i2b2 Design Document
Ontology Management (ONT) Cell
TABLE OF CONTENTS

DOCUMENT MANAGEMENT .................................................................................................................. 3

1 INTRODUCTION .......................................................................................................................... 4

2 RELATIONSHIP OF THE I2B2 ONTOLOGY STAR SCHEMA ......................................................... 5
   2.1 DATA STORAGE ......................................................................................................................... 5
   2.2 ONTOLOGY TABLE .................................................................................................................... 5
   2.3 DEFINITION OF FIELDS IN THE ONTOLOGY TABLE ................................................................. 6
       2.3.1 C_HLEVEL ......................................................................................................................... 6
       2.3.2 C_FULLNAME ................................................................................................................... 8
       2.3.3 C_NAME ............................................................................................................................ 8
       2.3.4 C_SYNONYM_CD ............................................................................................................... 8
       2.3.5 C_VISUALATTRIBUTES ................................................................................................. 8
       2.3.6 C_TOTALNUM ................................................................................................................... 10
       2.3.7 C_BASECODE .................................................................................................................... 10
       2.3.8 C_METADATAXML .............................................................................................................. 10
       2.3.9 C_FACTTABLECOLUMN ................................................................................................. 11
       2.3.10 C_TABLENAME ............................................................................................................... 12
       2.3.11 C_COLUMNNAME ............................................................................................................. 12
       2.3.12 C_COLUMNDATATYPE .................................................................................................... 12
       2.3.13 C_OPERATOR .................................................................................................................... 12
       2.3.14 C_DIMCODE .................................................................................................................... 12
       2.3.15 C_COMMENT ..................................................................................................................... 13
       2.3.16 C_TOOLTIP ........................................................................................................................ 13
       2.3.17 UPDATE_DATE ............................................................................................................... 13
       2.3.18 DOWNLOAD_DATE ........................................................................................................ 13
       2.3.19 IMPORT_DATE ............................................................................................................... 13
       2.3.20 SOURCESYSTEM_CD ....................................................................................................... 13
       2.3.21 VALUETYPE_CD ............................................................................................................. 13
       2.3.22 M_APPLIED_PATH .......................................................................................................... 14
       2.3.23 M_EXCLUSION_CD ......................................................................................................... 14
       2.3.24 C_PATH ............................................................................................................................ 14
       2.3.25 C_SYMBOL ....................................................................................................................... 14

3 SAMPLE ONTOLOGY QUERIES .................................................................................................... 15
   3.1 QUERY SAMPLE FOR DIAGNOSES ....................................................................................... 15
   3.2 QUERY SAMPLE FOR PROBLEMS ....................................................................................... 15
   3.3 QUERY SAMPLE FOR LABS .................................................................................................. 16
<table>
<thead>
<tr>
<th>Revision Number</th>
<th>Date</th>
<th>Author</th>
<th>Description of change</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.7.1</td>
<td>10/17/12</td>
<td>Janice Donahoe</td>
<td>Created 1.7 version of the document.</td>
</tr>
<tr>
<td>1.7.00-002</td>
<td>06/08/2015</td>
<td>Janice Donahoe</td>
<td>Fixed some spelling and grammar issues.</td>
</tr>
<tr>
<td>1.7.08-003</td>
<td>10/04/2016</td>
<td>Janice Donahoe</td>
<td>Fixed some spelling errors.</td>
</tr>
</tbody>
</table>
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 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, encounter (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

<table>
<thead>
<tr>
<th>Column Name</th>
<th>Data Type (Oracle)</th>
<th>Data Type (SQL)</th>
</tr>
</thead>
<tbody>
<tr>
<td>C_HLEVEL</td>
<td>INT</td>
<td>INT</td>
</tr>
<tr>
<td>C_FULLNAME</td>
<td>VARCHAR2(1500)</td>
<td>VARCHAR(700)</td>
</tr>
<tr>
<td>C_SYNONYM_CD</td>
<td>CHAR(1)</td>
<td>CHAR(1)</td>
</tr>
<tr>
<td>C_VISUALATTRIBUTES</td>
<td>CHAR(3)</td>
<td>CHAR(3)</td>
</tr>
<tr>
<td>C_TOTALNUM</td>
<td>INT</td>
<td>INT</td>
</tr>
<tr>
<td>C_BASECODE</td>
<td>VARCHAR2(50)</td>
<td>VARCHAR(50)</td>
</tr>
<tr>
<td>C_METADATAXML</td>
<td>CLOB</td>
<td>TEXT</td>
</tr>
<tr>
<td>C_FACTTABLECOLUMN</td>
<td>VARCHAR2(50)</td>
<td>VARCHAR(50)</td>
</tr>
<tr>
<td>C_TABLENAME</td>
<td>VARCHAR2(50)</td>
<td>VARCHAR(50)</td>
</tr>
<tr>
<td>C_COLUMNNAME</td>
<td>VARCHAR2(50)</td>
<td>VARCHAR(50)</td>
</tr>
<tr>
<td>Column Name</td>
<td>Data Type (Oracle)</td>
<td>Data Type (SQL)</td>
</tr>
<tr>
<td>--------------------------</td>
<td>--------------------</td>
<td>-----------------</td>
</tr>
<tr>
<td>C_COLUMNDATATYPE</td>
<td>VARCHAR2(50)</td>
<td>VARCHAR(50)</td>
</tr>
<tr>
<td>C_OPERATOR</td>
<td>VARCHAR2(10)</td>
<td>VARCHAR(10)</td>
</tr>
<tr>
<td>C_DIMCODE</td>
<td>VARCHAR2(700)</td>
<td>VARCHAR(700)</td>
</tr>
<tr>
<td>C_COMMENT</td>
<td>CLOB</td>
<td>TEXT</td>
</tr>
<tr>
<td>C_TOOLTIP</td>
<td>VARCHAR2(900)</td>
<td>VARCHAR(900)</td>
</tr>
<tr>
<td>UPDATE_DATE</td>
<td>DATE</td>
<td>DATETIME</td>
</tr>
<tr>
<td>DOWNLOAD_DATE</td>
<td>DATE</td>
<td>DATETIME</td>
</tr>
<tr>
<td>IMPORT_DATE</td>
<td>DATE</td>
<td>DATETIME</td>
</tr>
<tr>
<td>SOURCESYSTEM_CD</td>
<td>VARCHAR2(50)</td>
<td>VARCHAR(50)</td>
</tr>
<tr>
<td>VALUETYPE_CD</td>
<td>VARCHAR2(50)</td>
<td>VARCHAR(50)</td>
</tr>
<tr>
<td>M_APPLIED_PATH</td>
<td>VARCHAR2(700)</td>
<td>VARCHAR(700)</td>
</tr>
<tr>
<td>M_EXCLUSION_CD</td>
<td>VARCHAR2(25)</td>
<td>VARCHAR(25)</td>
</tr>
<tr>
<td>C_PATH</td>
<td>VARCHAR2(1300)</td>
<td>VARCHAR(700)</td>
</tr>
<tr>
<td>C_SYMBOL</td>
<td>VARCHAR2(200)</td>
<td>VARCHAR(50)</td>
</tr>
</tbody>
</table>

### 2.3 Definition of Fields in the Ontology Table

#### 2.3.1 C_HLEVEL

The **C_HLEVEL** is the hierarchical level of the term. The term at the highest level of a hierarchy has a value of 0 and 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
Example 1: C_HLEVELS 0 and 1

Example 2: C_HLEVELS 0, 1, and 2
2.3.2 **C_FULLNAME**

The *C_FULLNAME* is the hierarchical path that leads to the term. Below is an example of the *C_FULLNAME* for the term "Rheumatoid arthritis". It is shown on several lines but is actually one concatenated line in the *C_FULLNAME* column. Each back slash (\) 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**

The *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**

The *C_SYNONYM_CD* is a Boolean field that indicates whether the item 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 value is “N” so all terms start out with “N” and if a synonym is added it gets the value of “Y”.

Two or more synonyms of each other will have the same *C_BASECODE* (defined below).

2.3.5 **C_VISUALATTRIBUTES**

The *C_VISUALATTRIBUTES* column describes how the field looks in the user interface. It is a 3 character field with the following possible values:
1st character:
- C = Container
- F = Folder
- 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

Containers and folders are the yellow rectangles with plus signs next to them that can be expanded to display other folders or leaves.

Concept folders and containers have a magnifying glass in the icon.

Modifier folders and containers have a blue bulls eye in the icon.

The difference between a container and a folder is that a container cannot 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.

Concept leaves are depicted by a grey rectangle with a magnifying glass. Concept leaves are depicted by 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 $C_{\text{VISUALATTRIBUTES}}$ describes the status of the term.

- An **active** term is displayed normally.
- An **inactive** term is greyed out; it appears in the client 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 $C_{\text{VISUALATTRIBUTES}}$ indicates that the term can be edited. If a term is a folder or container then a child term can be added to it. Editable terms may also be deleted.

### 2.3.6 $C_{\text{TOTALNUM}}$

If available, the $C_{\text{TOTALNUM}}$ indicates the total number of patient have 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 $C_{\text{BASECODE}}$

The $C_{\text{BASECODE}}$ is the term that describes the ontological concept. This may be an ICD9 code (for diagnoses), and 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 $C_{\text{METADATAXML}}$

The $C_{\text{METADATAXML}}$ 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, $C_{\text{FACTTABLECOLUMN}}, C_{\text{TABLENAME}}, C_{\text{COLUMNNAME}}, C_{\text{OPERATOR}},$ and $C_{\text{DIMCODE}}$ 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

The **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, or PROVIDER_ID.

### 2.3.10 C_TABLENAME

The $C_{\text{TABLENAME}}$ is the name of the dimension table that holds the metadata to fact linking.

Typical entries will be CONCEPT_CD, PATIENT_NUM, ENCOUNTER_NUM, or PROVIDER_ID.

### 2.3.11 C_COLUMNNAME

The $C_{\text{COLUMNNAME}}$ is the name of the field in the $C_{\text{TABLENAME}}$ that holds the dimension code we are querying for.

Typical entries might be CONCEPT_PATH, BIRTH_DATE, INCOME_CD, INOUT_CD, LENGTH_OF_STAY, or PROVIDER_PATH.

### 2.3.12 C_COLUMNNDATATYPE

The $C_{\text{COLUMNNDATATYPE}}$ is either “T” for text or “N” for numeric and describes the data type of the concept or term.

### 2.3.13 C_OPERATOR

The $C_{\text{OPERATOR}}$ is any valid SQL operator used in the WHERE clause of the metadata SELECT SQL query.

Typical entries are: “LIKE”, “BETWEEN”, “IN”, or “=”

### 2.3.14 C_DIMCODE

The $C_{\text{DIMCODE}}$ is the actual value of the dimension table $C_{\text{COLUMNNAME}}$ that we are querying for.

Typical entries are an actual:

- CONCEPT_PATH like ('Diagnoses\Circulatory system\')
- BIRTH_DATE range ('getdate() - 365.25(10)'
- INOUT_CD like ('I')
2.3.15 C_COMMENT

The C_COMMENT is an optional column to store miscellaneous comments about the term.

2.3.16 C_TOOLTIP

The 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 backslash ("\") for readability.

2.3.17 UPDATE_DATE

The UPDATE_DATE is the date the data was updated.

2.3.18 DOWNLOAD_DATE

The DOWNLOAD_DATE is the date the data was downloaded.

2.3.19 IMPORT_DATE

The IMPORT_DATE is the date the data was imported.

2.3.20 SOURCESYSTEM_CD

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

2.3.21 VALUETYPE_CD

The VALUETYPE_CD is a coded value indicating the type of term. At present there are two values in use:

- DOC = indicates the terms represents documents or notes
- LAB = indicates the term is of a laboratory test nature
2.3.22 M_APPLIED_PATH

Introduced in 1.6 to support modifier term within the metadata table, the M_APPLIED_PATH is the CONCEPT_PATH that the term applies to. Traditional (non-modifier) concept terms have a 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 descendants, whereas a 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 descendants, whereas a M_APPLIED_PATH of ‘\Diagnoses\Circulatory system\’ applies to 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

The C_SYMBOL is a unique, abbreviated form of the node’s C_NAME. A nodes C_SYMBOL, prepended with its C_PATH (above) for the node’s C_FULLNAME.
3 SAMPLE ONTOLOGY QUERIES

3.1 Query Sample for Diagnoses

ICD-9 code is known:

Use this query to lookup the C_BASECODE and C_FULLNAME for ICD-9 diagnosis code 346.0

```
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:

Use this query to lookup the C_BASECODE and C_FULLNAME for the diagnosis of migraines.

```
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

Use this query to find all the patients that were diagnosed with migraines.

```
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\%'
)
```

Use this query to find the ages of all patients that were diagnosed with migraines.
3.3 Query Sample for Labs

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

```sql
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\%')
  )
)
```

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 patient 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\%'`