Informatics for Integrating Biology and the Bedside

i2b2 Design Document

Ontology Management (ONT) Cell

Document Version: *1.6.3*
I2b2 Software Version: *1.6*

Table of Contents

DOCUMENT MANAGEMENT

Revision Number	Date	Author	Description of change
1.6.1	07/22/10	Janice Donahoe	Created 1.6 version of document.
	07/15/11	Lori Phillips	Updated for modifiers
1.6.3	10/6/11	Mike Mendis	Minor Changes

1. INTRODUCTION

This document describes the functionality of the **Ontology Management (ONT) cell**. It is to be used as a guideline and continuing reference as the developers write the code.

2. RELATIONSHIP OF THE I2B2 ONTOLOGY TO STAR SCHEMA

2.1 Data Storage

The i2b2 data is stored in a relational database, usually either Oracle or SQL Server, and always in a **star schema** format. A star schema contains one fact and many dimension tables. The fact table contains the quantitative or factual data, while the dimension tables contain descriptors that further characterize the facts. Facts are defined by concept codes and the hierarchical structure of these codes together with their descriptive terms and some other information forms the i2b2 ontology (also called metadata).

i2b2 ontology data may consist of one or many tables. If there is one table, it will contain all the possible data types or categories. The other option is to have one table for each data type. Examples of data types are: diagnoses, procedures, demographics, lab tests, encounters (visits or observations), providers, health history, transfusion data, microbiology data and various types of genetics data. All metadata tables must have the same basic structure. This document will discuss the case of using one ontology table that holds all data types.

The structure of the metadata is integral to the visualization of concepts in the i2b2 workbench, as well as for querying the data. The next two sections are a representation of the i2b2 ontology table and a discussion of the fields therein.

2.2 Ontology Table

COLUMN NAME	DATA TYPE (ORACLE)	DATA TYPE (SQL)
C_HLEVEL	INT	INT
C_FULLNAME	VARCHAR2(1500)	VARCHAR(700)
C_NAME	VARCHAR2(2000)	VARCHAR(2000)
C_SYNONYM_CD	CHAR(1)	CHAR(1)
C_VISUALATTRIBUTES	CHAR(3)	CHAR(3)
C_TOTALNUM	INT	INT
C_BASECODE	VARCHAR2(50)	VARCHAR(50)
C_METADATAXML	CLOB	TEXT

C_FACTTABLECOLUMN	VARCHAR2(50)	VARCHAR(50)
C_TABLENAME	VARCHAR2(50)	VARCHAR(50)
C_COLUMNNAME	VARCHAR2(50)	VARCHAR(50)
C_COLUMNDATATYPE	VARCHAR2(50)	VARCHAR(50)
C_OPERATOR	VARCHAR2(10)	VARCHAR(10)
C_DIMCODE	VARCHAR2(700)	VARCHAR(700)
C_COMMENT	CLOB	TEXT
C_TOOLTIP	VARCHAR2(900)	VARCHAR(900)
UPDATE_DATE	DATE	DATETIME
DOWNLOAD_DATE	DATE	DATETIME
IMPORT_DATE	DATE	DATETIME
SOURCESYSTEM_CD	VARCHAR2(50)	VARCHAR(50)
VALUETYPE_CD	VARCHAR2(50)	VARCHAR(50)
M_APPLIED_PATH	VARCHAR2(700)	VARCHAR(700)
M_EXCLUSION_CD	VARCHAR2(25)	VARCHAR(25)
C_PATH	VARCHAR2(1300)	VARCHAR(700)
C_SYMBOL	VARCHAR2(200)	VARCHAR(50)

2.3 Definition of Fields in the Ontology Table

2.3.1 c_hlevel

c_hlevel is the hierarchical level of the term. The term at the highest level of a hierarchy has a value of 0, the next level has a value of 1 and so on.

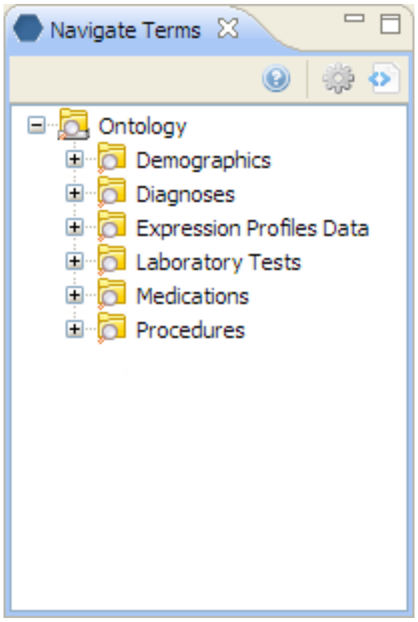
The screen shots below show how the values in *c_hlevel* determine the way ontology data looks in the user interface.

- The name of the ontology table is I2B2; the entry with c_hlevel 0 has c_name = ‘Ontology’ and is the root of the ontology tree.

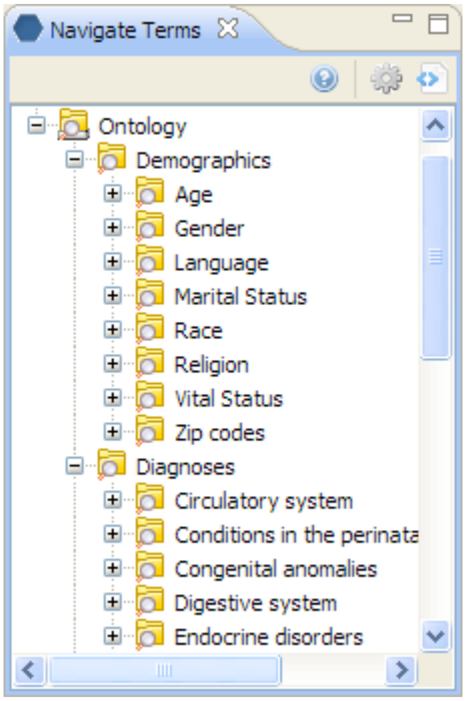
- The folders underneath Ontology all have c_hlevel =1.

- When a user clicks on a plus sign (⊞) to open a folder, the next level to open has the value c_hlevel =2.Thus the field c_hlevel keeps terms grouped in hierarchical order.

Example 1: **c_hlevels 0 and 1**



Example 2: c_hlevels 0, 1 and 2



- Ontology
  - Demographics
    - Age
    - Gender
    - Language
    - Marital Status
    - Race
    - Religion
    - Vital Status
    - Zip codes
  - Diagnoses
    - Circulatory system
    - Conditions in the perinata
    - Congenital anomalies
    - Digestive system
    - Endocrine disorders

2.3.2 c_fullname

c_fullname is the hierarchical path that leads to the term. Below is an example of *c_fullname* for the term ‘Rheumatoid arthiritis’. It is shown on several lines but is actually one concatenated line in the *c_fullname* field. Each ‘\’ represents another hierarchical level.

\i2b2
\Diagnoses
\Musculoskeletal and connective tissue (710-739)
\Arthropathies (710-719)
\(714) Rheumatoid arthritis and other arthropathies
\(714-0) Rheumatoid arthritis

2.3.3 c_name

c_name is the descriptive text value for the term. It is what is displayed in the user interface.

2.3.4 c_synonym_cd

c_synonym_cd is a boolean field that indicates whether the field is a synonym for another term or not A ‘Y’ in this field denotes that the field is a synonym, while an ‘N’ means this is the original term. The default values is ‘N’, so all terms start out with ‘N’ and if synonyms are added they get the value ‘Y’. . Two or more fields that are synonyms of each other will have the same c_basecode (defined below).

2.3.5 c_visualattributes

c_visualattributes describes how the field looks in the user interface. It is a 3 character field, with the following possible values:

1st character:

F = Folder
C = Container
M = Multiple
L = Leaf
O = Modifier container
D = Modifier folder
R = Modifier leaf

2nd character:

A = Active
I = Inactive
H = Hidden

3rd character:

E = editable

Folders (📁,📁) and **containers** (📁,📁) are the yellow rectangles with plus signs next to them that can be expanded to display other folders or leaves. Concept folders and containers contain a magnifying glass; modifier folders and containers contain a blue bulls eye.The difference between a container and a folder is that a container may not be dragged into a panel in the workbench as a query item, while a folder can be a query item. i2b2 primarily uses folders, which means that most terms can be used in queries.

Leaves (🔍, ◉) are the lowest level of a hierarchy. They cannot be expanded any further. Concepts are depicted by a grey rectangle with a magnifying glass, whereas modifiers contain a blue bulls eye.

Multiples are terms where there is more than one term mapped to an item, but only one is displayed. An example is under Gender in the Demographics folder – the term ‘Unknown’ has a black dot in the magnifying glass indicating that there are at least two terms that are considered to be ‘Unknown Gender’ and both are mapped to this one.

The second character of *c_visualattributes* describes the status of the term. An **active** term is displayed normally. An **inactive** term is greyed out. It appears in the interface to let the user know it is there, but it cannot be used. A **hidden** term is just that – it is hidden from the user entirely.

The third character of *c_visualattributes* indicates that the term is editable. If a term is a folder or container, a child term can be added to it. Editable terms may also be deleted.

2.3.6 c_totalnum

If available, *c_totalnum* indicates the total number of patients having that concept.

Since a single modifier can apply to more than one concept, this column is not used and does not apply for modifiers.

2.3.7 c_basecode

c_basecode this is the term that describes the ontological concept. This may be an ICD9 code (for diagnoses), or an NDC code (for medications) or a LOINC code (for lab tests). Or it may be any number of other coding systems, even home-grown ones.

2.3.8 c_metadataxml

c_metadataxml is an optional field to store extra information about the concept in xml format. It is currently used to describe value metadata associated with a lab finding.

The next several fields, *c_facttablecolumn*, *c_tablename*, *c_columnname*, *c_operator*, *c_dimcode*, are used to help construct a metadata SELECT SQL query that runs behind the scenes. The intent of this query is to link the dimension tables to the fact table for a given term. As a result every metadata SELECT SQL statement should return a fact table key.

In general the metadata SELECT SQL that is composed looks like the following:

select *c_facttablecolumn* from *c_tablename* where *c_columnname c_operator c_dimcode*

For most concept_dimension based queries this will appear as:

select concept_cd from concept_dimension where concept_path LIKE '\Diagnoses\Circulatory system\%'

For a patient_dimension based query this may appear as:

select patient_num from patient_dimension where birth_date BETWEEN 'getdate() AND getdate() – 365.25(10) '

For a visit_dimension based query this may appear as:

select encounter_num from visit_dimension where inout_cd = 'I'

For a provider_dimension based query this may appear as:

select provider_id from provider_dimension where provider_path LIKE '\Providers\Emergency\%'

2.3.9 c_facttablecolumn

c_facttablecolumn is the name of a key in the fact table (observation_fact) that links to the dimension code we are querying for.

Typical entries will be concept_cd, patient_num, encounter_num, provider_id

2.3.10 c_tablename

c_tablename is the name of the dimension table that holds the metadata to fact linkings.

Typical entries will be concept_dimension, patient_dimension, visit_dimension, provider_dimension

2.3.11 c_columnname

c_columnname is the name of the field in the c_tablename that holds the dimension code we are querying for.

Typical entries might be concept_path, birth_date or income_cd, inout_cd, length_of_stay or provider_path

2.3.12 c_columndatatype

c_columndatatype is either ‘T’ for text or ‘N’ for numeric and describes the datatype of the concept or term.

2.3.13 c_operator

c_operator is any valid SQL operator used in the WHERE clause of the metadata SELECT SQL query.

Typical entries are: ‘LIKE’, ‘BETWEEN’, ‘IN’, ‘=”.

2.3.14 c_dimcode

c_dimcode is the actual value of the dimension table c_columnname that we are querying for.

Typical entries are an actual: concept_path (\Diagnoses\Circulatory system\), birth_date range (‘getdate() AND getdate() – 365.25(10) ’ , inout_cd (‘I’) or provider_path (\Providers\Emergency\)

2.3.15 c_comment

c_comment is an optional field to store miscellaneous comments

2.3.16 c_tooltip

c_tooltip is the tooltip that appears in the user interface for a given term. It is usually the c_fullname with spaces around the ‘\’ for readability.

2.3.17 update_date

update_date is the date the data was updated.

2.3.18 download_date

download_date is the date the data was downloaded.

2.3.19 import_date

import_date is the date the data was imported.

2.3.20 sourcesystem_cd

sourcesystem_cd is a coded value for the source system from which the data was loaded or derived.

2.3.21 valuetype_cd

valuetype_cd is a coded value indicating the term type. At present there are two values in use: the type 'DOC' indicates terms that represent documents or notes; the type 'LAB' indicates terms of a laboratory test nature.

2.3.22 m_applied_path

Introduced in 1.6 to support modifier terms within the metadata table, *m_applied_path* is the concept path that the term applies to. Traditional (non-modifier) concept terms have an m_applied_path of '@'.

An m_applied_path of '\Diagnoses\Circulatory system\%' means that the term is a modifier that applies to the term(s) with c_fullname of \Diagnoses\Circulatory system\ and all its descendents, whereas an m_applied_path of '\Diagnoses\Circulatory system\' applies to the term with c_fullname of '\Diagnoses\Circulatory system\' only.

2.3.23 m_exclusion_cd

Introduced in 1.6 to support modifier terms within the metadata table, a non-null ('X') *m_exclusion_cd* indicates the modifier is to be excluded from the specified applied path. Traditional concept terms and non-exclusion modifiers have an m_exclusion_cd of null. An m_applied_path of '\Diagnoses\Circulatory system\%' and m_exclusion_cd of 'X' means that the term is a modifier that is excluded from the term(s) with c_fullname of \Diagnoses\Circulatory system\ and all its descendents, whereas an m_applied_path of '\Diagnoses\Circulatory system\' would be excluded from the term with c_fullname of '\Diagnoses\Circulatory system\' only.

2.3.24 c_path

A subset of c_fullname; its meant to contain the c_fullname of the node’s parent. A node’s c_path, concatenated with its c_symbol (below) form the node’s c_fullname.

2.3.25 c_symbol

c_symbol is a unique, abbreviated form of the node’s c_name. A node’s c_symbol, prepended with its c_path (above) form the node’s c_fullname.

3. SAMPLE ONTOLOGY QUERIES

3.1 Query Sample for Diagnoses

ICD-9 code is known:

To lookup the *c_basecode* and *c_fullname* for ICD-9 diagnosis code 346.0, use this query:

```
Select c_basecode, c_fullname
From rpdr
Where c_basecode ='3460'
```

The *c_basecode* returned in the results can then be joined to the *concept_cd* in the Observation_Fact table to find all patients diagnosed with ICD-9 code 346.0. Note that the c_basecode 3460 has no decimal point, these are removed.

ICD-9 code is unknown, but the diagnosis description is known:

To lookup the *c_basecode* and *c_fullname* for the diagnosis of migraines, use this query:

```
Select c_basecode, c_fullname
From rpdr
where c_fullname like '%diagnoses%migraine%'
```

The *c_basecodes* returned in the results could then be joined to the *concept_cd* in the Observation_Fact table to find all patients diagnosed with migraines.

3.2 Query Sample for Problems

To find all the patients that were diagnosed with migraines, use this query:

```
Select distinct(patient_num)
From  observation_fact
Where concept_cd in
(select concept_cd
from concept_dimension
where concept_path like '%Neurologic Disorders (320-389)\(346) Migraine\%')
```

To find the ages of all patients that were diagnosed with migraines, use this query:

```
Select concept_cd
From  observation_fact
where concept_cd like 'DEM|Age%'
and patient_num in
(select patient_num from observation_fact
where concept_cd in
(select concept_cd
from concept_dimension
where concept_path like '%Neurologic Disorders (320-389)\(346) Migraine\%'))
```

3.3 Query Sample for Labs

If we wanted to get all the ages for patients having a Cholesterol lab, we could run the following query:

```
Select concept_cd
From  observation_fact
where concept_cd like 'DEM|Age%'
and patient_num in
(select patient_num from observation_fact
where concept_cd in
(select concept_cd
from concept_dimension
where concept_path like '%LAB\(LLB16) Chemistry\(LLB17) Lipid Tests\CHOL\%'))
```

Notice how the path of the concept is used to query all concept ids that fall into the Cholesterol group. If we only wanted to query for patients with Plasma Cholesterol only, we would use the same query with the following path joined against c_fullname:

‘LAB\(LLB16) Chemistry\(LLB17) Lipid Tests\CHOL\MCSQ-PCHOL\%’

Or

‘LAB\(LLB16) Chemistry\(LLB17) Lipid Tests\CHOL\MCPCHOL\%’