Informatics for Integrating Biology and the Bedside

User Guide

NCBO Integration Tool

Document Version: *1.0.1*

i2b2 Software Version: *1.0*

Table of Contents

DOCUMENT MANAGEMENT

Revision Number	Date	Author	Description of change
1.0.0	06/09/12	Lori Phillips	Original document
1.0.1	08/02/12	Lori Phillips	Corrections plus feedback from Janice

ABSTRACT

This is a User's Guide for the NCBO Integration Tool. This guide will help you integrate mapping data into a target ontology metadata table. The end result will be a single metadata table consisting of a target ontology with mapped terms merged within.

1. BEFORE YOU BEGIN

1.1 Prequisites

1.1.1 Overview

Please read the document titled “NCBO Integration Overview”. It gives an overview of what the integraton tool does and what its intended usage is.

For the purpose of a demonstration example, consider an institution that is moving from ICD-9 to ICD10. Their legacy data is in ICD-9, while future data will be reported as ICD-10. This tool allows i2b2 users to create an ICD-10 hierarchical metadata table with ICD-9 terms merged within.

1.1.2 Terms used throughout this document

1.1.2.1 SOURCE TABLE

This is an i2b2 metadata table containing the desired hierarchy for the final table. In our example, this would be ICD-10. It is also referred to as the “mappedTo” table within this document.

1.1.2.2 TARGET TABLE

An i2b2 metadata integration table containing the mapped ICD-9 terms within an ICD-10 hierarchy.

1.1.2.3 MAPPING TABLE

This is the mapping table that contains the ICD-9 to ICD-10 mappings. This table is defined in the Ontology Mapping cell design documentation and is named PROJECT_ONT_MAPPING.

1.1.3 Concept path length considerations

Some users are very concerned with the maximum concept path (c_fullname) length of their metadata. Others don’t care. To support each user’s different needs we have made the concept path length configurable. The concept_path is made up of terms from each level of the path: \Level 1 term\Level 2 term\ …. and so on.

The integration tool requires a specified path format. If the source ontology (ICD-10) was generated via the NCBO Extraction tool, obtain the pathFormat information from the report output file associated with the extraction and use that. If not, choose one from the descriptions below.

We allow:

“As short as possible”	term at each level is reduced to a 4-char hash code. [S]
“Readable”	term at each level each term is reduced to 32-char [R]
“Something in between”	term at each level each term is reduced to 20-char [M]

If you decide to use the Short format be sure to read the note on page 8 about the possibility of duplicate non-synonymous c_fullname entries.

1.2 Software

1.2.1 Java JDK

JDK 6.0 is recommended and can be downloaded from the java website: http://java.sun.com/products/

1. Install the SDK into a directory of your choice.

   Example: /opt/java/jdk1.6.0 or *YOUR_JAVA_HOME_DIR*

1.2.2 Update Environment Variables

Be sure to set the JAVA_HOME, home directories you set up in the previous sections.

Example:

```
# Sample environment variables
JAVA_HOME=/opt/java/jdk1.6.0
export JAVA_HOME
```

2. INSTALLATION AND PREPARATION

2.1 Preparing the database

This package contains a folder called DatabaseScripts. Locate it now

- cd NCBOIntegration_1.0/DatabaseScripts

Scripts are provided for both oracle and sqlserver to create the target table populated by this program. Open create_{sqlserver/oracle}_metadata_tables.txt and

a. create target table INTEGRATION
b. create final table I2B2_INTEGRATED
c. If you have your own mapping table use that, otherwise create mapping table PROJECT_ONT_MAPPING and install mapping data into it.
d. If you have your own source (mappedTo) table use that, otherwise create one using the standard i2b2 metadata create table scripts for table I2B2. This table is also referred to as the ontology we are mapping to.

2.2 Run the Integration command line utility

This package contains a folder called Release_1_0. Locate it now

- cd NCBOIntegration_1.0/Release_1_0/

2.2.1 Configure database parameters

Configure the IntegrationApplicationContext.xml file to point to your source, target and mapping tables above. This configuration shown below is for SQLServer .. Insert the driverClassName and url for your database here. Database drivers for sqlserver and oracle have been included in folder lib/jdbc/.

```
<bean id="dataSource" class="org.apache.commons.dbcp.BasicDataSource"
destroy-method="close">
    <propertyname="driverClassName"
            value="com.microsoft.sqlserver.jdbc.SQLServerDriver"/>
    <property name="url" value="jdbc:sqlserver://your_db:port"/>
    <property name="username" value="your u-name"/>
    <property name="password" value="your password"/>
```

```
    <property name="defaultAutoCommit" value="true"/>
    <property name="defaultReadOnly" value="false"/>
  </bean>
```

Next, configure this for the schema that your integration tables reside in and the table names. Example shown here is for sqlserver.

```
<bean id="database" class="edu.harvard.i2b2.ncbo.model.DBInfoType">
  <property name="db_fullSchema" value="i2b2demodataICD10.dbo"/>
  <property name ="db_serverType" value="SQLSERVER"/>
  <property name="sourceTable" value="YOUR_SOURCE_TABLE"/>
  <property name="targetTable" value="integration"/>
  <property name="mappingTable" value="project_ont_mapping"/>
</bean>
```

2.2.2 Run the integration processing program

- java -classpath endorsed_lib/*:genlib/i2b2Common-core.jar:lib/commons/*:lib/log4j/*:lib/jdbc/*:lib/jdbc/sqlserver2005/*:lib/spring/*:* edu.harvard.i2b2.ncbo.integration.NCBOOntologyIntegrateAll –pathFormat Short

where

-pathFormat is an optional parameter that affects overall concept fullname length. (see sections 1.3.1 and 1.3.2) It may be set to:

S[hort] results in shortest possible concept path length

M[edium] results in a somewhat readable concept_path

R[eadable] results in a readable concept_path

If not specified, the tool defaults to M[edium].

A note about pathFormat = S[hort]

A pathFormat of S[hort] results in a 4-character symbol for each level of a c_fullname. While careful consideration was made in creating an algorithm that generates the 4-character symbol, it cannot absolutely guarantee a unique c_fullname for each term at a given level. We therefore recommend that you query for and manually edit duplicate non-synonymous c_fullname entries in your final metadata table.

```
select c_fullname, count(1) from ncbo_i2b2 where c_synonym_cd = 'N'
        group by c_fullname having count(1) > 1
```

Status will appear on screen….

NCBO Integration tool Version 1.0 May 2012

Creating symbols for records 1 to 10000

Creating symbols for records 10001 to 20000

Creating symbols for records 20001 to 20623

Creating tooltips for records 1 to 10000

Creating tooltips for records 10001 to 20000

Creating tooltips for records 20001 to 27947

Cleaning up fixed parameters and visual attributes

Integration complete

When processing is complete, the INTEGRATION table will contain your mapped terms within the hierarchy of your source (mapped to) ontology

2.2.3 Create a final metadata table

Copy the data from tables INTEGRATION and your source (I2B2 metadata) to table I2B2_INTEGRATED.

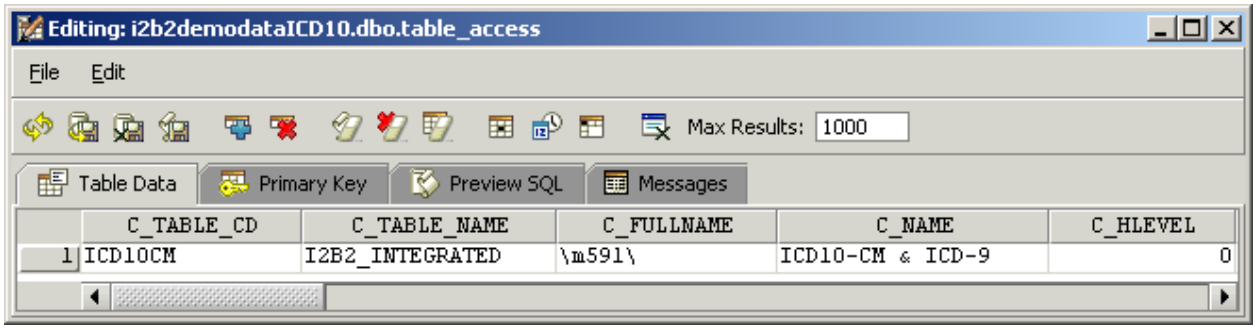
There is one step remaining. You will want to clean up any prior I2B2 metadata leaf nodes that now have integrated child concepts by making them folders.

```
update i2b2_integrated
set c_visualattributes = 'FA'
where c_visualattributes = 'LA'
and c_fullname in (select c_path from i2b2_integrated)
```

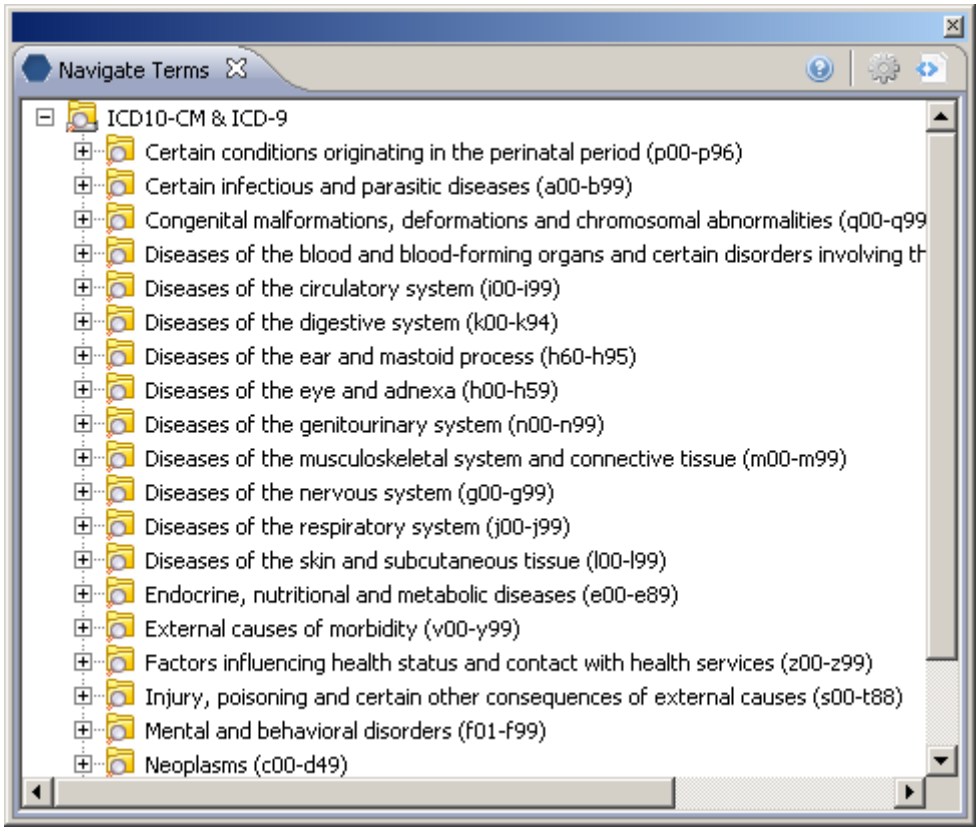
3. CONFIGURING I2B2 TO USE YOUR NEW METADATA

3.1 TABLE_ACCESS

Reconfigure table_access to include the root nodes of your final integrated metadata. There should be one entry per root node. The following example shows a configuration for our example. In this example our final integrated i2b2 metadata table was named 'I2B2_INTEGRATED', as shown in c_table_name column.

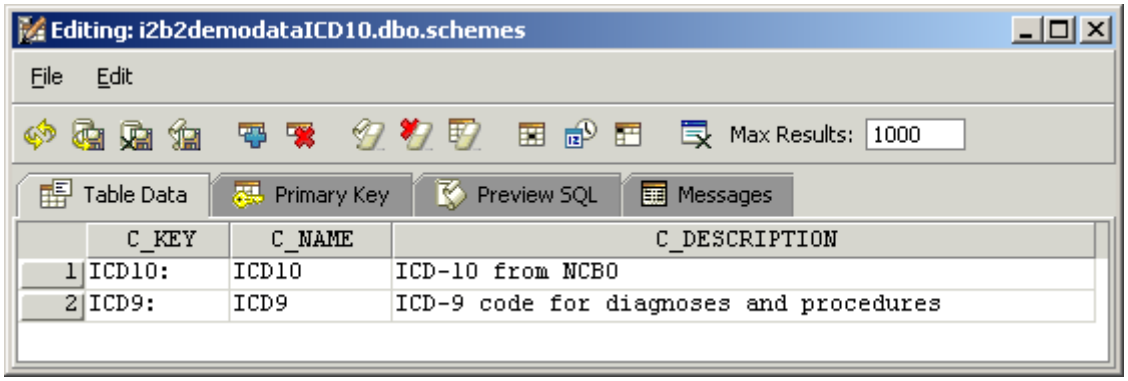


	C_TABLE_CD	C_TABLE_NAME	C_FULLNAME	C_NAME	C_HLEVEL
1	ICD10CM	I2B2_INTEGRATED	\m591\	ICD10-CM & ICD-9	0



3.2 SCHEMES

Reconfigure the SCHEMES table to include your integrated ontology's scheme or *prefix*. There should be one entry per scheme. The following example shows a configuration including the both the ICD9 and the ICD10 prefix.



	C_KEY	C_NAME	C_DESCRIPTION
1	ICD10:	ICD10	ICD-10 from NCBO
2	ICD9:	ICD9	ICD-9 code for diagnoses and procedures

3.3 CONCEPT_DIMENSION

You will need to synchronize your concept_dimension table so it contains the terms in your new metadata. Synchronization of metadata and concept_dimension is a feature found in the Edit View tool of the workbench. A user must have role of EDITOR in order to perform the synchronization process.

1. Click on the synchronize icon ( )or ( ) at the top of the view.

2. The **Update Ontology window** will open.

3. Click on the **OK button** to start the process.