|  |  |  |
| --- | --- | --- |
| Nebraska Lexicon© CDMV6 / i2b2 metadata release notes | | |
| A service of the Greater Plains Collaborative Research Network | | |
| Date | 20210623 | |
| Version | | 20210623 |



Amendment History

|  |  |  |  |
| --- | --- | --- | --- |
| **Version** | **Date** | **Editor** | **Comments** |
| 20180330 | 20180330 | James Campbell  Jay Pedersen | First release of SNOW SHRINE metadata |
| 20181001 | 20181207 | James Campbell  Jay Pedersen  Yeshwanth Narayana | Second release of SNOW SHRINE metadata; revisions to ICD-10-CM, RxNorm, NDC, SNOMED CT, LOINC |
| 20200930 | 20200930 | James Campbell  Jay Pedersen | ONC standard ontologies Nebraska Lexicon release for i2b2 and CDMV51 support |
| 20210301 | 20210428 | James Campbell  Jay Pedersen | All terminologies updated for Spring 2021 including COVID-19 codes; metadata (ontology) tables expanded to support PATH and TRANSITIVE CLOSURE folder searches |
| 20210623 | 20210623 | James Campbell | Added metadata and reference tables for ICD-O-3 and SNOMED CT anatomy |

Review Timetable

|  |  |  |
| --- | --- | --- |
| **Review date** | **Responsible owner** | **Comments** |
|  |  |  |

© University of Nebraska Medical Center

Nebraska Medicine

Table of Contents

[1 Purpose 4](#_Toc58337273)

[2 Audience of interest for this document 4](#_Toc58337274)

[3 Content scope and history 4](#_Toc58337275)

[4 Quality assurance, versioning and maintenance 6](#_Toc58337276)

[5 Description of Release Datasets and Instructions for use 6](#_Toc58337277)

[5.1 Nebraska Lexicon metadata (ontology) files 6](#_Toc58337278)

[5.1.1 SNOMED CT Conditions metadata 6](#_Toc58337279)

[5.1.2 LOINC laboratory metadata 6](#_Toc58337280)

[5.1.3 RXNORM/NDC medication metadata 7](#_Toc58337281)

[5.1.4 ICD-10-CM 7](#_Toc58337282)

[5.2 Metadata build procedures 7](#_Toc58337283)

[6 Deployment Instructions 7](#_Toc58337284)

[6.1 About I2b2 metadata set design 7](#_Toc58337285)

# Purpose

This document summarizes the datasets, use cases and deployment of i2b2 metadata and PCORnet CDMV6 supporting tables to support interoperation of the research data network employing i2b2 and in service to the larger community of ACT/SHRINE affiliated datamarts.

# Audience of interest for this document

Interested parties may include:

* GPC and collaborating ACT/SHRINE datamart information managers, programming staff and administrators
* I2b2 user community members with interest in interoperation of queries with others employing Office of the National Coordinator (ONC) for Health IT terminology standards
* Research network data managers evaluating resources for use by their collaborators

# Content scope and history

Dr Isaac Kohane, Dr Shawn Murphy and a team at Harvard University led the project which developed i2b2. Staff at Harvard developed the protocols of the SHRINE network which allows i2b2 datamarts to share queries across research networks. In order for those queries to interoperate, i2b2 concept metadata and concept coding of observation facts must be aligned or extensively mapped at every site supporting a datamart. In view of the fact that a large fraction of i2b2 research data originate in the electronic health record (EHR), the semantic interoperability framework imposed by ONC pragmatically defines a major portion of the terminology standards that could contribute to rapid development of interoperability in any US research data network.

Harvard promoted a metadata set for SHRINE networks which it named in reference to their information management scheme “Scalable Collaborative Infrastructure for a Learning Healthcare System” (SCILHS). In fall of 2017 the Greater Plains Collaborative (GPC) network evaluated SCILHS metadata for support of interoperation within the SHRINE network it was deploying with hub at the University of Wisconsin, Madison. GPC identified opportunities for enhancing SCILHS 3.1b metadata content for medications, laboratory results and diagnoses(problems/conditions) to assure comprehensive use of ONC data standards. GPC consulted with Harvard in 2017 to propose an expanded metadata set which UNMC would support on a regular maintenance cycle in order to maintain currency with ONC terminology standards evolution and provide versioning and synchronization of SHRINE networks for interoperation.

UNMC began publishing metadata 20180330 and supports FULL implementations consisting of metadata build procedures employing MS SQL and ONC released dataset, comma delimited CSV metadata files which can build all version-specific tables. First release is also accompanied by a DELTA dataset which consists of TXT files of SQL commands that may be executed to convert an existing deployment of (previous) version to the new release version. The first DELTA to be published 20180330 consisted of version changes to move from SCILHS(ARCH) 3.1b to SNOW\_SHRINE\_20180330.

|  |  |  |
| --- | --- | --- |
| **Publication date** | **Version** | **Content** |
| 3/30/2018 | 20180330 | SCILHS 3.1b content AND enhancements which include added content:   1. MEDICATIONS includes historically comprehensive RXNORM (IN/SCD/GPCK/SBD/BPCK) and NDC codes indexed by ingredient 2. LABORATORY with expanded LOINC 2.61 lab metadata covering 2484 tests in lab, anatomic and molecular pathology |
| 12/7/2018 | 20181001 | New content development:   1. ENCOUNTER DIAGNOSES: ICD-10-CM updated with additional 1485 concepts from CMS to 20181001; 30 code revisions this release 2. PROBLEM LIST: SNOMED CT conditions added with US extension 20180930 3. LABORATORY expanded to include 218 more molecular pathology and microbiology 4. MEDICATIONS updated with 1461 new RxNorm and 9781 NDC from NLM release files 20181001 |
| 9/30/2020 | ONC versions inclusive of 20200801 | New content development:   1. SNOMED CT International core 20200701, US Extension(NLM) 20200301 and Nebraska Lexicon© 20200918 2. MEDICATIONS RXNORM(NLM) 20200801 supplemented by NDC history files maintained by Nebraska 3. Lab LOINC© (Regenstrief) v2.68 with LOINC parts mapping to SNOMED CT and Observables ontology build supporting LOINC-on-OWL ontology supports i2b2 LOINC metadata 4. ICD-10-CM (NCHS\CDC) 2021 updates |
| 4/29/2021 | ONC versions inclusive of 20210301 | New content development:   1. SNOMED CT International core 20210101, US extension (NLM) 20210301 and Nebraska Lexicon extension 20210301 2. MEDICATIONS: RXNORM(NLM) 20210301 supplemented by NDC history files maintained by Nebraska 3. Lab LOINC(Regenstrief) v2.69 with LOINC Parts mapping and support for LOINC-on-OWL 4. ICD-10-CM(NCHS/CDC) 2021 updates |
| 6/23/2021 | 20210623 | Added metadata for ICD-O-3 morphology/behavior and topology as well as SNOMED CT anatomy to support TUMOR table browsing and cohort selection. |

# Description of Release Datasets and Instructions for use

## Nebraska Lexicon metadata (ontology) and transitive closure files

Each metadata set, commonly called an ontology within the i2b2 community, consists of two data tables which must be deployed within i2b2 for the metadata to appear within the i2b2 client. The Metadata file in csv format consists of the hierarchical data view of the related terminology useful for the i2b2 client. These csv files support both PATH- and TRANSITIVE\_CLOSURE-searching in i2b2 depending upon the metadata root reference installed in the TABLE\_ACCESS table of the ONTOLOGY cell.

The Transitive Closure tables (TRC) consist in csv files which summarize all of the hierarchical relationships in use in the terminology and are required for TRANSITIVE\_CLOSURE searching. They may also be employed as supplemental tables for concept aggregation searches for their respective terminologies. A transitive closure table includes, for each concept in the ontology, identified in the table as a SUPERTYPE\_CONCEPT, all SUBTYPE\_CONCEPTS( data column CONCEPT\_CD) that exists in the source ontology. For purposes of aggregation queries, TRC tables published by Nebraska also include the SUPERTYPE\_CONCEPT itself as a ‘SUBTYPE\_CONCEPT’.

The metadata table must be copied to the database which hosts the I2B2 ONTOLOGY cell and the TRC tables must be placed in the database that hosts the I2B2 CRC cell. TABLE\_ACCESS in the ONTOLOGY cell must be updated to employ the i2b2 search algorithms of interest to the site manager.

Laboratory metadata(ontology) includes a third table which is the METADATAXML\_SNAPSHOT used with each reload of LOINC laboratory. Metadataxml supports query by value in the i2b2 client and should be loaded ino the ONTOLOGY database.

Included in the release files is TABLES\_20210430.zip. See section 6.1 for instructions on creating this zip file from the set of smaller files named 20210430.zip.001 to 012 that it was split into due to GITHUB file size restrictions. This zip file contains a folder named TABLES\_20210430 which itself contains the metadata and transitive closure files which are described in this section.

### Download the metadata from the GITHUB repository.

The March 2021 metadata release files are in the following GITHUB repository: https://github.com/i2b2plugins/ontology-wg. See section 6.1 in the appendix for instructions on downloading the contents to your computer. See section 6.2 in the appendix for instructions on creating the zip file TABLES\_20210430.zip from the files obtained by following section 6.1 instructions.

### Metadata files in TABLES\_20210430.zip

NOTE: the acronym TRC is used in the Description section to indicate a file with Transitive Closure definitions.

|  |  |  |
| --- | --- | --- |
| ***Terminology*** | ***Filename*** | ***Description*** |
| ***SNOMED CT*** *conditions* | *SNOMEDCT\_METADATA.csv* | *Metadata* |
|  | *SNOMEDCT\_TRC.csv* | *TRC* |
|  |  |  |
| ***LOINC*** *laboratory* | *LABS\_METADATA.csv* | *Metadata* |
|  | *LOINC\_TRC.csv* | *TRC* |
|  | *METADATAXML\_SNAPSHOT.csv* | *C\_METADATAXML definitions for LOINC metadata* |
|  |  |  |
| ***RXNORM/NDC*** *medications* | *MEDICATION\_METADATA.csv* | *Metadata* |
|  | *RXNORM\_NDC\_TRC.csv* | *TRC* |
|  | *MEDICATION\_META\_ROOTS.csv* | *Needed for metadata build procedure* |
|  |  |  |
| ***ICD-10-CM*** | *ICD10CM\_METADATA.csv* | *Metadata* |
|  | *ICD10CM\_TRC.csv* | *TRC* |
| ***ICD-O-3***  ***SNOMED CT Body structures*** | *ANATOMY\_METADATA.csv*  *SNOMEDCT\_TRC.csv* | *Metadata*  *ICD-O-3 + SNOMEDCT TRC table* |

### SNOMED CT Conditions metadata files

SNOMEDCT\_METADATA.csv and SNOMEDCT\_TRC.csv

### LOINC laboratory metadata

LABS\_METADATA.csv, LOINC\_TRC.csv and METADATAXML\_SNAPSHOT.csv.

### RXNORM/NDC medication metadata

MEDICATION\_METADATA.csv, RXNORM\_NDC\_TRC.csv and MEDICATION\_META\_ROOTS.csv(table required for the metadata build); MEDICATION\_META\_ROOTS.csv is a small dataset needed to BUILD MEDICATION\_METADATA.

### ICD-10-CM

ICD10CM\_METADATA.csv and ICD10CM\_TRC.csv.

### ICD-O-3 + SNOMED CT Body structures

ANATOMY\_METADATA.csv; SNOMEDCT\_TRC.csv

## Metadata build procedures

METADATA\_BUILD\_N\_COUNT.7z contains the zipped SQL files for each of the four metadata build routines and FACT\_COUNT.sql which is a procedure built by Jay Pedersen which will fact count i2b2 metadata tables once OBSERVATION\_FACT is populated, performing fact counting which is accurate based upon BOTH the CONCEPT\_CD and MODIFIER\_CD deployed for a fact instance. It requires MODIFIER\_DIMENSION has been built and is current.

# Deployment Instructions

## About I2b2 metadata set design

Metadata tables, sometimes referenced by users as ‘ontologies’, support i2b2 aggregate query functionality in the user client and across SHRINE networks. They do this by organizing code references stored in OBSERVATION\_FACT.CONCEPT\_CD in hierarchical arrays in the metadata tables where the code reference is contained in the attribute C\_BASECODE. The hierarchical structure viewable in the i2b2 client is configured in the attribute C\_FULLNAME which specifies all the hierarchical paths which may lead to the concept of interest. Steps in the path which may be queried as ‘parent’ concepts – supporting aggregation queries - are called ‘folders’ in the client and are the ontological supertypes of all nodes descendant in the ontology.

Aggregation (Folder) searching in the client is supported by two different algorithms. Metadata authored at Harvard employs the classical ‘path-based’ searching using the LIKE operator and comparing the C\_FULLNAME path for each metadata row to the C\_FULLNAME of the folder parent. This is expedient for monohierarchical ontologies but becomes inefficient to poly-hierarchical schemes. ‘Transitive Closure searching’ supporting aggregation requires an additional Transitive Closure table to be installed in the CRC cell database which basically has one row for each SUPERTYPE-SUBTYPE concept pair in the ontology. TRC searching uses the ‘=’ operator and for each concept designated as folder in the metadata queries for all SUBTYPE concepts in OBSERVATION\_FACT. TRC searching supports polyhierarchies and has been demonstrated to execute more efficiently for complex ontologies.

Metadata are deployed in the i2b2 metadata schema as a set of relational tables, one for each data type stored in i2b2, such as demographics, diagnoses, medications, etc. Each metadata row in the table specifies one path to a particular C\_BASECODE concept reference and includes C\_METADATAXML which supports query functionality for the concept as well as details of aggregate searching in C\_COLUMNNAME, C\_OPERATOR and C\_DIMCODE.

The metadata deployment for a datatype is enabled when a row is created in TABLE\_ACCESS specifying the table name in C\_TABLE\_NAME and the root of the ontology in C\_FULLNAME. Installing an i2b2 metadata schema therefore, consists of creating a metadata table for each datatype to be supported, adding the table name and metadata details to TABLE\_ACCESS and then running a utility to create the CONCEPT\_DIMENSION table which supports client searching. Metadata tables may specify the use of certain code sets for indexing OBSERVATION\_FACT.CONCEPT\_CD. The code sets must be identified as coding systems in a SCHEMES metadata table.

Metadata tables may also contain rows which define modifiers for the datatype. Modifiers are an additional i2b2 search tool which support query of subsets of concepts, effectively enforcing a data model in i2b2. Modifier paths are specified in C\_BASECODE and are applied to hierarchies in CONCEPT\_DIMENSION as specified in the attribute M\_APPLIED\_PATH. A table row with M\_APPLIED\_PATH not equal to ‘@’ indicates that that row is a modifier definition. Once the metadata table is installed, modifiers are deployed in the schema by running a utility which creates the table LOAD\_MODIFIER\_DIM.

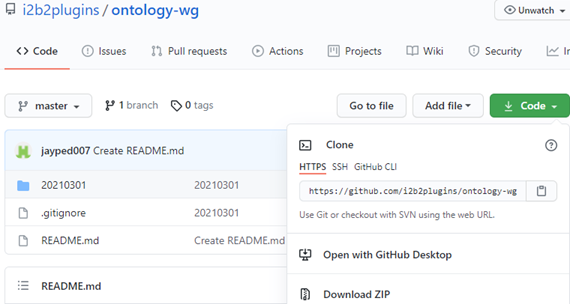
Sites which are participating in SHRINE networks have an additional requirement for installing adaptor mappings, which support mapping a network query to the specific paths and structures of the site metadata.

# APPENDIX

## Downloading the metadata from the GITHUB repository.

The GITHUB repository containing the metadata is located at <https://github.com/i2b2plugins/ontology-wg>.

After navigating to the GITHUB repository, select the green “Code” button and select the “Download ZIP” option that appears in the context menu. This will download the file ontology-wg-master.zip to your computer.



Extract the contents of that zip file into a folder and that folder will have the same structure that is seen on GITHUB. There will be a “20210623” subfolder which holds the metadata release files for March 2021.

## How to create TABLES\_20210623.zip from 20210623.zip.001 thru .015

Within the 20210623 folder are a set of files which can be used to create the zip file TABLES\_20210623.zip which holds the June 2021 release of metadata files. These files are named 20210623.zip.001, 20210623.zip.002, etc.

The WinZip program can be used to combine these smaller files into the single zip file TABLES\_20210623.zip. Winzip is available for Windows and Macintosh operating systems at https://www.winzip.com/. Highlight the zip files in a file explorer window and use the context menu (right-click button on windows) and select the “7-zip -> Extract Here” option. This will create the file TABLES\_20210623.zip in the same folder. That zip file can then be expanded with Winzip or other ZIP file program to create the TABLES\_20210623 folder with the metadata tables and TRC files discussed in section 4.1.1.

## Why TABLES\_20210623.zip was converted into multiple files.

The TABLES\_20210623.zip file could not be stored ‘as is’ within the repository due to GITHUB file size restrictions (100 megabyte limit for any file). Another mechanism was needed.

The Winzip program allows splitting an existing large file into multiple smaller pieces. It is available for Windows and Mac platforms. The following is a screenshot of splitting a large ZIP file into 48 megabyte pieces.

