalEntity** and **Substance**. The entities which compose a tumor are categorized as **Tumor** whereas the others are categorized as **NormalEntity**; any molecular entity is categorized as **Substance**. Based on the two compartments of a tumor, **Tumor** are further divided into **Parenchyma** and **Stroma**. Tissues and other anatomical structures (e.g. a cavity or duct) are described using

Other_AnatomicalEntity (e.g. **Other_NormalEntity**). Correspondingly, **EntityReference** has four child classes which are **Other_AnatomicalEntityReference**, **CellReference**, **CellularComponentReference**, and **SubstanceReference**. **Tumor**, **Parenchyma**, **Stroma** and **NormalEntity** don't have corresponding **EntityReference** classes.

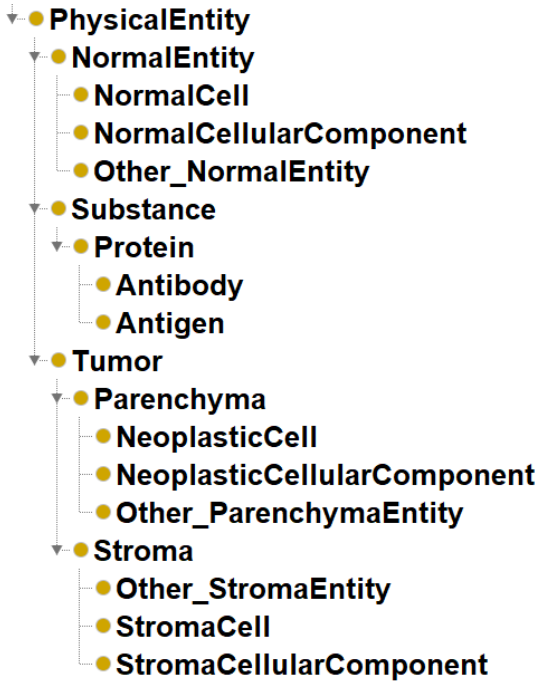

Figure 2-7

2.5.1 Tumor

Tumor refers to a benign or malignant pathologic structure in any part of the body, resulting from a neoplastic accumulation of cells. The definition of **Tumor** is shown in Figure 2-8.

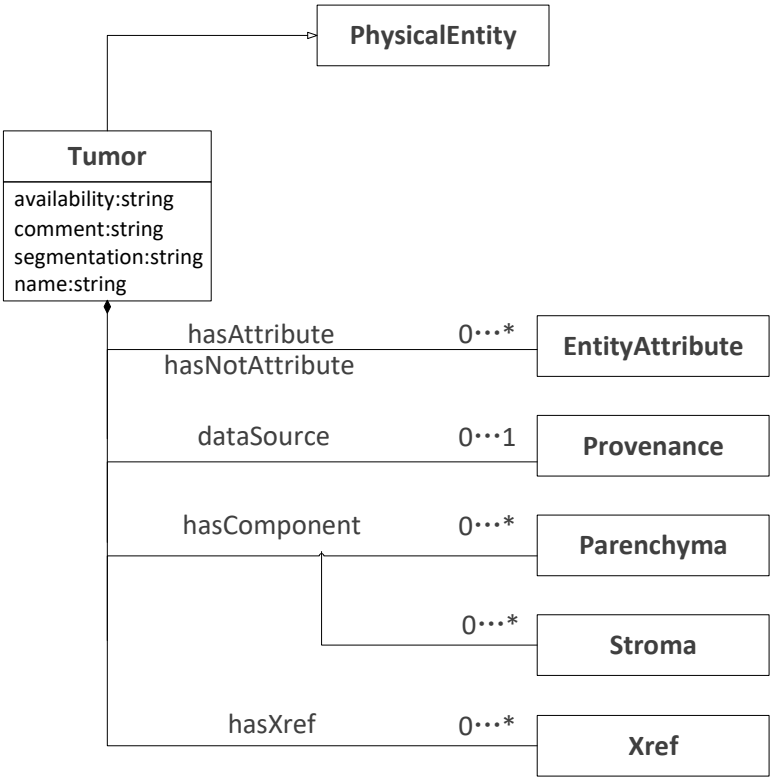

Figure 2-8

hasComponent: (0 or more object:Prenchyma or object:Stroma) Used to define tumor components such parenchyma or stroma.

hasXref: (0 or more object:Xref) Values of this property define external cross-references from this entity to entities in external databases (e.g. controlled vocabulary). PathoML recommends using "Tumor Tissue Sample" as well as its sub-classes from Tumor Pathology Ontology.

2.5.2 Parenchyma

The parenchyma of a tumor is one of the two distinct compartments in a tumor; the other compartment is the stroma. The parenchyma is made up of neoplastic cells. Neoplastic cells and morphologic structures in the parenchyma are described using **Parenchyma**. **Parenchyma** has three child classes which are **NeoplasticCell**, **NeoplasticCellularComponent** and **Other_ParenchymaEntity**. The definition of

Parenchyma is shown in Figure 2-9.

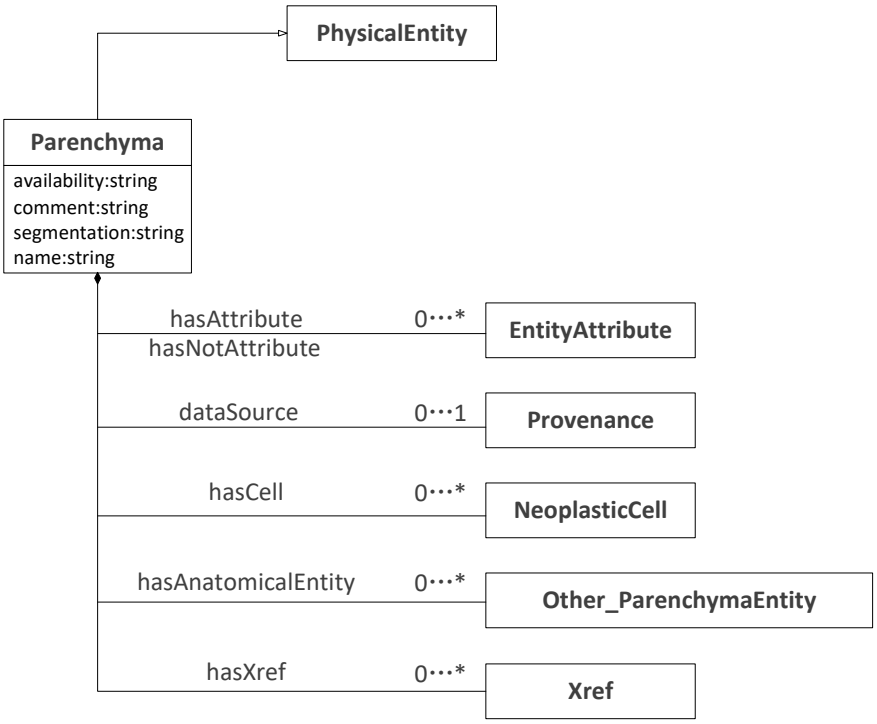

Figure 2-9

hasCell: (0 or more object:NeoplasticCell) Used to define tumor cells in the parenchyma.

hasAnatomicalEntity: (0 or more object:Other_ParenchymaEntity) Used to define morphologic structures in the parenchyma.

hasXref: (0 or more object:Xref) Values of this property define external cross-references from this entity to entities in external databases (e.g. controlled vocabulary). PathoML recommends using "Tumor Parenchyma" from Tumor Pathology Ontology.

（1）NeoplasticCell

An instance of **NeoplasticCell** refers to a tumor cell. Its definition is shown in Figure 2-10.

hasCellularComponent: (0 or more object:NeoplasticCellularComponent) Used to define cellular components of the cell including its cytoplasm, nucleus, nucleolus.

hasAnatomicalEntity: (0 or more object:Other_ParenchymaEntity) Used to define morphologic structures (e.g. cavity) in the tumor cell.

hasProduct: (0 or more object:Substance) Used to define the product of the cell.

hasReserve: (0 or more object:Substance) Used to define the reserve of the cell.

entityReference: (0 or 1 object: CellReference) The entity reference stores the base definition of the cell.

hasXref: (0 or more object:Xref) Values of this property define external cross-references from this entity to entities in external databases (e.g. controlled vocabulary). PathoML recommends using "Neoplastic Cell" as well as its sub-classes from Tumor Pathology Ontology.

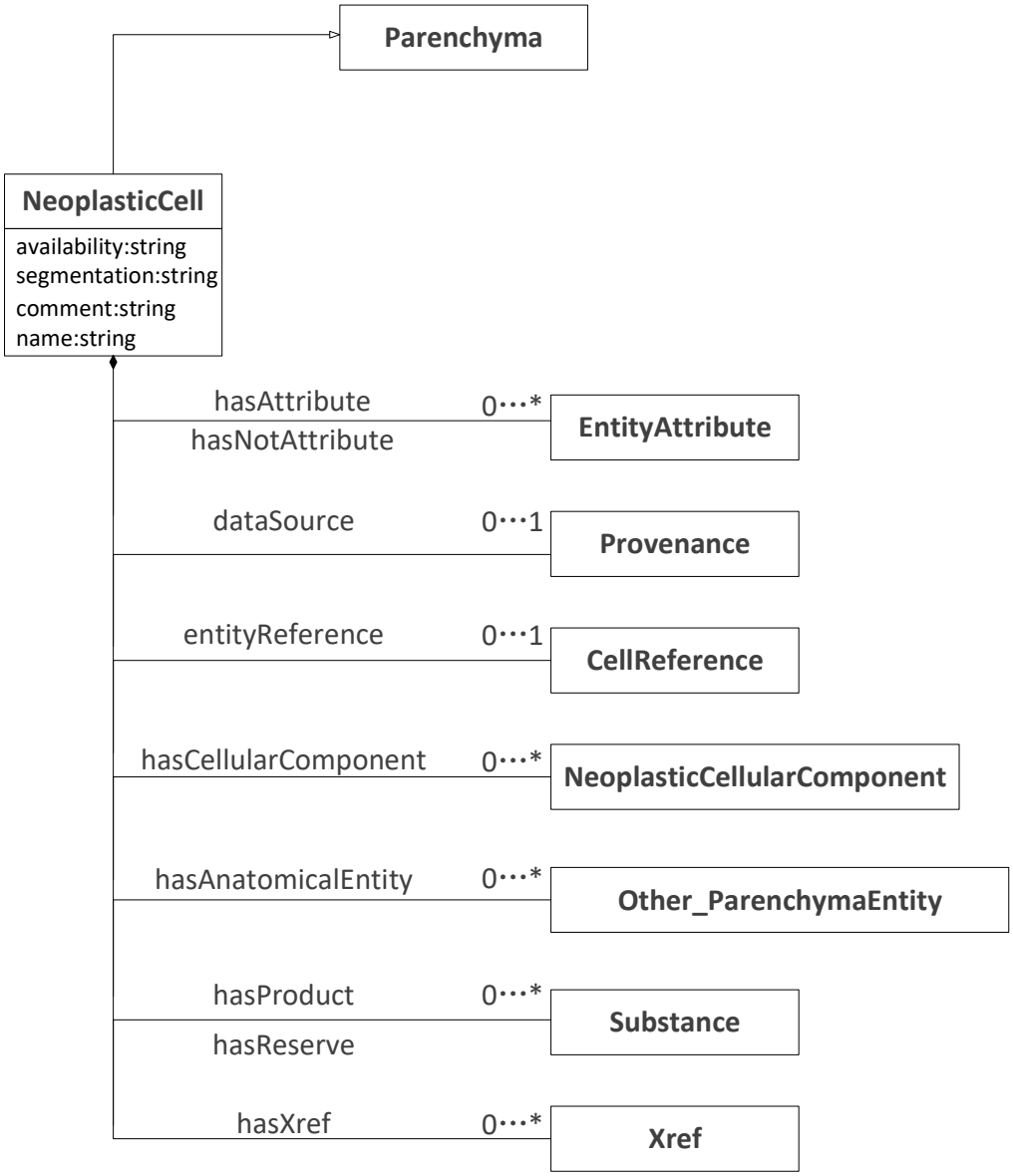

Figure 2-10

（2）NeoplasticCellularComponent

An instance of **NeoplasticCellularComponent** refers to one of the cellular components of a tumor cell. Its definition is shown in Figure 2-11.

hasCellularComponent: (0 or more object:NeoplasticCellularComponent) Some cellular components are even formed by several parts. For example, nucleus has chromatin and nucleolus etc. *hasCellularComponent* is used to define these

components.

hasAnatomicalEntity: (0 or 1 object:Other_ParenchymaEntity) Used to define morphologic structures in the cellular component.

entityReference: (0 or 1 object: CellularComponentReference) The entity reference stores the base definition of the cellular component.

hasXref: (0 or more object:Xref) Values of this property define external cross-references from this entity to entities in external databases (e.g. controlled vocabulary). PathoML recommends using "Neoplastic Cellular Component" as well as its sub-classes from Tumor Pathology Ontology.

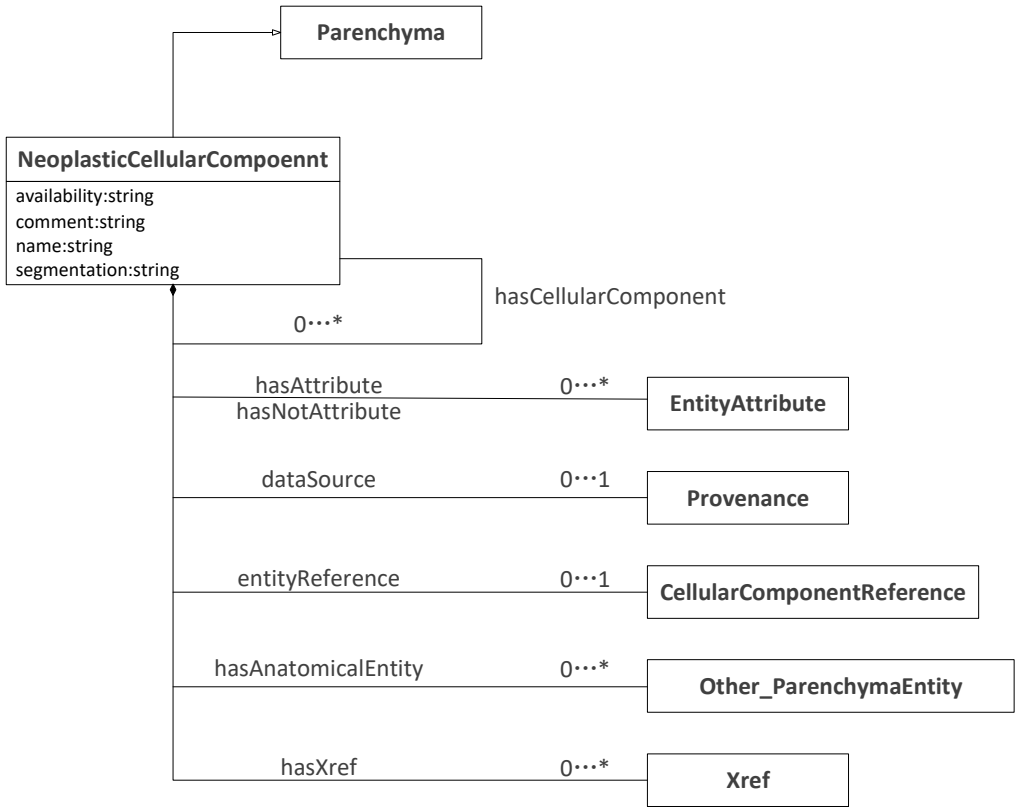

Figure 2-11

（3）Other_ParenchymaEntity

Other_ParenchymaEntity is used to describe immaterial morphologic structures within tumor cells like scattered small cystic spaces in the cytoplasm. Its definition is shown in Figure 2-12.

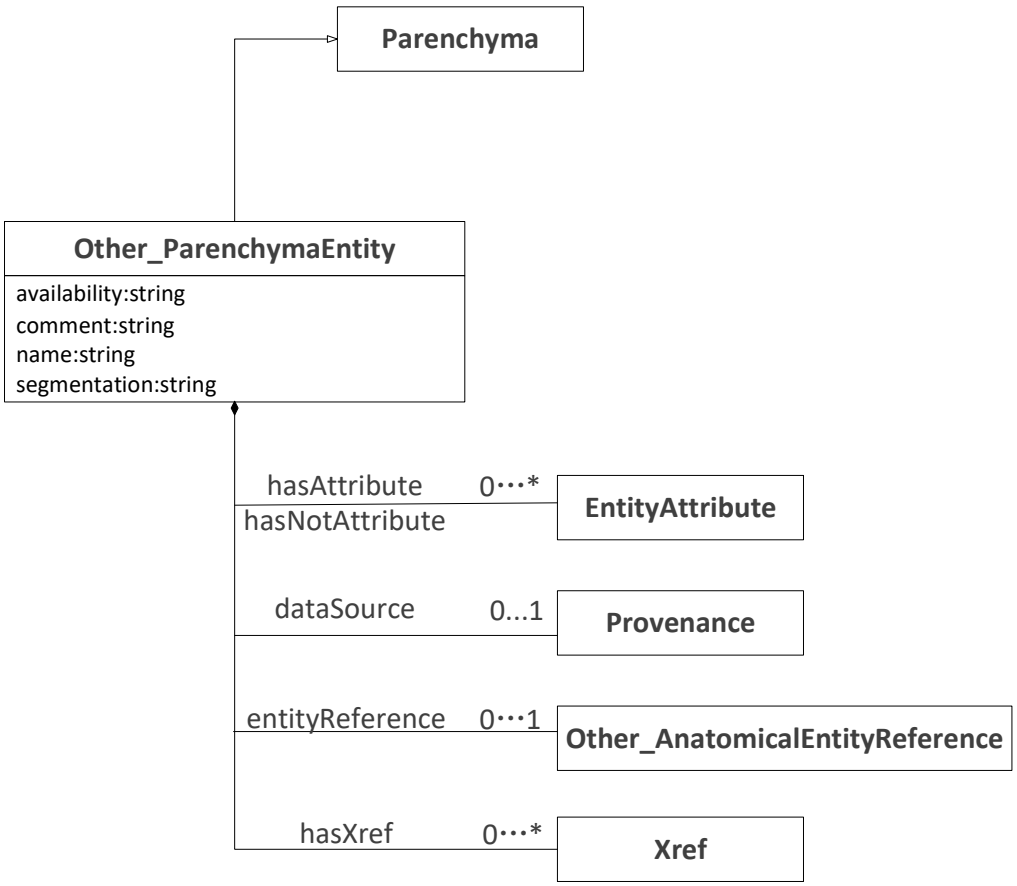

Figure 2-12

entityReference: (0 or 1 object:Other_AnatomicalEntityReference) The entity reference stores the base definition of the morphologic structure.

2.5.3 Stroma

Tumor stroma is primarily composed of the basement membrane, fibroblasts, extracellular matrix, immune cells, and blood vessels. Cells and their cellular

components, tissues and morphologic changes existed in the stroma should be described using **Stroma**. **Stroma** has three child classes which are **Other_StromaEntity**, **StromaCell**, **StromaCellularComponent**. Its definition is shown in Figure 2-13.

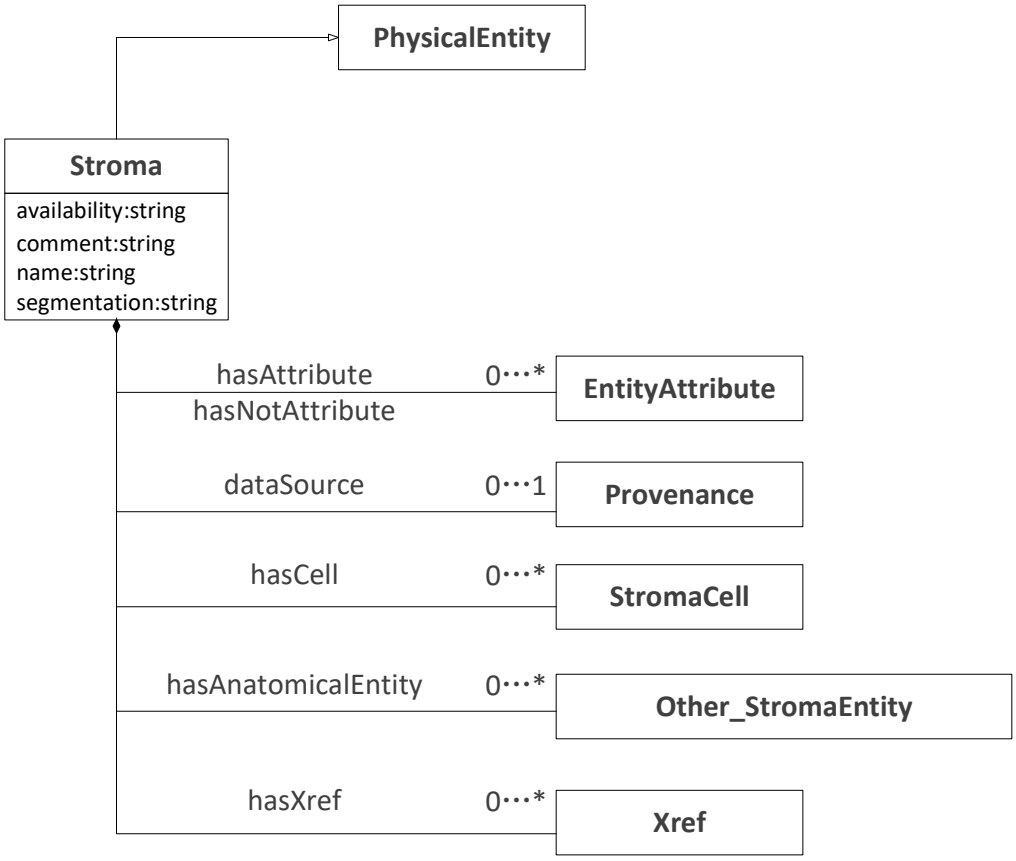

Figure 2-13

hasCell: (0 or more object:StromaCell) Used to define cells in the stroma such as immune cells, fibroblasts, endothelial cells etc.

hasAnatomicalEntity: (0 or more object:Other_StromaEntity) Used to define tissues (e.g. blood vessels) or morphologic structures in the stroma.

hasXref: (0 or more object:Xref) Values of this property define external cross-references from this entity to entities in external databases (e.g. controlled vocabulary). PathoML recommends using "Tumor Stroma" as well as its sub-classes from Tumor Pathology Ontology.

（1） Other_StromaEntity

Within the stroma, other than cells and their cellular components, there are tissues which are induced by the tumor, such as blood vessels and fibers; moreover, there are immaterial morphological structures which are phenotypically meaningful, such as cavity or duct. These entities should be represented using **Other_StromaEntity**, Its definition is shown in Figure 2-14.

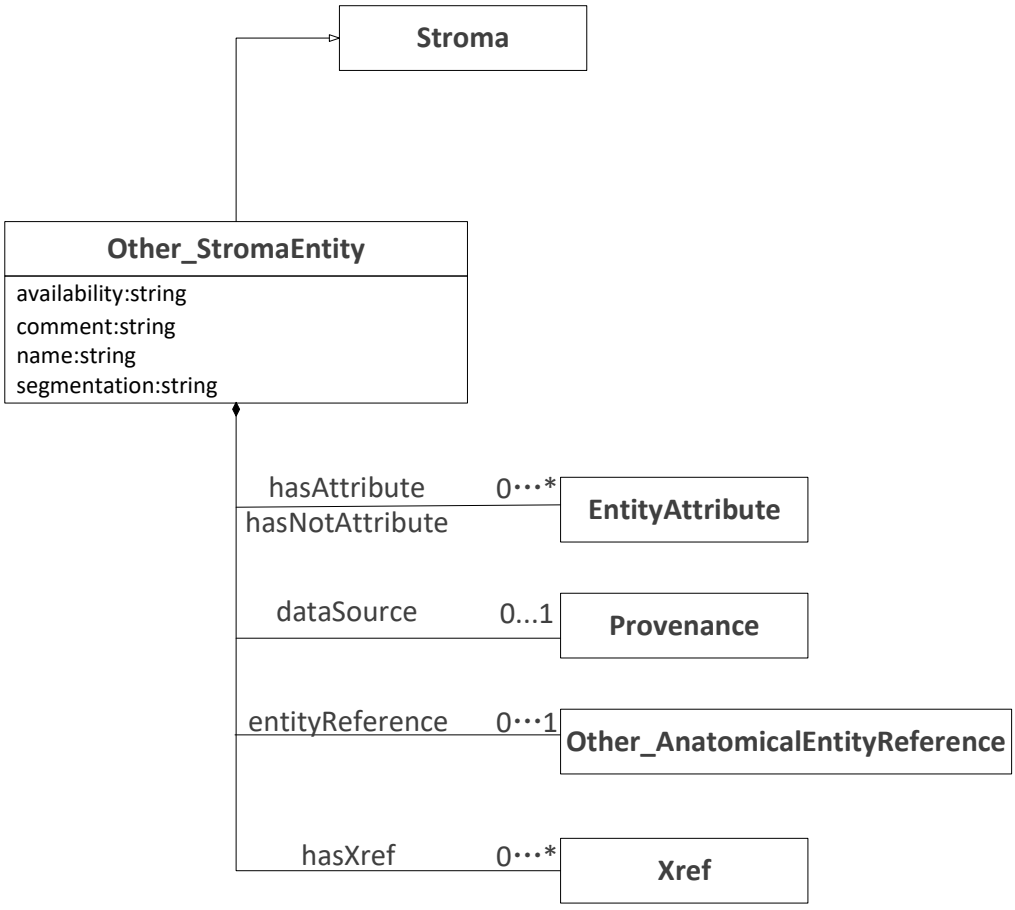

Figure 2-14

（2） StromaCell

Cells in the stroma should be represented using **StromaCell** such as immune cells, fibroblasts, endothelial cells etc. Its definition is shown in Figure 2-15.

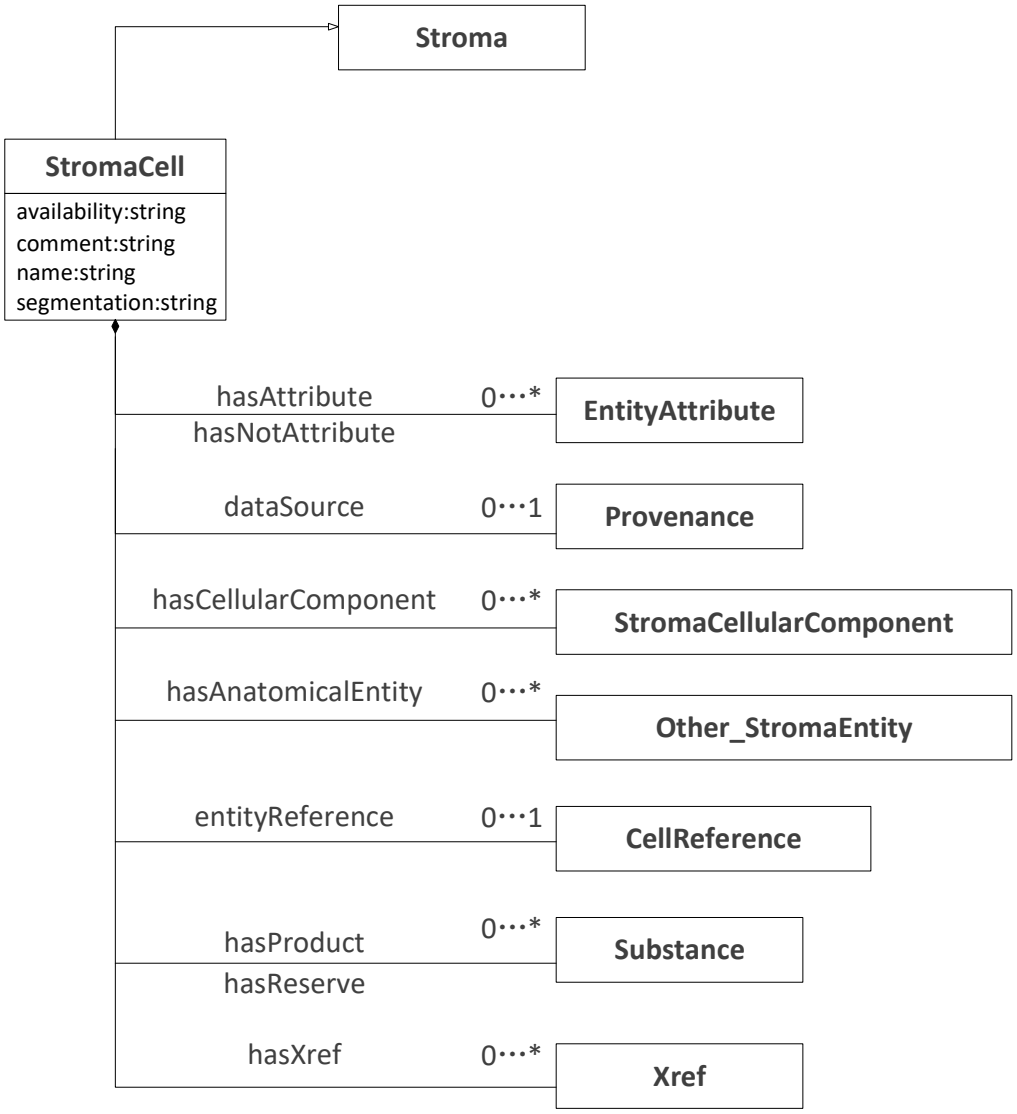

Figure 2-15

hasCellularComponent: (0 or more object:StromaCellularComponent) Used to define cellular components of a stroma cell.

hasAnatomicalEntity: (0 or more object:Other_StromaEntity) Used to define morphologic structures in a stroma cell.

hasProduct: (0 or more object:Substance) Used to define the product of the cell.

hasReserve: (0 or more object:Substance) Used to define the reserve of the cell.

（3）StromaCellularComponent

Cellular components of the stroma cells should be represented using **StromaCellularComponent**. Its definition is shown in Figure 2-16.

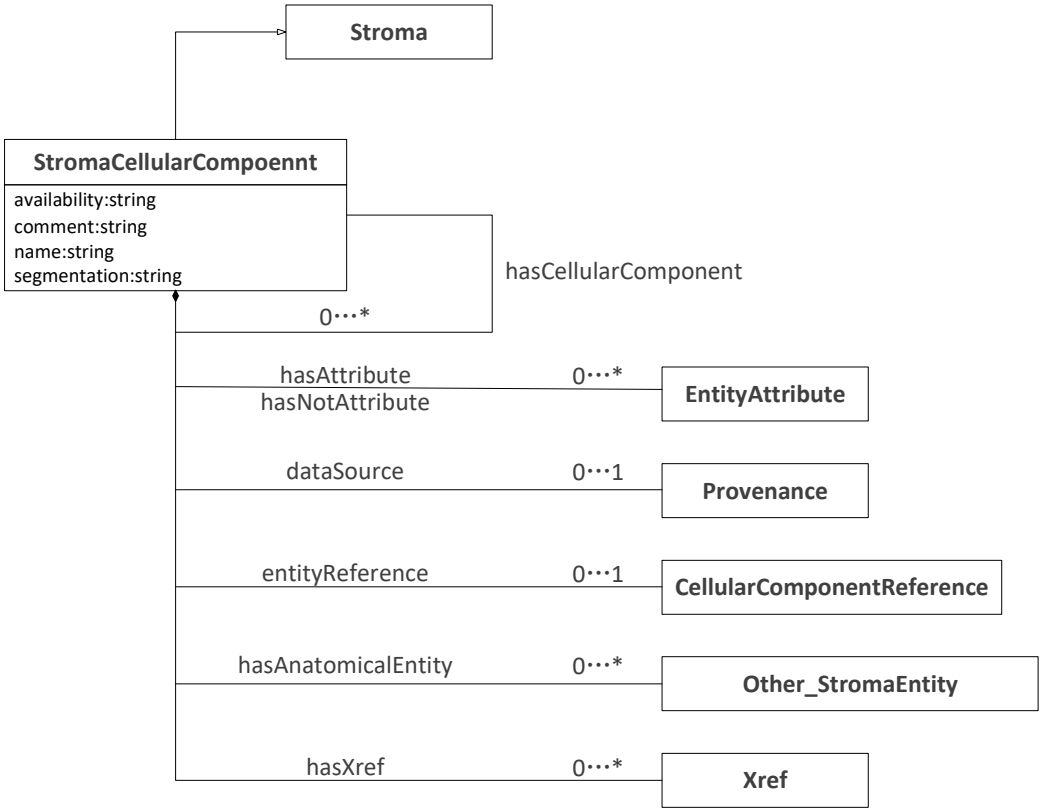

Figure 2-16

hasCellularComponent: (0 or more object:StromaCellularComponent) Some cellular components are even formed by several parts. For example, nucleus has chromatin and nucleolus etc. *hasCellularComponent* is used to define these components.

hasAnatomicalEntity: (0 or more object:Other_StromaEntity) Used to define morphologic structures in the cellular component.

2.5.4 NormalEntity

NormalEntity is used to describe normal cells, their cellular components, normal

tissues, parts of normal tissues, parts of organs, acellular structures and immaterial morphological structures (e.g. cavity or duct) in the normal regions other than tumors.

（1） NormalCell

Its definition is shown in Figure 2-17.

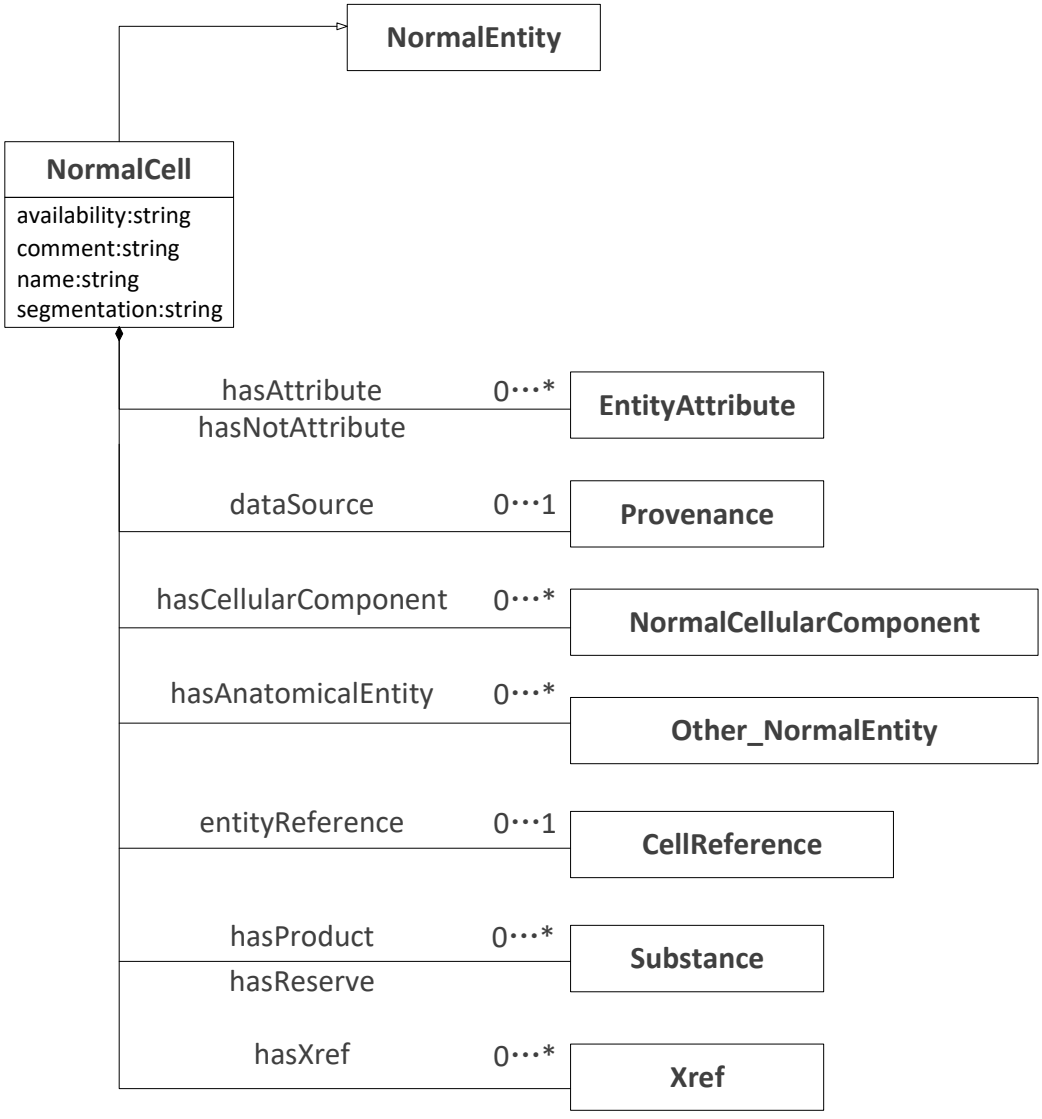

Figure 2-17

hasCellularComponent: (0 or more object:NormalCellularComponent) Used to define cellular components of a normal cell.

hasAnatomicalEntity: (0 or more object:Other_NormalEntity) Used to define morphologic structures in a normal cell.

hasProduct: (0 or more object:Substance) Used to define the product of the cell.

hasReserve: (0 or more object:Substance) Used to define the reserve of the cell.

hasXref: (0 or more object:Xref) Values of this property define external cross-references from this entity to entities in external databases (e.g. controlled vocabulary). PathoML recommends using "Cell" as well as its sub-classes from Tumor Pathology Ontology.

（2） NormalCellularComponent

Its definition is shown in Figure 2-18.

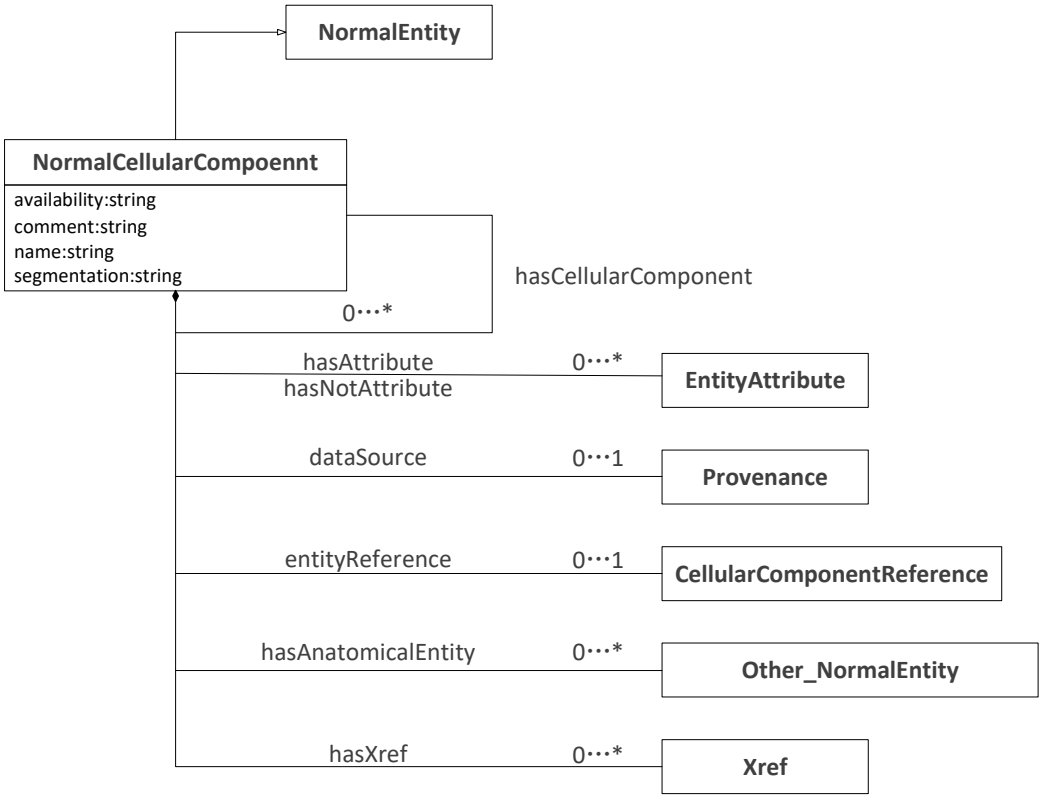

Figure 2-18

hasCellularComponent: (0 or more object:NormalCellularComponent) Some cellular

components are even formed by several parts. For example, nucleus has chromatin and nucleolus etc. *hasCellularComponent* is used to define these components.

hasAnatomicalEntity: (0 or more object:Other_NormalEntity) Used to define morphologic structure in the cellular component.

hasXref: (0 or more object:Xref) Values of this property define external cross-references from this entity to entities in external databases (e.g. controlled vocabulary). PathoML recommends using "cellular_component" as well as its sub-classes from Tumor Pathology Ontology.

（3）Other_NormalEntity

Normal tissues, parts of tissues, parts of organs, acellular structures and immaterial morphological structures (e.g. cavity or duct) in the normal regions are represented using **Other_NormalEntity**. Its definition is shown in Figure 2-19.

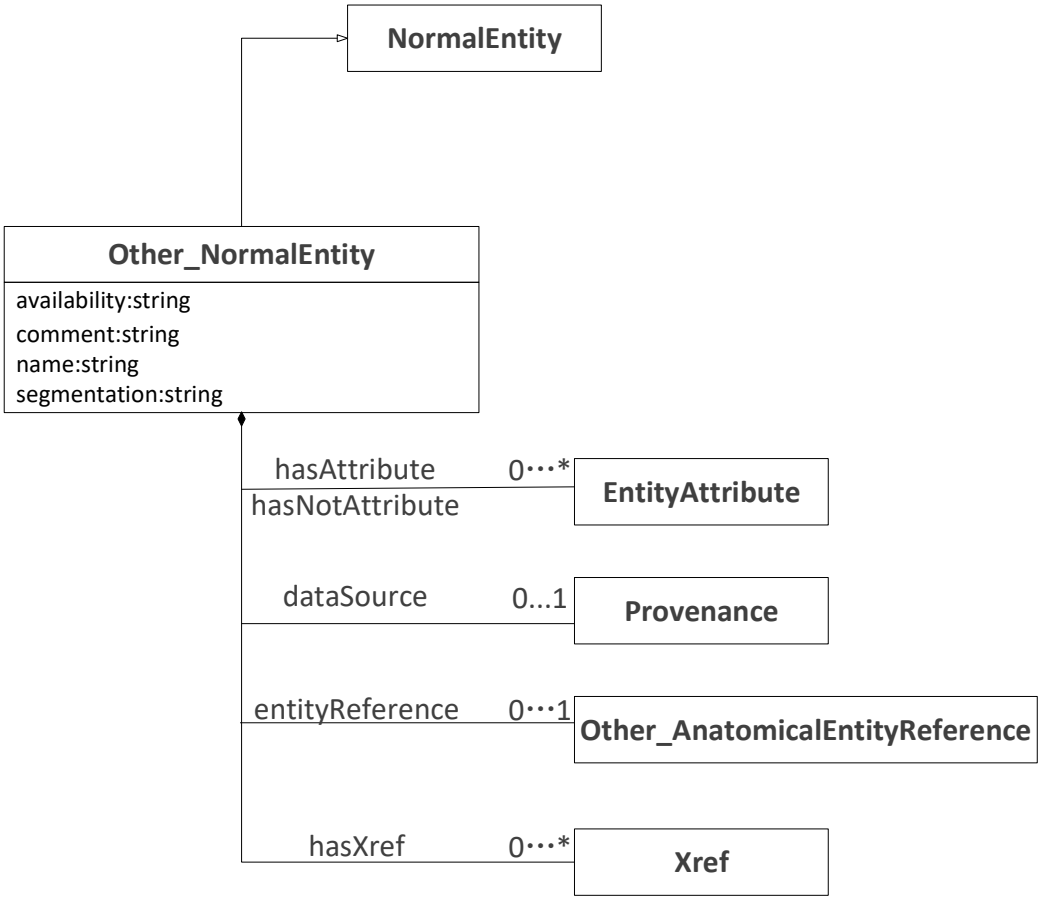

Figure 2-19

hasCell: (0 or more object:NormalCell) Used to define cells of a tissue in a normal region.

hasAnatomicalEntity: (0 or more object:Other_NormalEntity) Used to define tissues and morphologic structures in a normal region.

hasXref: (0 or more object:Xref) Values of this property define external cross-references from this entity to entities in external databases (e.g. controlled vocabulary). PathoML recommends using "Other_NormalEntity" as well as its sub-classes from Tumor Pathology Ontology.

2.5.5 Substance

Substance is used to describe any chemical substance including a cell's product

or reserve, antigen and antibody. Its definition is shown in Figure 2-20.

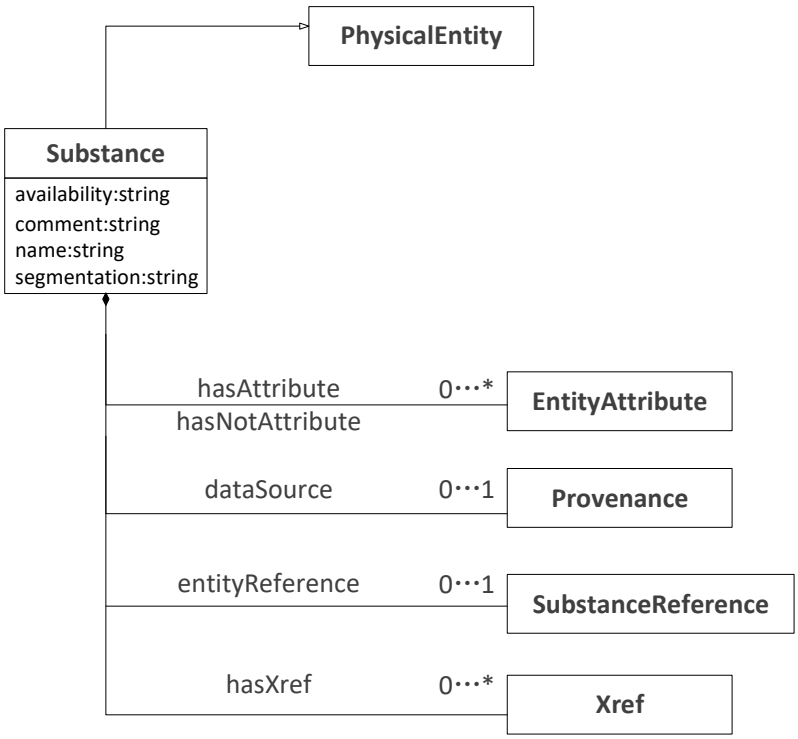

Figure 2-20

hasXref: (0 or more object:Xref) Values of this property define external cross-references from this entity to entities in external databases (e.g. controlled vocabulary). PathoML recommends using "Biological macromolecule" and "Portion of body substance" as well as its sub-classes from Tumor Pathology Ontology and terms from UniProt.

2.5.5.1 Antigen

An individual of **Antigen** represents the antigen expressed on the specific location of a cell in an IHC slide. The structure of **Antigen** is shown in Figure 2-21.

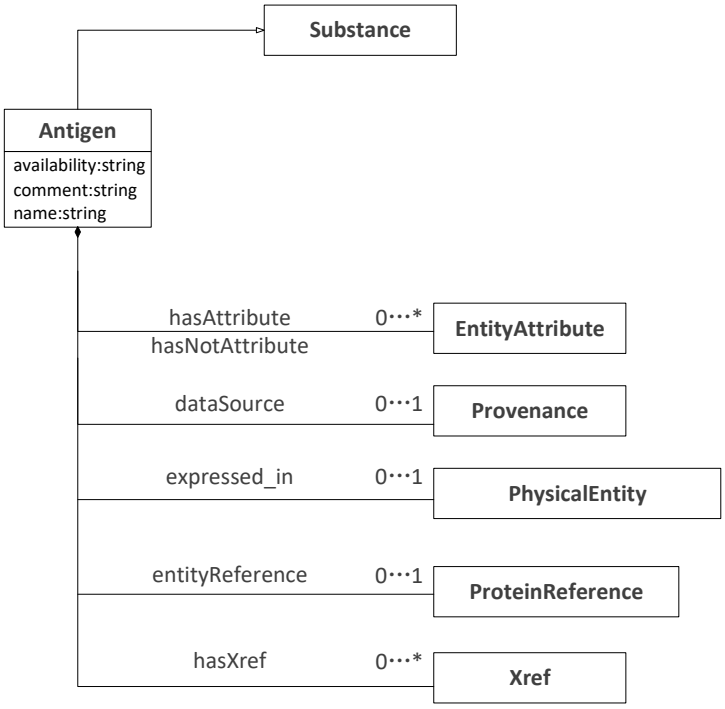

Figure 2-21

hasAttribute: (0 or more object:EntityAttribute) This property is used to the extent and intensity of an antigen's expression.

hasXref: (0 or more object:Xref) Values of this property define external cross-references from this entity to entities in external databases (e.g. controlled vocabulary). PathoML recommends using UniProt.

2.5.5.2 Antibody

An individual of **Antibody** represents the antibody used to stain an IHC slide.

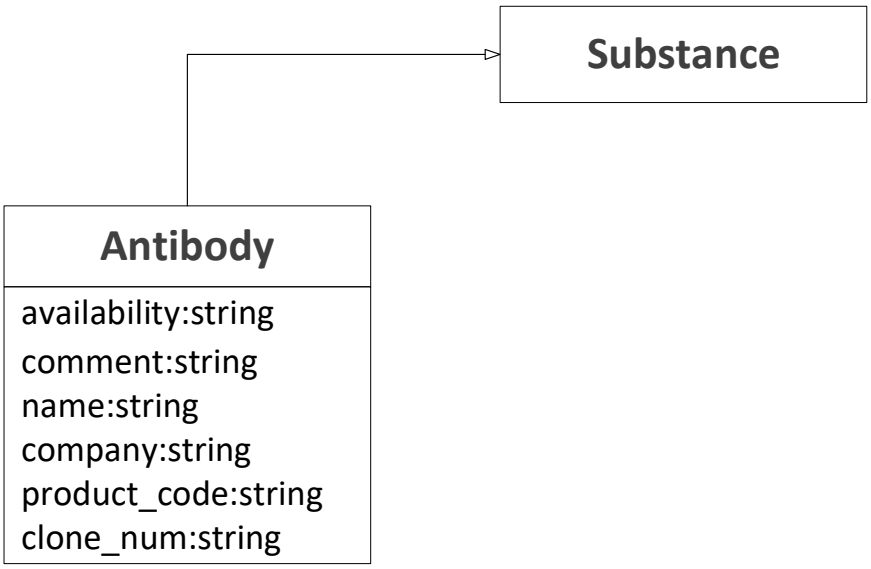

Figure 2-22

company: (xsd:string) The brand of the antibody.

product_code: (xsd:string) The product code of the antibody, if available.

clone_num: (xsd:string) The clone number of this antibody.

2.6 Phenotype Subclass

Phenotype has five child classes including **Architectural_Pattern**, **Cellular_Appearances** and **Product_or_Reserve**, **Immunophenotype** and **Quantitative_Metric**. The structure of Phenotype is shown in Figure 2-23.

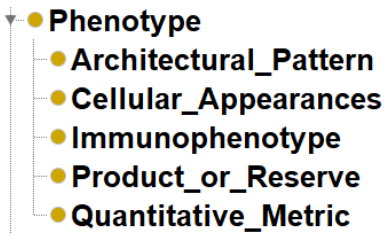

Figure 2-23

2.6.1 Cellular_Appearances

Cellular_Appearances describes a physical change of a cell or a subcellular structure which are usually observed microscopically at high power. Its definition is shown in Figure 2-24.

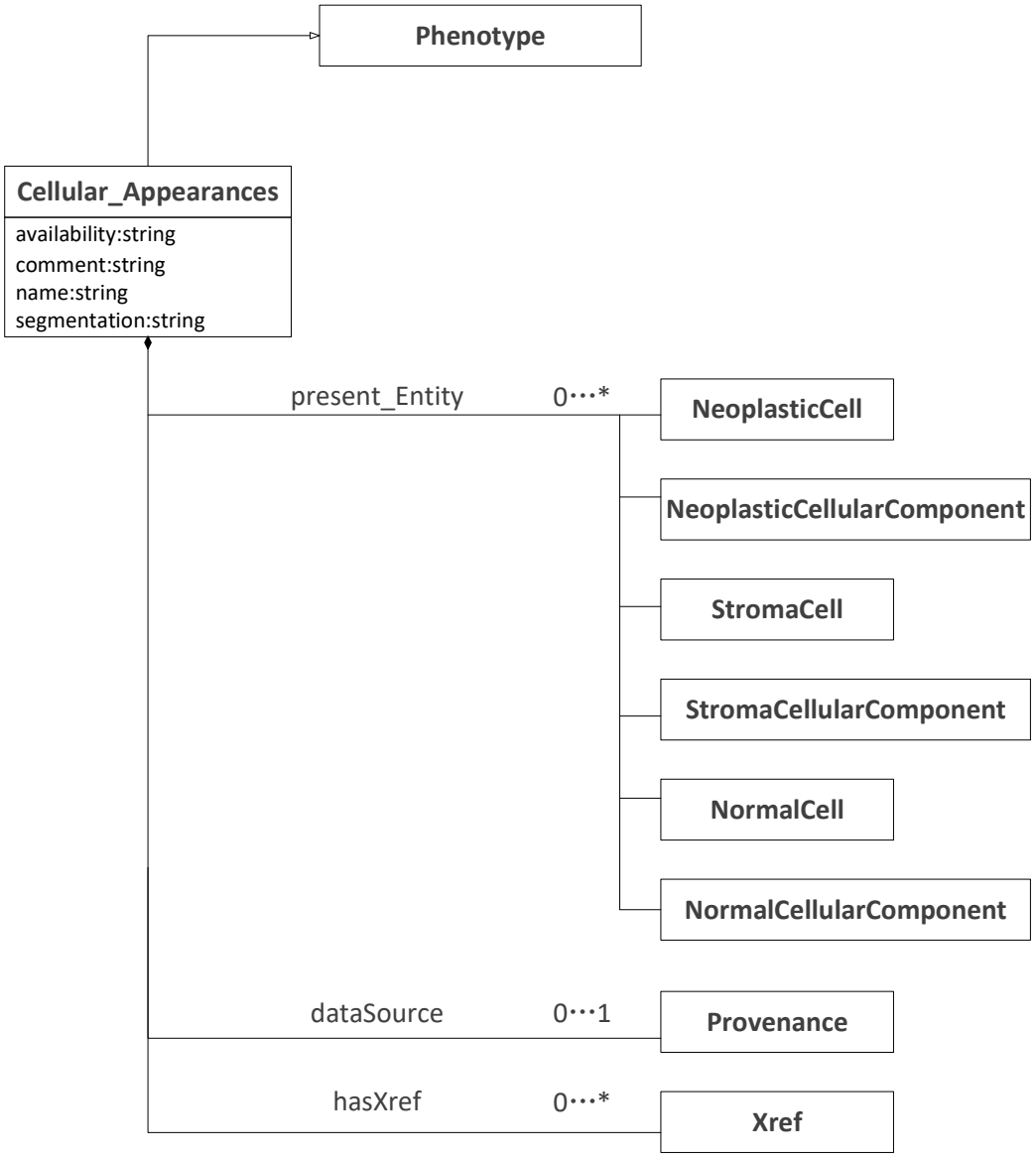

Figure 2-24

present_Entity: (0 or more object:Cell or object:CellularComponent) Used to describe the appearance by linking it to the corresponding **PhysicalEntity** instance

which defines the cell or cellular component.

hasXref: (0 or more object:Xref) Values of this property define external cross-references from this entity to entities in external databases (e.g. controlled vocabulary). PathoML recommends using "Cellular phenotype" as well as its sub-classes from Tumor Pathology Ontology.

2.6.2 Architectural_Pattern

Architectural_Pattern describes histologic patterns of cell populations and tumor behaviors (e.g. extension, invasion) which are usually observed microscopically at medium and low power. Its definition is shown in Figure 2-25.

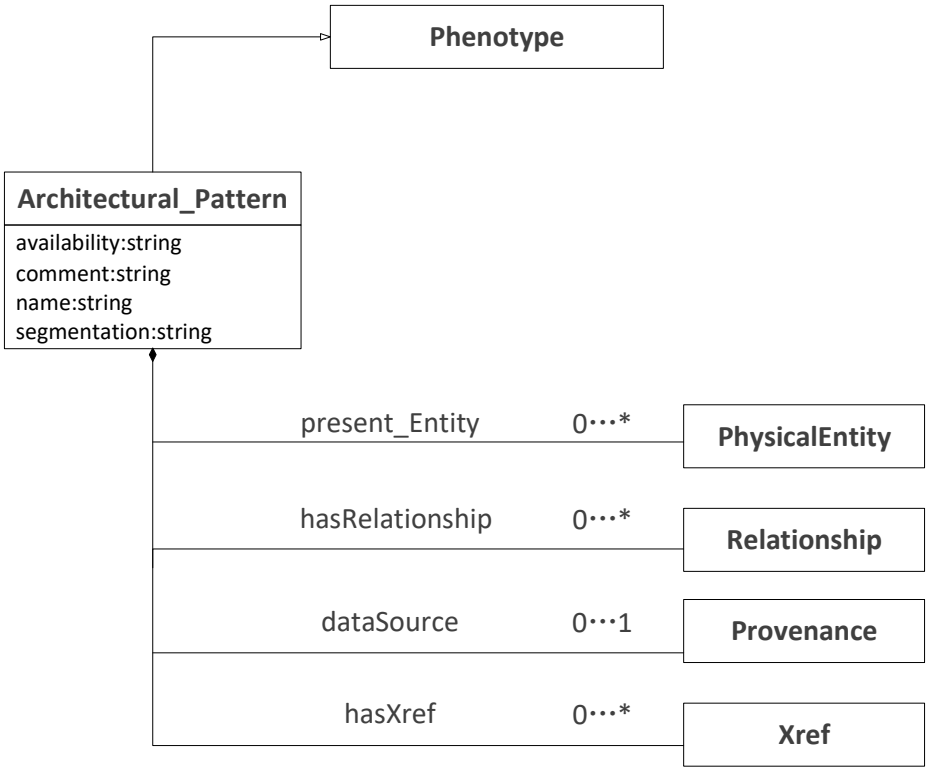

Figure 2-25

present_Entity: (0 or more object:PhysicalEntity) The physical entity that is present in the architectural pattern, e.g., in an acinar pattern, the present entities are erythrocytes, acidophilic fluid, a lumen, tumor cells, a capillary and endothelial cells.

hasRelationship: (0 or more object:Relationship) The relationships between the present entities which are used to describe tumor behaviors such as invading, extending, being limited to etc.

hasXref: (0 or more object:Xref) Values of this property define external cross-references from this entity to entities in external databases (e.g. controlled vocabulary). PathoML recommends using "Morphologic Finding" as well as its sub-classes from Tumor Pathology Ontology.

2.6.3 Product_or_Reserve

Some phenotypes observed under the slide refer to the products or reserves of tumor cells such as neutral fat, glycogen, pigment or crystal etc. These phenotypes are described using **Product_or_Reserve**. Its definition is shown in Figure 2-26.

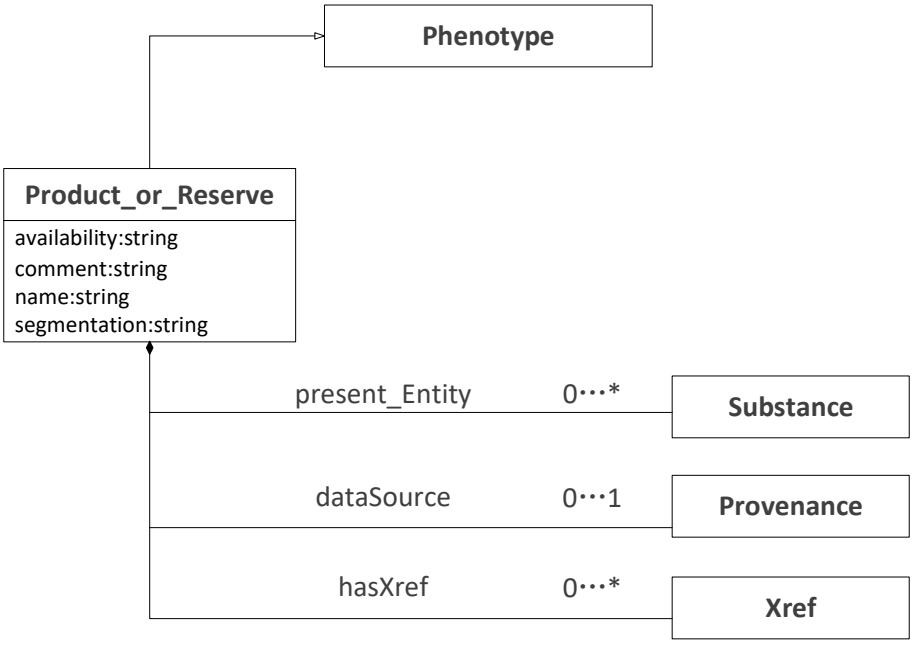

Figure 2-26

present_Entity: (0 or more object:Substance) The substance that is present in the slide, e.g., in the area of calcification, the present substance is calcium.

2.6.4 Immunophenotype

Immunophenotypes refer to the expression patterns of antigens in the context of cells and tissues. The type of antigen, the extent and intensity of expression and the expression location are the key to characterize an immunophenotype. Immunophenotypes are described using **Immunophenotype**. Its definition is shown in Figure 2-27.

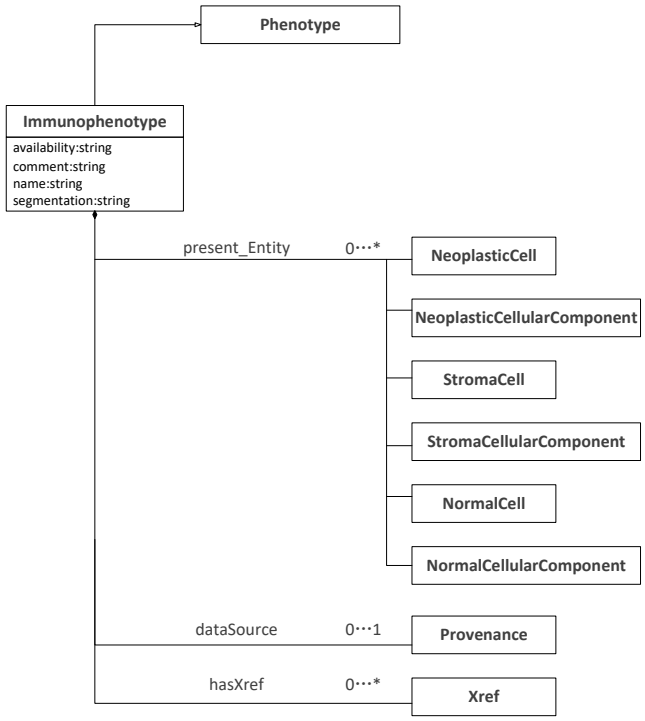

Figure 2-27

present_Entity: (0 or more object:Cell or object:CellularComponent) The cell or cellular component stained by the antibody either in stromal area or parenchyma.

hasXref: (0 or more object:Xref) Values of this property define external cross-references from this entity to entities in external databases (e.g. controlled vocabulary). PathoML recommends using " Immunophenotypic Finding " as well as its sub-classes from Tumor Pathology Ontology.

2.6.5 Quantitative_Metric

There exist quantitative measurements of phenotypes such as tumor-stroma ratio for characterizing intra-tumoral stroma and the semi-quantitative scoring for assessing HER2 expression. They have been shown to provide prognostic and predictive value and regarded as diagnostic parameters by many diagnostic criteria. PathoML represents such measurement using **`Quantitative_Metric`**. Its definition is shown in Figure 2-28.

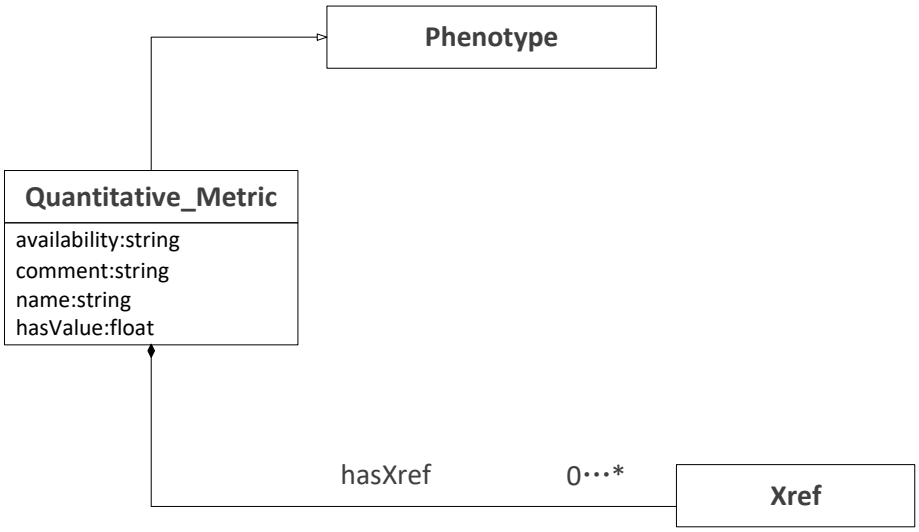

Figure 2-28

`hasValue`: (xsd:float) The value of this measurement.

hasXref: (0 or more object:Xref) Values of this property define external cross-references from this entity to entities in external databases (e.g. controlled vocabulary). PathoML recommends using "Quantitative_Metric" as well as its sub-classes from Tumor Pathology Ontology.

2.7 Diagnosis

After examining a tissue's histopathological phenotypes and immunophenotypes, a pathologist makes pathological diagnoses which evaluate the impact of these factors to

the patient's future outcome. Pathological diagnoses refer to not only the classification of a patient as having a particular disease or not, but also the evaluations of some prognostic parameters such as tumor grade and stage, with immunophenotypes and histopathological phenotypes as the main basis for decision-making. Such information is regarded as the primary guide to treatment and prognosis and contained in pathology reports. An individual of **Diagnosis** class represents a diagnostic decision made by a pathologist. Its definition is shown in Figure 2-29.

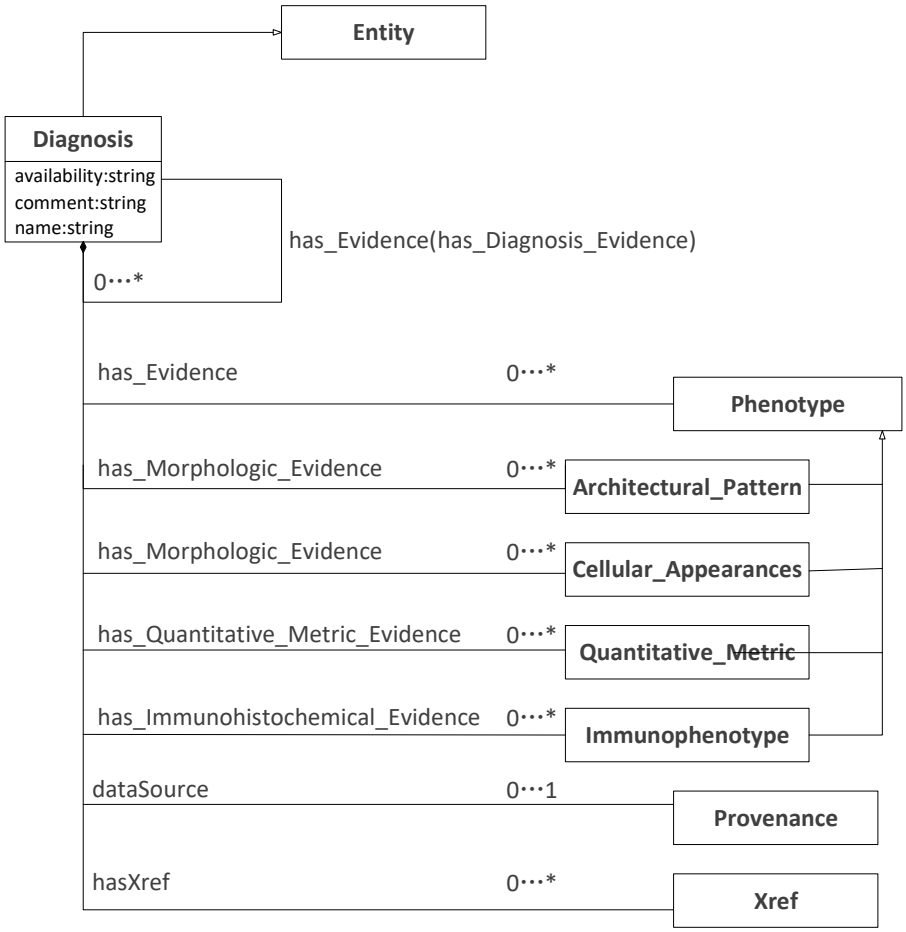

Figure 2-29

has_Evidence: (0 or more object:Diagnosis and object:Phenotype) Used to link an individual of **Diagnosis** to one or several **Phenotype** individuals representing that the presence of these phenotypes are the supporting evidence of the diagnosis; one

phenotype could be an evidence of several diagnoses in PathoM. Additionally, a diagnosis which has been made previously could be regarded as facts to support the up-coming diagnoses. According to the different types of supporting evidence *has_Evidence* links to, including histopathological phenotypes, immunophenotypes , quantitative metrics of phenotypes and diagnoses, *has_Evidence* is further categorized into three sub-properties including *has_Morphologic_Evidence*, *has_Quantitative_Metric_Evidence*, *has_Immunohistochemical_Evidence* and *has_Diagnosis_Evidence*. The use of these sub-properties is similar to *has_Evidence*.

hasXref: (0 or more object:Xref) Values of this property define external cross-references from this entity to entities in external databases (e.g. controlled vocabulary). PathoML recommends using "Pathological Diagnosis" as well as its sub-classes from Tumor Pathology Ontology.

Diagnosis has two child classes including **Differential_Diagnosis**, **Fina_Diagnosis.** The structure of **Diagnosis** is shown in Figure 2-30.

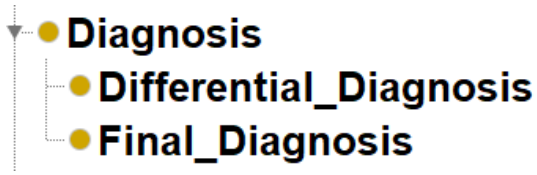

Figure 2-30

2.7.1 Differential_Diagnosis and Final_Diagnosis

In real-world practice, two different diagnoses could be attached to a patient who exhibits histopathological phenotypes fitting into either one. Under this circumstance, pathologists tend to use a differential diagnostic method to reduce candidate conditions. PathoML represents a differential diagnostic process by using **Differential_Diagnosis** and **Final_Diagnosis**. An individual of **Final_Diagnosis** represents a confirming diagnosis while an individual of

Differential_Diagnosis represents an unlikely diagnosis. The structure of **Differential_Diagnosis** is shown in Figure 2-31; the structure of **Final_Diagnosis** is similar.

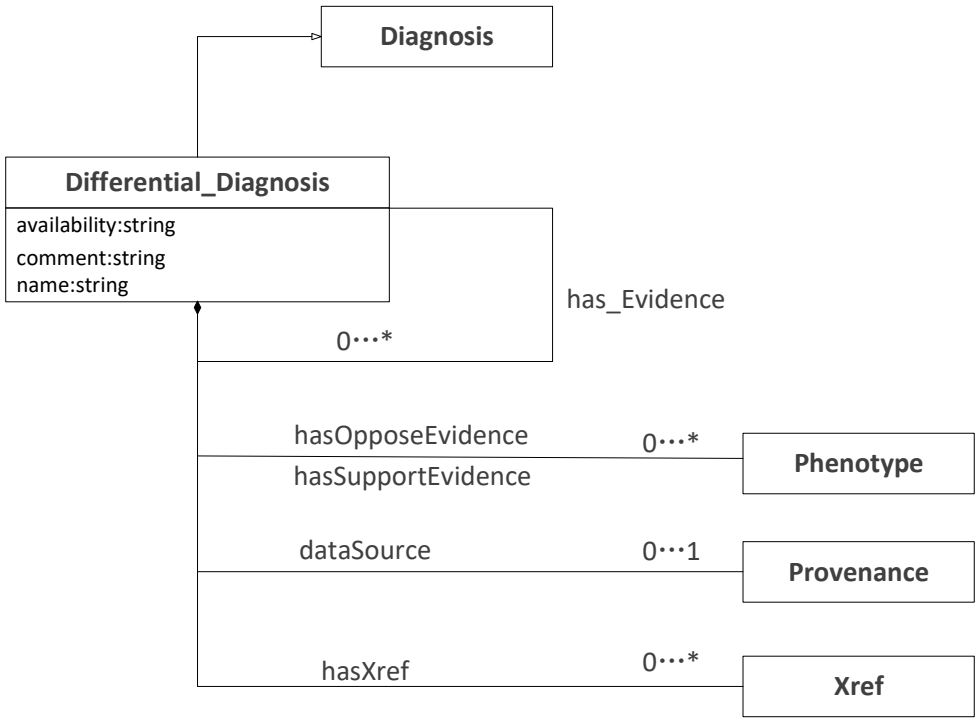

Figure 2-31

hasOpposeEvidence and *hasSupportEvidence*: (0 or more object:Diagnosis and object:Phenotype) *hasOpposeEvidence* and *hasSupportEvidence*: are used to link a diagnosis to the phenotypes which are for and against the diagnosis respectively, representing the reasons why a pathologist determines this possible condition true or false. They are other two sub-properties of *has_Evidence* in addition to *has_Morphologic_Evidence*, *has_Quantitative_Metric_Evidence*, *has_Immunohistochemical_Evidence* and *has_Diagnosis_Evidence*.

hasXref: (0 or more object:Xref) Values of this property define external cross-references from this entity to entities in external databases (e.g. controlled vocabulary). PathoML recommends using " Pathological Diagnosis" as well as its sub-classes from

Tumor Pathology Ontology.

2.8 Utility

Several object properties of **Entity** subclasses accept instances of **Utility** classes as values. **Utility** classes are used to annotate the **Entity** subclasses. Examples include references to external databases, controlled vocabularies, evidence and provenance. The structure of Utility class is shown in Figure 2-32.

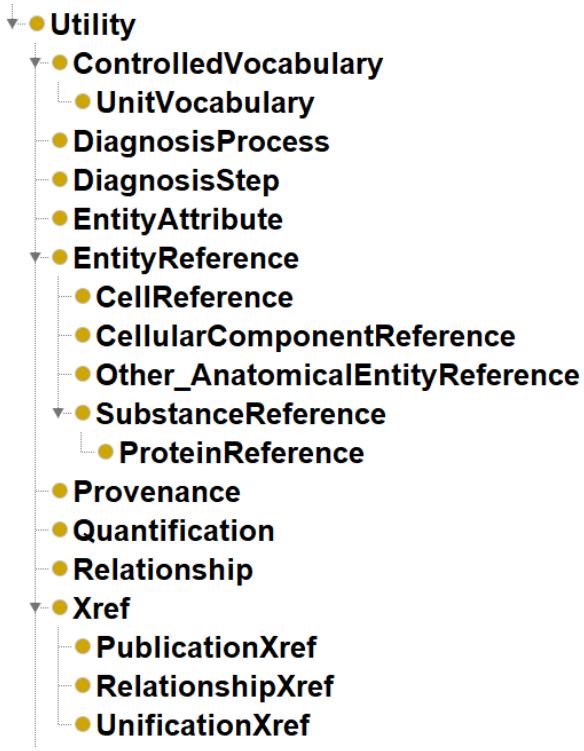

Figure 2-32

2.8.1 ControlledVocabulary

ControlledVocabulary is used to define PathoML's own controlled vocabulary terms. **UnitTypeVocabulary** is the sub-class of it which provides reference to the Units of measurement ontology (UO) to quantitatively describe the attributes of an entity. Its definition is shown in Figure 2-33.

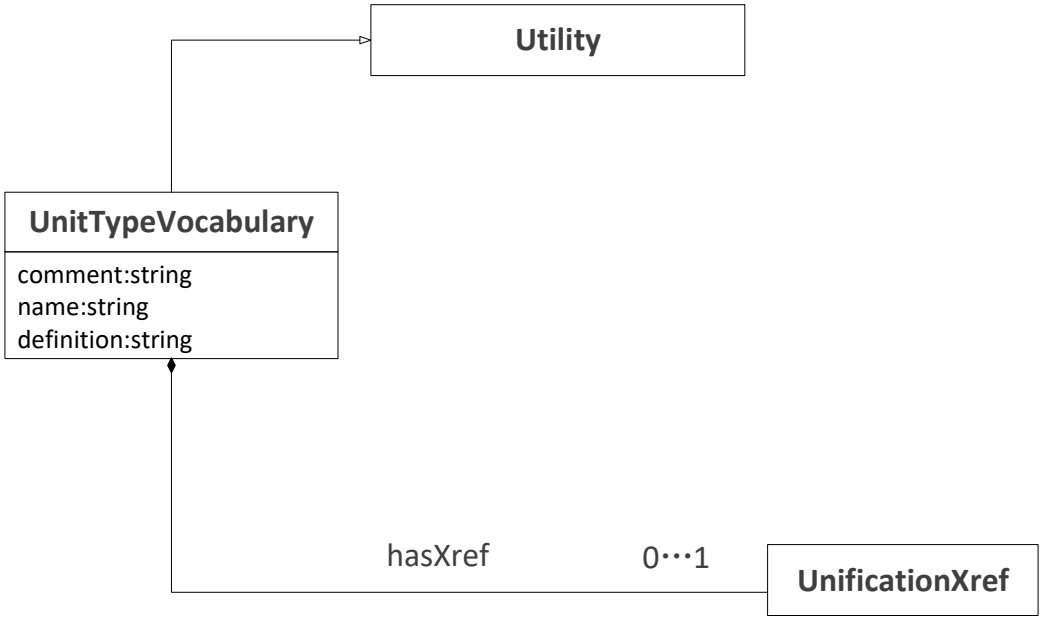

Figure 2-33

hasXref: (0 or 1 object:UnificationXref) the cross-reference to the term in Units of measurement ontology (UO).

`name`: (xsd:string) name of this unit.

`definition`: (xsd:string) definition of this unit.

2.8.2 DiagnosisProcess

PathoML could further represent a diagnostic decision-making process of a pathologist consisting of several diagnoses by using **DiagnosisProcess** and **DiagnosisStep**, which depict the cause-and-effect relationships between diagnoses. **DiagnosisProcess** contains a set of diagnoses involving in a diagnostic procedure. Its definition is shown in Figure 2-34.

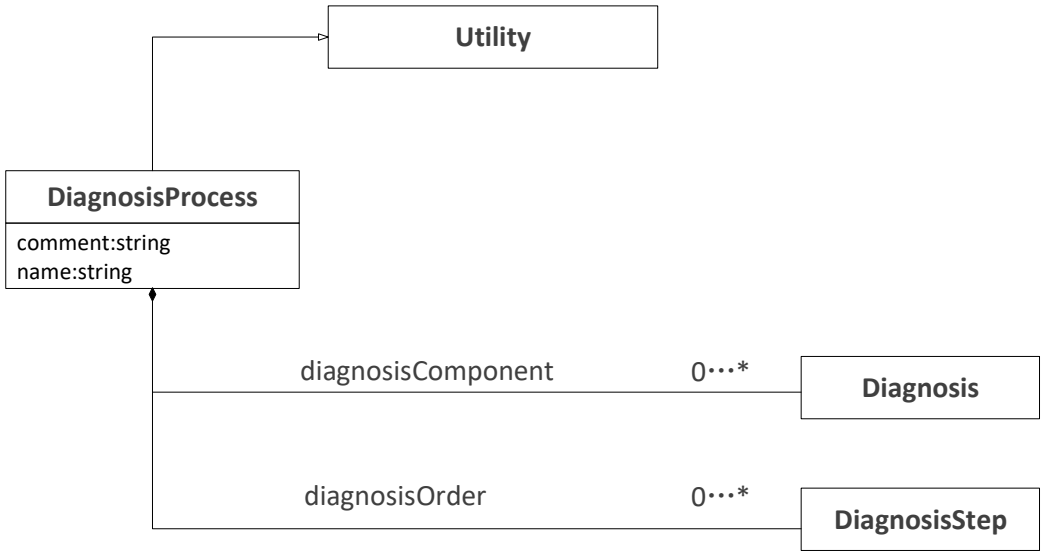

Figure 2-34

diagnosisComponent: (0 or more object:Diagnosis) Used to link an individual of **DiagnosisProcess** to a set of diagnoses involved in the diagnostic procedure.

diagnosisOrder: (0 or more object:DiagnosisStep) Used to links an individual of **DiagnosisProcess** to a **DiagnosisStep** individual(s) which represent the order of diagnoses made by a pathologist toward the final diagnosis.

2.8.3 DiagnosisStep

DiagnosisStep represents the order of diagnoses to arrive at the final diagnosis. Its definition is shown in Figure 2-35.

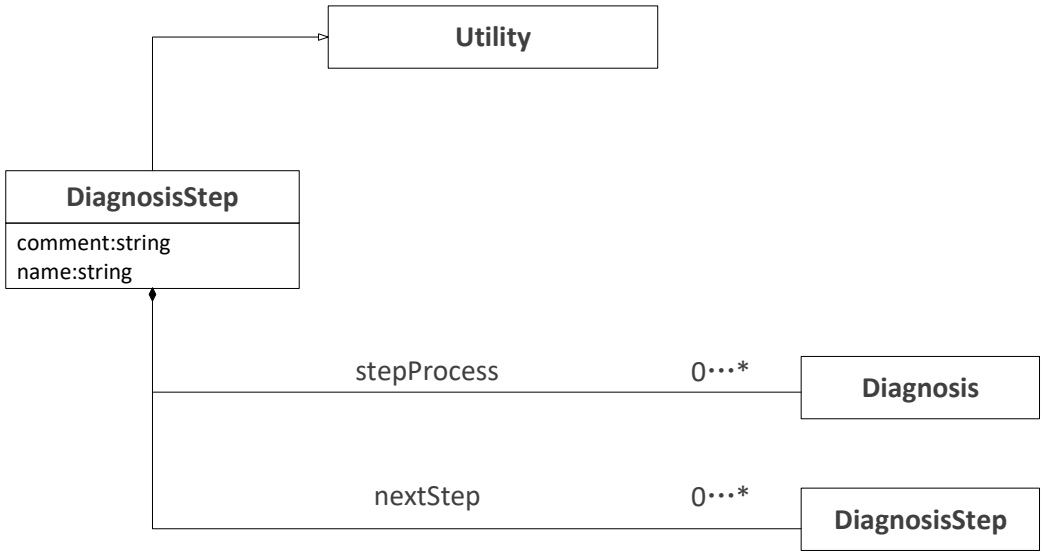

Figure 2-35

stepProcess: (0 or more object:Diagnosis) Used to indicate the diagnosis made by a pathologist currently.

nextStep: (0 or more object:DiagnosisStep) Used to represents the next diagnosis following the pathologist's logic.

2.8.4 EntityAttribute

An attribute of a physical entity that can be changed while the entity still retains its biological identity. For example, when representing a histopathological phenotype, **EntityAttribute** could be used to describe the morphologic features of an entity; when representing an immunophenotype, **EntityAttribute** could be used to describe the extent and intensity of a protein expression. Entity attributes could also be generic across physical entities that are in a generic grouping. This allows generic attributes to be defined on a generic physical entity in **EntityReference**. Its definition is shown in Figure 2-36.

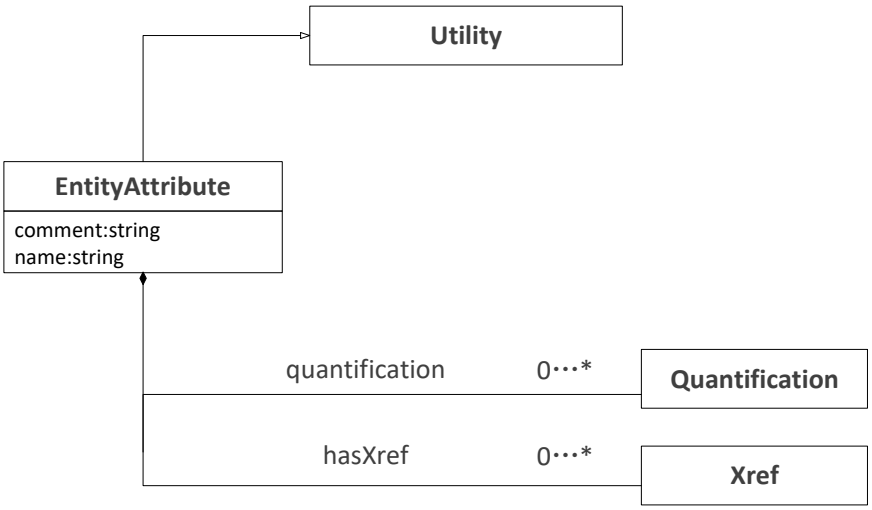

Figure 2-36

quantification: (0 or more object:Quantification) Used to quantitatively describe physical entities' attributes.

hasXref: (0 or more object:Xref) Used to clarify the attributes by providing their semantic cross-references to the controlled vocabulary such as The Phenotype And Trait Ontology ("quality" as well as their sub-classes from Tumor Pathology Ontology).

2.8.5 EntityReference

An entity reference is a grouping of several physical entities with multiple states that are often named and treated as a single entity by pathologists. **EntityReference** instances store the information common to a set of entities in various states. For example, the morphology of a tumor cell of clear cell renal cell carcinoma (ccRCC) undergoes changes as it evolves including nuclear changes, cytoplasmic changes; any ccRCC tumor cell in one of the states is represented as an individual of **NeoplasticCell** while a generic ccRCC tumor cell is represented as an individual of **CellReference** (child of **EntityReference**). The **EntityReference** is important because it explicitly links multiple physical entities representing different states of a

generic entity, which would otherwise be difficult to recognize as related. The definition of EntityRefrence is shown in Figure 2-37.

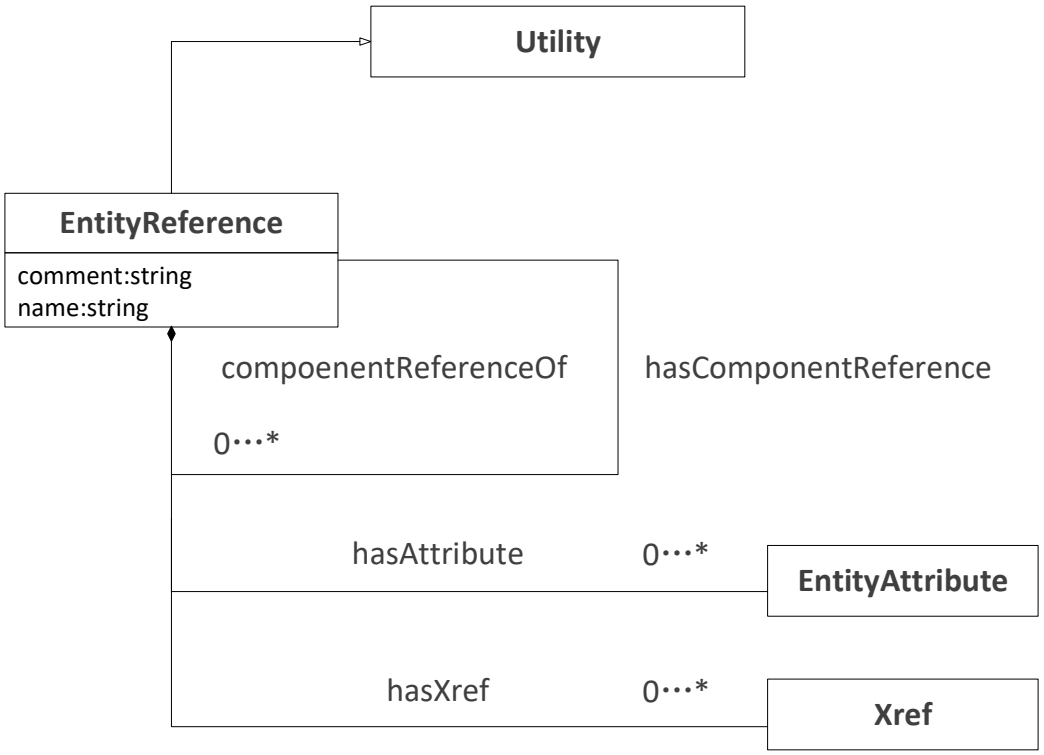

Figure 2-37

hasAttribute: (0 or more object:EntityAttribute) This property is used to define generic attributes for a generic physical entity. Other attributes which distinguish the specific physical entities representing different states of the generic entity are defined in **PhysicalEntity**.

hasComponentReference: (0 or more object:EntityReference) Whereas *hasComponent* is used to define components of a physical entity in a specific state, *hasComponentReference* is used to define components of the generic entity. It has 3 sub-properties which are *hasAnatomicalEntityReference*, *hasCellularComponentReference* and *hasCellReference*. For example, in order to represent a tumor cell with eosinophilic cytoplasm, a tumor cell is represented as a **NeoplasticCell**, and the cytoplasm is represented as a

NeoplasticCellularComponent; *hasCellularComponent* is used to link the tumor cell to the cytoplasm. While *hasCellularComponentReference* is used to link the entity reference of the tumor cell to the reference of the cytoplasm since cytoplasm, in any morphologic state, is the component of a cell.

compoenentReferenceOf: (0 or more object:EntityReference) Inverse property of *hasComponentReference*. Similar to *hasComponent* and *componentOf*, either *hasComponentReference* or *compoenentReferenceOf* is specified, the other could be inferred.

hasXref: (0 or more object:Xref) Values of this property define external cross-references from this entity to entities in external databases (e.g. controlled vocabulary). PathoML recommends using "Anatomical entity" from Tumor Pathology Ontology.

（1）CellReference

Used to store shared information about a set of related cells differing in states. Its definition is shown in Figure 2-38.

hasXref: (0 or more object:Xref) Values of this property define external cross-references from this entity to entities in external databases (e.g. controlled vocabulary). PathoML recommends using "Cell" and "Neoplastic Cell" as well as their sub-classes from Tumor Pathology Ontology.

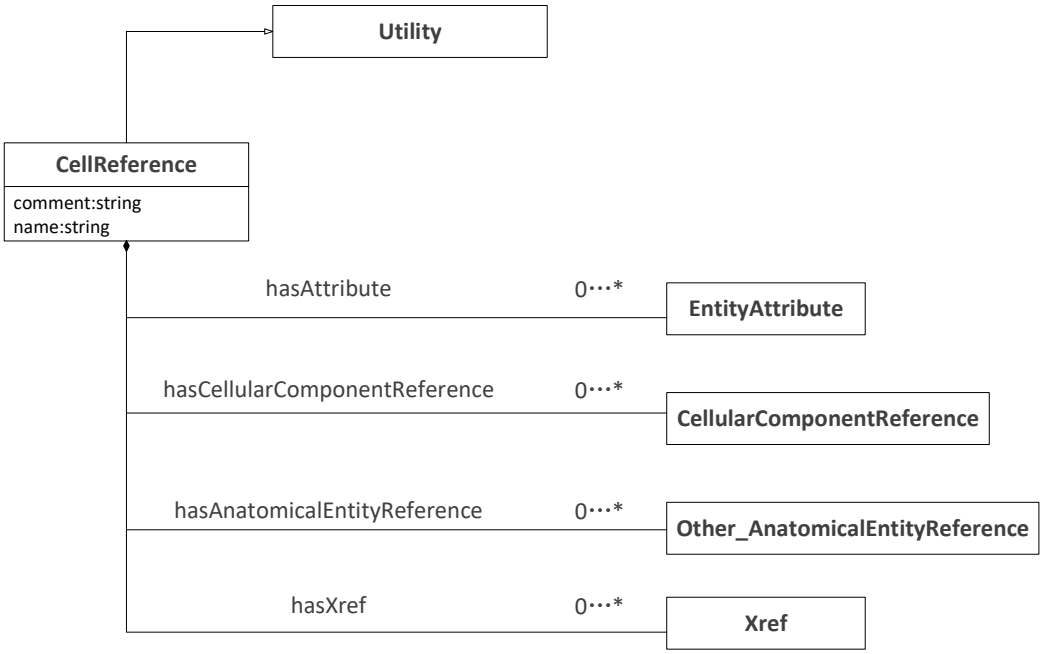

Figure 2-38

（2） CellularComponentReference

Used to store shared information about a set of related cellular components differing in states. Its definition is shown in Figure 2-39.

hasXref: (0 or more object:Xref) Values of this property define external cross-references from this entity to entities in external databases (e.g. controlled vocabulary). PathoML recommends using "cellular_component"and "Neoplastic Cellular Component" as well as their sub-classes from Tumor Pathology Ontology.

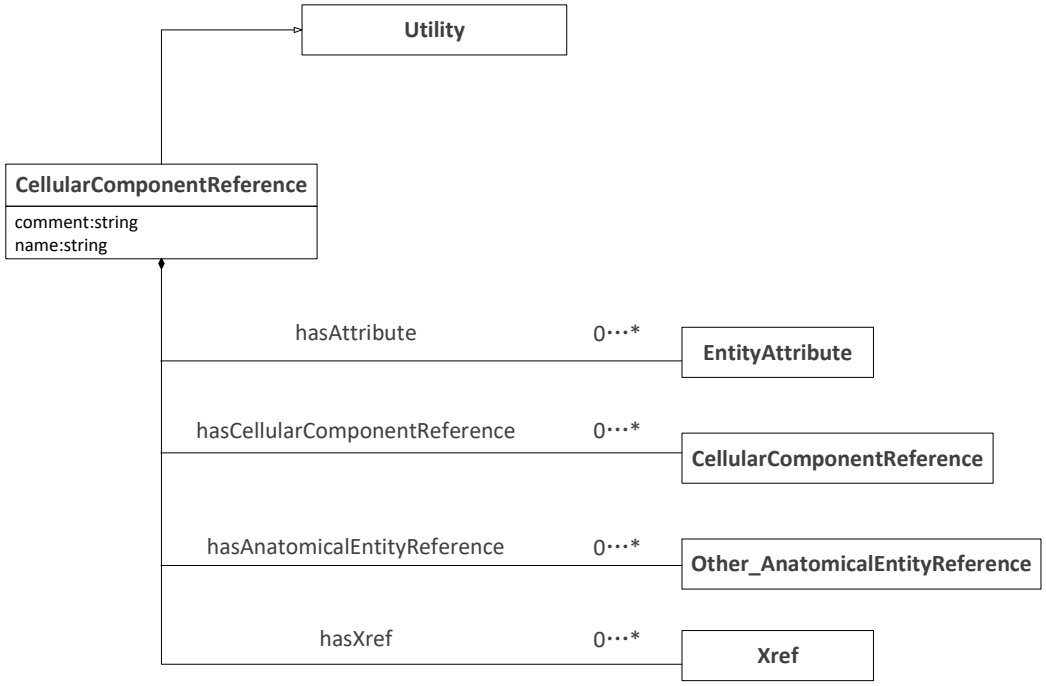

Figure 2-39

（3）Other_AnatomicalEntityReference

Used to store shared information about a set of related tissues, parts of organs or immaterial morphologic structures differing in states. Its definition is shown in Figure 2-40.

hasXref: (0 or more object:Xref) Values of this property define external cross-references from this entity to entities in external databases (e.g. controlled vocabulary). PathoML recommends using "Anatomical space", "Cardinal organ part", "Cardinal tissue part", "Tumor Stroma" and "Tumor Parenchyma" as well as their sub-classes from Tumor Pathology Ontology.

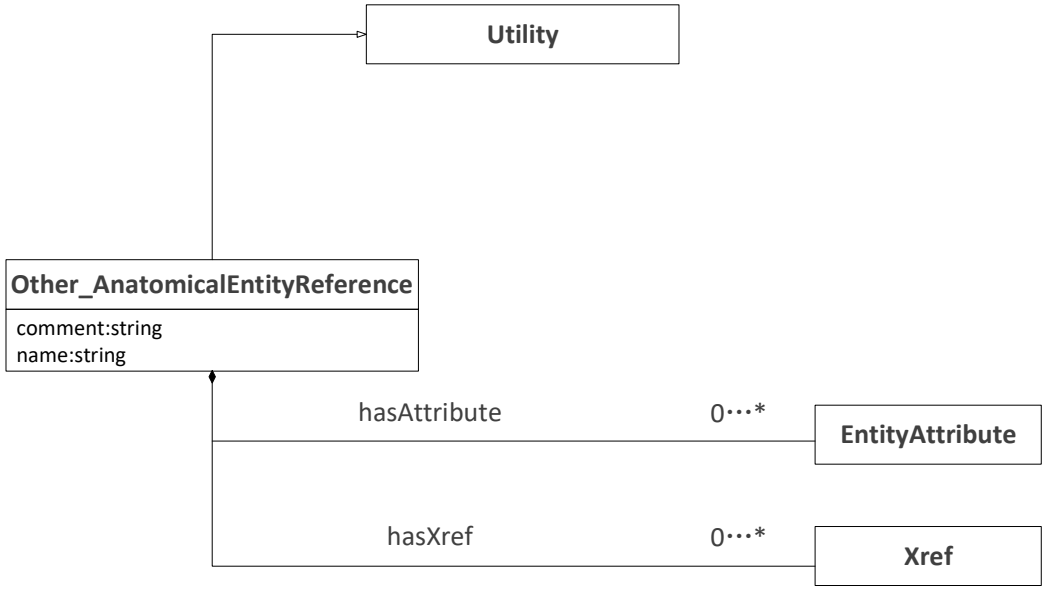

Figure 2-40

（4） SubstanceReference

Used to store shared information about a set of related substance differing states such as protein molecules encoded by the same gene. Its definition is shown in Figure 2-41. **SubstanceReference** has one sub-class which is **ProteinReference** of which the structure is the same. This class is used to define a generic antigen or a generic antibody.

hasXref: (0 or more object:Xref) Values of this property define external cross-references from this entity to entities in external databases (e.g. controlled vocabulary). PathoML recommends using "Biological macromolecule" and "Portion of body substance" from Tumor Pathology Ontology or UniProt.

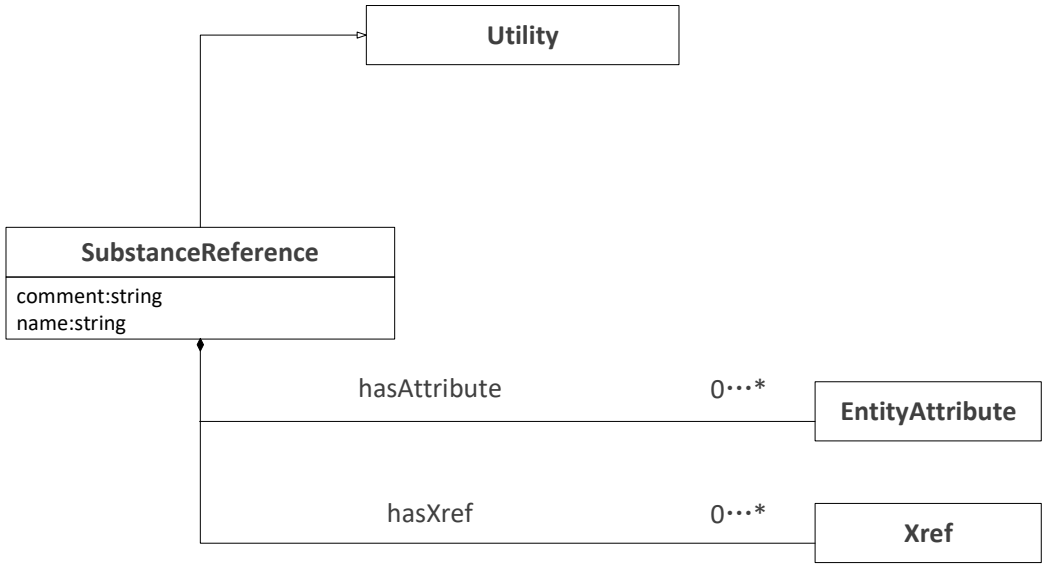

Figure 2-41

2.8.6 Provenance

The direct source of the data represented using PathoML such as a database. The *hasXref* property may contain a **PublicationXref** referencing a publication describing the data source (e.g. a database publication). Its definition is shown in Figure 2-42.

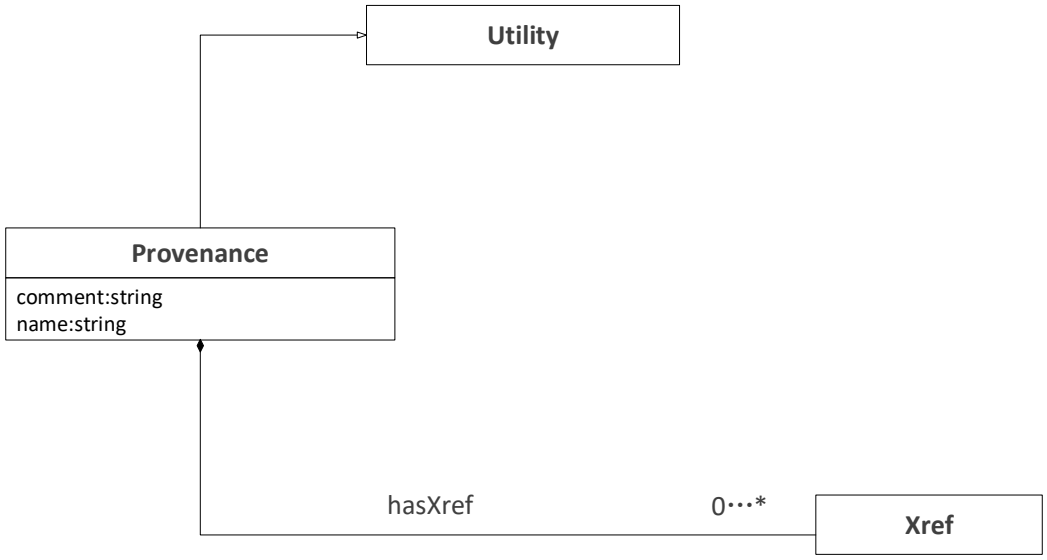

Figure 2-42

2.8.7 Quantification

Quantification is used to quantitatively describe an attribute of an entity which provides the absolute amount or the numerical range of the attribute. Its definition is shown in Figure 2-43.

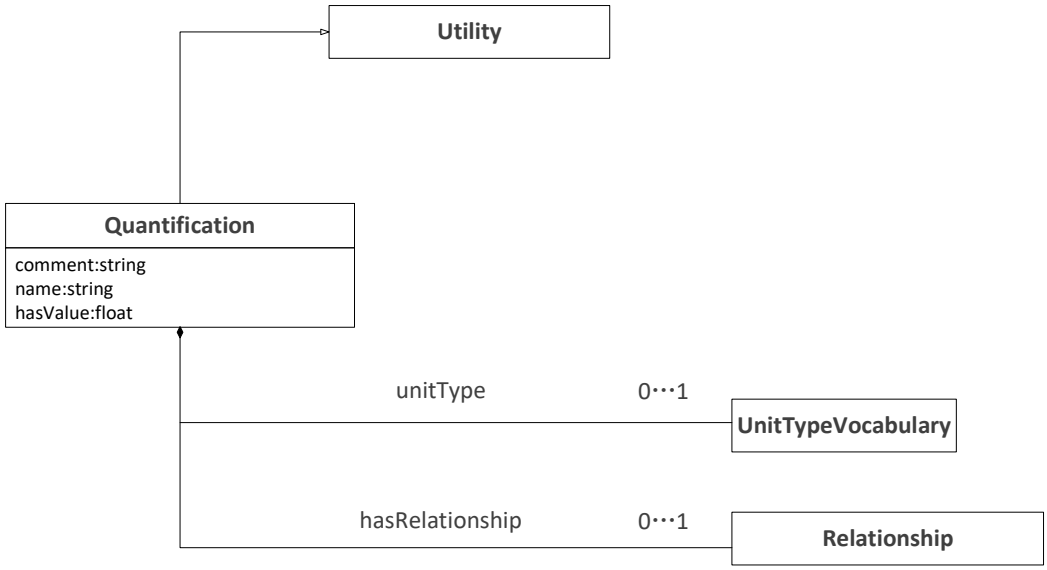

Figure 2-43

hasValue: (xsd:float) The absolute amount of the attribute.

unitType: (0 or 1 object:UnitTypeVocabulary) The unit of the attribute. The unit depends on the quality of the attribute. For example, if the quality is the diameter of a cell, then the unit is micrometer.

hasRelationship: (0 or 1 object:Relationship) The numerical relationship between the attribute and the amount such as “is equal to”, “is greater than or equal to”, and “is less than or equal to” etc. *hasXref* of the **Relationship** would reference the term of controlled vocabulary referring to this numerical relationship.

2.8.8 Relationship

The relationships between entities. The relationships are represented as triple. The head entity and tail entity are represented using *subject* and *object* respectively. Its definition is shown in Figure 2-44.

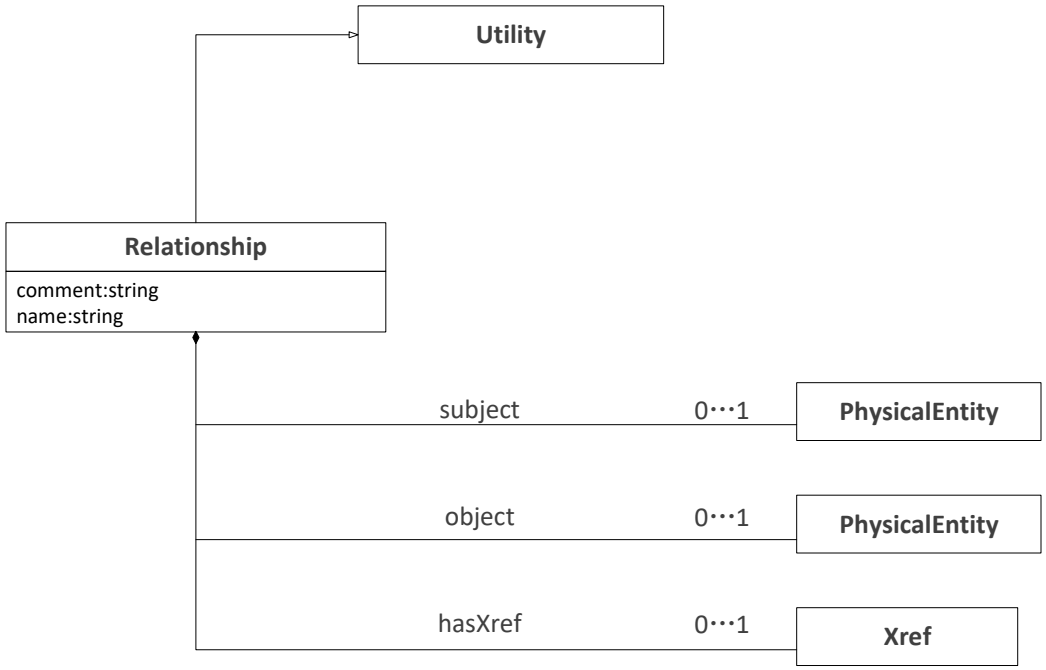

Figure 2-44

subject: (0 or 1 object:PhysicalEntity) The head entity of the relationship.

object: (0 or 1 object:PhysicalEntity) The tail entity of the relationship.

hasXref: (0 or more object:Xref) Values of this property define external cross-references from this entity to entities in external databases (e.g. controlled vocabulary). PathoML recommends using "Qualifier" as well as its sub-classes from Tumor Pathology Ontology.

2.8.9 Xref

A reference from an instance of a class in PathoML to an object in an external resource (e.g. controlled vocabulary). Its definition is shown in Figure 2-45. **Xref** has three sub-classes which are **PublicationXref**, **RelationshipXref** and **UnificationXref**.

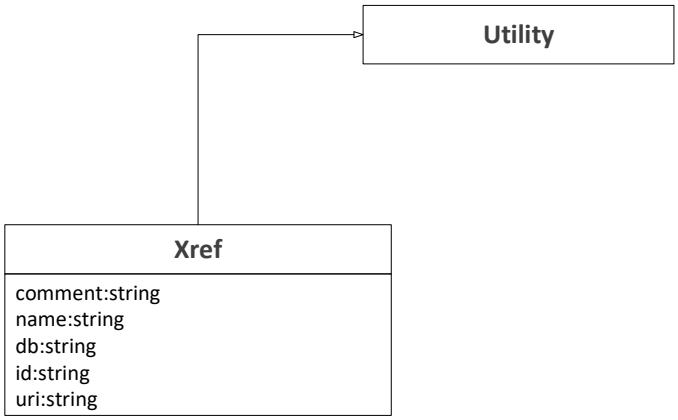

Figure 2-45

db: (xsd:string) The name of the external database to which this xref refers.

id: (xsd:string) The identifier in the external database of the object to which this xref refers.

uri: (xsd:string) Uniform resource identifier of referencing term, such as IRI (Internationalized Resource Identifier) of an ontology term.

（1）PublicationXref

An xref that defines a reference to a publication such as a journal article, book, web page, or software manual. References to PubMed are preferred when possible. Its definition is shown in Figure 2-46.

author: (xsd:string) The authors of this publication, one per property value.

db: (xsd:string) PubMed or ISBN.

id: (xsd:string) The PubMed of an academic paper or ISBN number of book if it is available.

title: (xsd:string) The title of the publication.

year: (xsd:string) The year when this publication published.

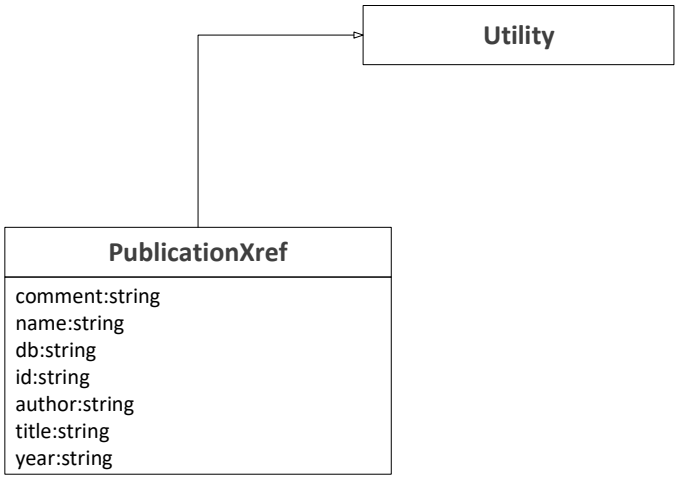

Figure 2-46

（2）RelationshipXref

An xref that defines a reference to an entity in an external resource that does not have the same biological identity as the referring entity. Its definition is shown in Figure 2-47.

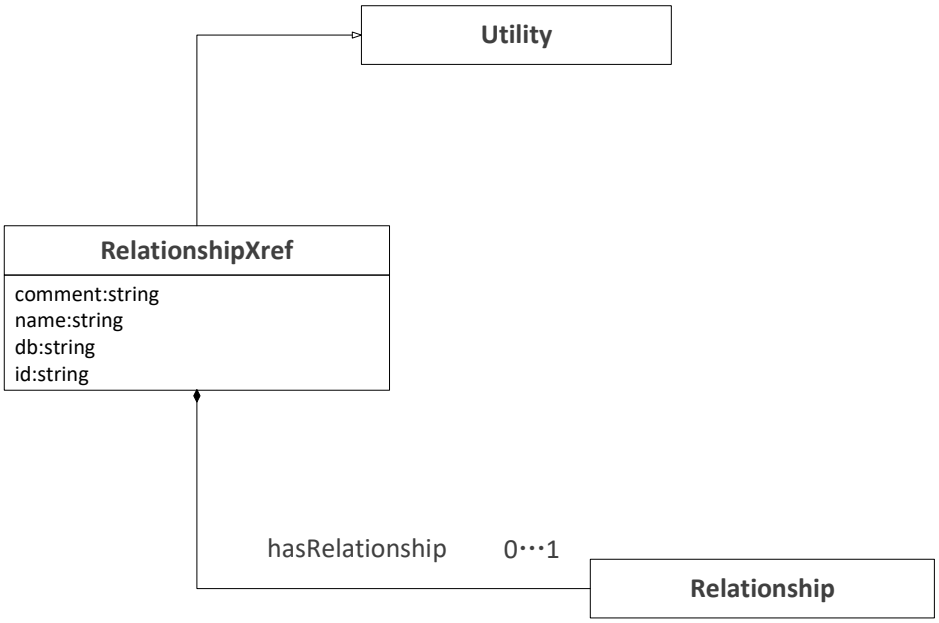

Figure 2-47

hasRelationship: (0 or 1 object:Relationship) This property names the type of relationship between the PathoML object linked from and the external object linked to.

（3）UnificationXref

A **UnificationXref** defines a reference to an entity in an external resource that has the same biological identity as the referring entity. **UnificationXref** should be used whenever possible since it improves data integration and semantic interoperability. Its definition is shown in Figure 2-48.

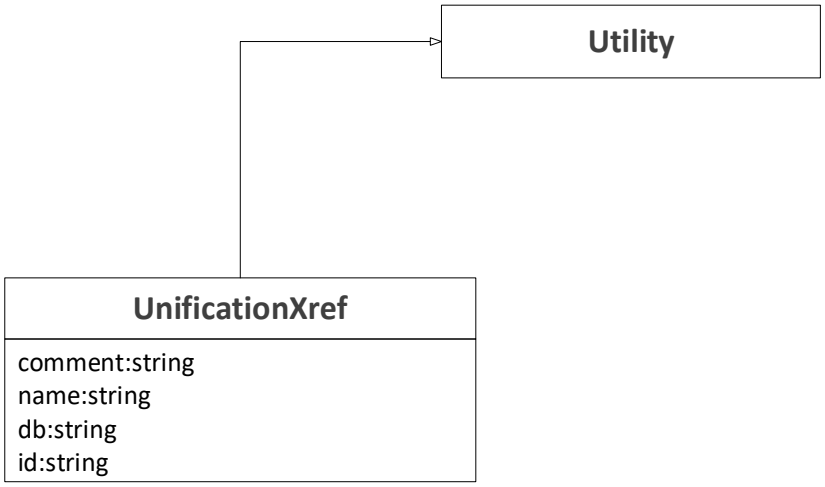

Figure 2-48

2.9 Data

Entity and **Utility** represent pathological features of pathology data, while **Data** stores metadata of the data files. Take digital slide as an example, the metadata include height, width, magnification, brand of the scanner and equipment settings used to capture the slide. **Data** has three sub-classes which are **IHC_Slide**, **HE_Slide** and **Pathology_Report**.

2.9.1 IHC_Slide

IHC_Slide stores metadata of an IHC slide. Its definition is shown in Figure 2-49.

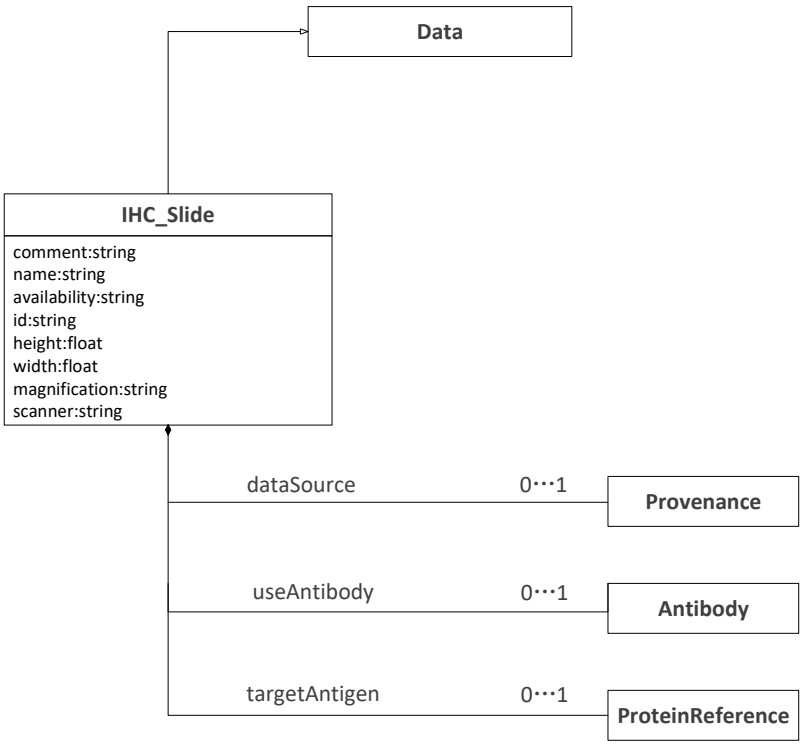

Figure 2-49

dataSource: (0 or more object:Provenance) A description of the source of this data such as a database name (The Cancer Genome Atlas).

useAntibody: (0 or more object:Antibody) A description of the antibody used to stained the slide.

targetAntigen: (0 or more object:ProteinReference) A description of the target antigen.

`id`: (xsd:string) Identifier of this slide from the database.

`height`: (xsd:float) Height of the slide.

`width`: (xsd:float) Width of the slide.

`magnification`: (xsd:string) Magnification at which the slide was scanned.

`scanner`: (xsd:string) Model name of the scanner.

2.9.2 HE_Slide

HE_Slide stores metadata of a HE slide. Its definition is shown in Figure 2-50.

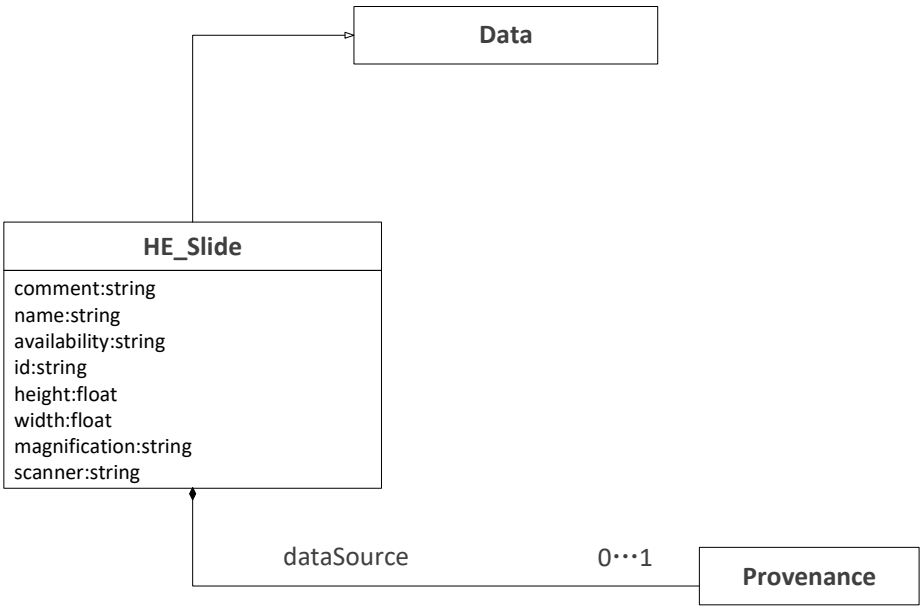

Figure 2-50

dataSource: (0 or more object:Provenance) A description of the source of this data such as a database name (The Cancer Genome Atlas).

id: (xsd:string) Identifier of this slide from the database.

height: (xsd:float) Height of the slide.

width: (xsd:float) Width of the slide.

magnification: (xsd:string) Magnification at which the slide was scanned.

scanner: (xsd:string) Model name of the scanner.

2.9.3 Pathology_Report

Pathology_Report stores metadata of a pathology report. Its definition is shown in Figure 2-51.

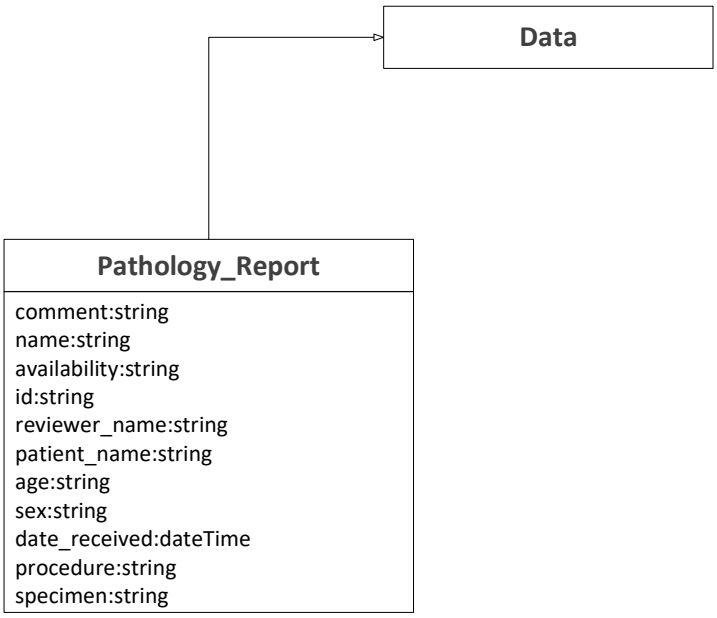

Figure 2-51

dataSource: (0 or more object:Provenance) A description of the source of this data such as a database name (The Cancer Genome Atlas).

`id`: (xsd:string) Identifier of this report from the database.

`patient_name`: (xsd:string) Name of the patient.

`procedure`: (xsd:string) Type of procedure obtaining the tissue such as biopsy or surgery.

`specimen`: (xsd:string) Specimen laterality.