Ontology Approaches for PCORI

Shawn Murphy MD, Ph.D.
Lori Phillips MS
i2b2/SHRINE AUG
July 11, 2014
Approach – Expressing Data

- I2b2 does not require any specific ontology
  - Queries can be created using any system of local codes
  - Allows i2b2 to attach data of nearly any kind to patients, visits, and providers
Represents Hierarchies

**SNOMED Clinical Terms**

- **CORE Subset of SNOMED CT**
- **SNOMED Anatomy**
- **SNOMED Clinical Findings**
  - **Description:** The Clinical Finding subtree of SNOMED CT
  - **Definition:** Class subtree of ClinicalFinding
  - **Ontology ID:** 2018
  - **Definition Language:** Manual

<table>
<thead>
<tr>
<th>Version</th>
<th>Base Version</th>
<th>Created</th>
<th>Created By</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.2</td>
<td>2009_07_31</td>
<td>07/16/2010</td>
<td>Tania Tudorache, <a href="mailto:tudorache@stanford.edu">tudorache@stanford.edu</a></td>
</tr>
<tr>
<td>1.1</td>
<td>2009_07_31</td>
<td>03/23/2010</td>
<td>Tania Tudorache, <a href="mailto:tudorache@stanford.edu">tudorache@stanford.edu</a></td>
</tr>
<tr>
<td>1.0</td>
<td>2009_01_31</td>
<td>09/09/2009</td>
<td>Tania Tudorache, <a href="mailto:tudorache@stanford.edu">tudorache@stanford.edu</a></td>
</tr>
</tbody>
</table>

- **SNOMED Ethnic Group**
The flexibility of i2b2 becomes a weakness when constructing data networks

- There is no guarantee that people will choose the same ontologies at every site.
- These can be roughly approximated using mappings
SHRINE Mappings
Scope of SHRINE Mappings
• Full data representations in i2b2, often needing modifiers and values, are labeled “Information Models”
• These can be managed in i2b2 with relatively minor impact on the local observation codes
• However, the modifiers and values of the basic concepts need to be represented in similar ways to allow simple mappings to be effective.
• Also, for performance, it is useful to represent concept hierarchies in similar ways in the mappings
Represent scores as concepts
Represent scores as values
Scope of solution to allow information model equivalencies
Arrive at consensus through distributed ontologies

- **Approach**
  - Bring together existing ontologies of the group to understand the best ontology representation of a new information model
  - Develop a consensus ontology and attach to local instance
  - Raw codes in observation-fact table are mapped to “Consensus” ontology
  - SHRINE mapping capabilities used for “finishing touches”
New Information Model Ontology

Consensus Ontology can live alongside other ontologies

(For example: PCORNet CDM ontology and the i2b2 demo ontology in this case)
Supporting Full PCORI SCILHS workflow in i2b2

- To support the PCORNet CDM Information Model, we are distributing full SCILHS ontologies that are to be attached to the local SCILHS instance of i2b2.
- The SCILHS ontologies can accommodate the information model with simple mappings.
- The SCILHS ontologies will support data transformations to the PCORNet CDM.
Contributors

- Ontology Contributors
  - Jeff Klann
  - Dan Connolly
  - Lori Phillips
  - Nathan Wilson

- Release
  - [https://www.i2b2.org/webclient/](https://www.i2b2.org/webclient/)
  - Change the username to pcori
  - Public release early fall
PCORI-centric Use Case

A requirement in PCORI is that all i2b2 instances have to map their ontologies to a central PCORI ontology.

- Local codes need to be integrated into the PCORI hierarchy
- Tools and strategies to assist with this effort
Constructing queries from metadata

```
select patient_num from observation_fact where [c_facttablecolumnname] IN (select [c_facttablecolumnname] from [c_tablename] where [c_columnname] [c_operator] [c_dimcode])
```

METADATA

<table>
<thead>
<tr>
<th>Column</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>C_HLEVEL</td>
<td>INT NULL</td>
</tr>
<tr>
<td>C_FULLNAME</td>
<td>VARCHAR(900) NULL</td>
</tr>
<tr>
<td>C_NAME</td>
<td>VARCHAR(2000) NULL</td>
</tr>
<tr>
<td>C_SYNONYM_CD</td>
<td>CHAR(1) NULL</td>
</tr>
<tr>
<td>C_VISUALATTRIBUTES</td>
<td>CHAR(3) NULL</td>
</tr>
<tr>
<td>C_TOTALNUM</td>
<td>INT NULL</td>
</tr>
<tr>
<td>C_BASECODE</td>
<td>VARCHAR(450) NULL</td>
</tr>
<tr>
<td>C_METADATAXML</td>
<td>TEXT NULL</td>
</tr>
<tr>
<td>C_FACTTABLECOLUMN</td>
<td>VARCHAR(50) NULL</td>
</tr>
<tr>
<td>C_TABLENAME</td>
<td>VARCHAR(50) NULL</td>
</tr>
<tr>
<td>C_COLUMNNAME</td>
<td>VARCHAR(50) NULL</td>
</tr>
<tr>
<td>C_COLUMNDATATYPE</td>
<td>VARCHAR(50) NULL</td>
</tr>
<tr>
<td>C_OPERATOR</td>
<td>VARCHAR(10) NULL</td>
</tr>
<tr>
<td>C_DIMCODE</td>
<td>VARCHAR(900) NULL</td>
</tr>
<tr>
<td>C_COMMENT</td>
<td>TEXT NULL</td>
</tr>
<tr>
<td>C_TOOLTIP</td>
<td>VARCHAR(900) NULL</td>
</tr>
<tr>
<td>UPDATE_DATE</td>
<td>DATETIME NULL</td>
</tr>
<tr>
<td>DOWNLOAD_DATE</td>
<td>DATETIME NULL</td>
</tr>
<tr>
<td>IMPORT_DATE</td>
<td>DATETIME NULL</td>
</tr>
<tr>
<td>SOURCESYSTEM_CD</td>
<td>VARCHAR(50) NULL</td>
</tr>
<tr>
<td>VALUETYPE_CD</td>
<td>VARCHAR(50) NULL</td>
</tr>
</tbody>
</table>
i2b2 Star Schema

patient_dimension
PK Patient_Num
- Birth_Date
- Death_Date
- Vital_Status_CD
- Age_Num*
- Gender_CD*
- Race_CD*
- Ethnicity_CD*

observation_fact
PK Patient_Num
PK Encounter_Num
PK Concept_CD
PK Observer_CD
PK Start_Date
PK Modifier_CD
PK Instance_Num
PK End_Date
PK ValType_CD
PK TVal_Char
PK NVal_Num
PK ValueFlag_CD
PK Observation_Blob

visit_dimension
PK Encounter_Num
PK Start_Date
PK End_Date
PK Active_Status_CD
PK Location_CD*

concept_dimension
PK Concept_Path
PK Concept_CD
PK Name_Char

observer_dimension
PK Observer_Path
PK Observer_CD
PK Name_Char
Visit_dimension types of queries: Encounter types

```
select patient_num
from visit_dimension
where inout_cd IN ('ED', 'E')
```

Edit c_dimcode to match your inout_cds.
A word about missing/unknown values

The PCORnet CDM uses the HL7 conventions of “Null Flavors” (http://hl7.org/implement/standards/fhir/v3/NullFlavor/) as a basis for representing missing or unknown values. Specifically, we will populate these values as follows:

- ‘No Information’ means that the data field is present in the source system, but the source value is null or blank.
- ‘Unknown’ means that the data field is present in the source system, but the source value explicitly denotes an unknown value.
- ‘Other’ means that the data field is present in the source system, but the source value cannot be mapped to the choices presented in the PCORI CDM ontology.
select patient_num from patient_dimension where birth_date BETWEEN sysdate – (365.25*18) AND sysdate – (365.25*10)
select patient_num from patient_dimension where sex_cd IN ('F')

Edit c_dimcode to match your sex_cds.
Patient_dimension types of queries: Race

select patient_num from patient_dimension where race_cd IN ('01','amer. Indian','i','na','nat. am')

Edit c_dimcode to match your race_cds.
Patient_dimension types of queries: Ethnicity

```
select patient_num
from
patient_dimension
where ethnicity_cd IN ('HISPANIC')
```

Edit c_dimcode to match your ethnicity_cds.
select concept_cd from concept_dimension where concept_path like '\PCORI\DEMOGRAPHIC\BIOBANK_FLAG\Y\%'

Edit c_basecode to match your code for biobank specimens.
Broader concept_dimension based queries

Diagnoses, Procedures typically require a merge of local terms within the tree.
ICD-9 Example

- If your institution uses ICD-9 …
  - Change the c_basecode to match your code format.

<table>
<thead>
<tr>
<th>PCORI_BASECODE</th>
<th>C_BASECODE</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICD9:250.1</td>
<td>MY_ICD9:250.1</td>
</tr>
</tbody>
</table>

update pcornet
set c_basecode = replace(pcori_basecode, 'ICD9:', 'MY_ICD9:')
where c_basecode is not null
and pcori_basecode like 'ICD9:%'
and c_fullname like '\PCORI\DIAGNOSIS\09\%';
If your institution uses ICD-9 but in a plain_code format...

- Change the `c_basecode` to match your code format.

```
PCORI_BASECODE   C_BASECODE
ICD9:250.1       2501
```

```sql
update pcornet
set c_basecode = substring(pcori_basecode, 6, 25)
where pcori_basecode like 'ICD9:%'
and c_fullname like '\PCORI\DIAGNOSIS\09\%';
```

```sql
update pcornet
set c_basecode = replace(c_basecode, '.', '')
where pcori_basecode like 'ICD9:%'
and c_fullname like '\PCORI\DIAGNOSIS\09\%';
```
Local diagnoses codes (non-ICD9)

- Local codes need to be mapped/merged.
- Mapper cell / mapping tool plugins exist for workbench.
<table>
<thead>
<tr>
<th>Source Field</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>SOURCE_CODING_SYSTEM</td>
<td>VARCHAR(50)</td>
</tr>
<tr>
<td>SOURCE_BASECODE</td>
<td>VARCHAR(50)</td>
</tr>
<tr>
<td>SOURCE_NAME</td>
<td>VARCHAR(2000)</td>
</tr>
<tr>
<td>SOURCE_FULLNAME</td>
<td>VARCHAR(700)</td>
</tr>
<tr>
<td>SOURCE_TOOLTIP</td>
<td>VARCHAR(700)</td>
</tr>
<tr>
<td>SOURCE_TABLE_CD</td>
<td>VARCHAR(25)</td>
</tr>
<tr>
<td>SOURCE_KEY</td>
<td>VARCHAR(725)</td>
</tr>
<tr>
<td>DESTINATION_CODING_SYSTEM</td>
<td>VARCHAR(50)</td>
</tr>
<tr>
<td>DESTINATION_BASECODE</td>
<td>VARCHAR(50)</td>
</tr>
<tr>
<td>DESTINATION_NAME</td>
<td>VARCHAR(2000)</td>
</tr>
<tr>
<td>DESTINATION_FULLNAME</td>
<td>VARCHAR(700)</td>
</tr>
<tr>
<td>DESTINATION_TABLE_CD</td>
<td>VARCHAR(25)</td>
</tr>
<tr>
<td>DESTINATION_KEY</td>
<td>VARCHAR(725)</td>
</tr>
<tr>
<td>MAPPING_SOURCE</td>
<td>VARCHAR(50)</td>
</tr>
<tr>
<td>VARFLAG</td>
<td>VARCHAR(25)</td>
</tr>
<tr>
<td>FLAG</td>
<td>INT</td>
</tr>
<tr>
<td>STATUS_CD</td>
<td>VARCHAR(25)</td>
</tr>
<tr>
<td>UPDATE_DATE</td>
<td>DATETIME</td>
</tr>
<tr>
<td>C_TOTALNUM</td>
<td>INT</td>
</tr>
</tbody>
</table>

Source fields are local terms.
Destination fields are PCORI terms.
Automated mappings

- If your local codes already exist within an ICD9 hierarchy..
- Find the c_fullname, c_basecode and c_name of parent of the term and use that as your mapping’s destination.
Automated mappings, cont.

WITH joined as(
  select r2.c_basecode as sourceCode, r1.c_fullname as destination, r1.c_basecode as destCode, r1.c_name as destName from local_ont r1 inner join local_ont r2 on r1.c_fullname = r2.c_path where r2.c_basecode in (select source_basecode from project_ont_mapping where destination_fullname is null))

update project_ont_mapping
set destination_fullname = joined.destination,
  destination_basecode = joined.destCode,
  destination_name = joined.destName
from joined
where joined.sourceCode = source_basecode
Integration tool

Select path symbol format (S,M,L)
\texttt{i2b2\textbackslash Diagnoses\textbackslash Metabolic and immunity disorders (270-279)\textbackslash (277) Other and unspecified disorder\textbackslash (277-0) Cystic fibrosis}\n
Start integration
Refresh process status until integration is complete.
Export tool

- Exports result of integration process as a single, delimited i2b2 metadata file.
Alternate map/merge method

- If you have small number of local codes to merge, use the Edit Terms view.
  - User must be set up with EDITOR role
Edit Terms View
New term added

No need to export file, term is added to table automatically.
<entry>
  <key>\SHRINE\SHRINE\Diagnoses\Endocrine, nutritional, and metabolic diseases and immunity disorders\Cystic fibrosis</key>
  <value>
    <local_key>\i2b2\Diagnoses\Metabolic and immunity disorders (270-279)(277) Other and unspecified disor~\(277-0) Cystic fibrosis</local_key>
  </value>
</entry>
<entry>
  <key>\SHRINE\SHRINE\Demographics\Gender\Female</key>
  <value>
    <local_key>\i2b2\Demographics\Gender\Female</local_key>
  </value>
</entry>
Meaning behind <key> 

- Concept key == \c_table_cd\c_fullname\ 

<concept> 
   <level>1</level> 
   <key>\i2b2_DIAG\i2b2\Diagnoses\</key> 
   <name>Diagnoses</name> 
   ... 
</concept> 

TABLE_ACCESS 

<table>
<thead>
<tr>
<th>C_TABLE_CD</th>
<th>C_TABLE_NAME</th>
<th>C_FULLNAME</th>
<th>C_NAME</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>BIRN</td>
<td>\BIRN\</td>
<td>Clinical Trials</td>
</tr>
<tr>
<td>2</td>
<td>CUST</td>
<td>\CUSTOM_META\</td>
<td>Custom Metadata</td>
</tr>
<tr>
<td>3</td>
<td>i2b2_DEMO</td>
<td>\i2b2\Demographics\</td>
<td>Demographics</td>
</tr>
<tr>
<td>4</td>
<td>i2b2_DIAG</td>
<td>\i2b2\Diagnoses\</td>
<td>Diagnoses</td>
</tr>
<tr>
<td>5</td>
<td>i2b2_EXPR</td>
<td>\i2b2\Expression Profiles Data\</td>
<td>Expression Profiles Data</td>
</tr>
<tr>
<td>6</td>
<td>i2b2_LABS</td>
<td>\i2b2\Labtests\</td>
<td>Laboratory Tests</td>
</tr>
<tr>
<td>7</td>
<td>i2b2_MEDS</td>
<td>\i2b2\Medications\</td>
<td>Medications</td>
</tr>
<tr>
<td>8</td>
<td>i2b2_PROC</td>
<td>\i2b2\Procedures\</td>
<td>Procedures</td>
</tr>
<tr>
<td>9</td>
<td>i2b2_PROV</td>
<td>\i2b2\Providers\</td>
<td>Providers</td>
</tr>
<tr>
<td>10</td>
<td>i2b2_REP</td>
<td>\i2b2\Reports\</td>
<td>Reports</td>
</tr>
<tr>
<td>11</td>
<td>i2b2_VISIT</td>
<td>\i2b2\Visit Details\</td>
<td>Visit Details</td>
</tr>
<tr>
<td>12</td>
<td>ICD10_ICD9</td>
<td>\Diagnoses\</td>
<td>Diagnoses (ICD10)</td>
</tr>
</tbody>
</table>
SHRINE Adapter Mapping Tool

Version of Mapping Tool for editing Adapter Mapping File.
Export new file
entry>
   <key>\SHRINE\SHRINE\Diagnoses\Endocrine, nutritional, and metabolic diseases and immunity disorders\Cystic fibrosis</key>
   <value>
      <local_key>\i2b2\i2b2_DIAG\Diagnoses\Metabolic and immunity disorders (270-279)\(277) Other and unspecified disor~\(277-0) Cystic fibrosis</local_key>
   </value>
</entry>
<entry>
<entry>
<entry>
   <key>\SHRINE\SHRINE\Demographics\Gender\Female</key>
   <value>
   <local_key>\i2b2\i2b2_DEMO\Demographics\Gender\ Female</local_key>
   </value>
</entry>
Where can I find these tools?

- Edit terms view is distributed in standard release of i2b2.
- Mapping tools will be available soon on wiki:
  
  https://community.i2b2.org/wiki/display/NCBO/Mapping+tools+version+1.0
Fields used to construct queries

Select * from observation_fact where c_facttablecolumn in (select c_facttablecolumn from c_tablename where c_columnname c_operator ‘c_dimcode %’)

Select * from observation_fact where concept_cd in (select concept_cd from concept_dimension where concept_path like ‘\RPDR\Demographics\Gender\Female\%’)

<table>
<thead>
<tr>
<th>c_facttablecolumn</th>
<th>c_tablename</th>
<th>c_columnname</th>
<th>c_columnndatatype</th>
<th>c_operator</th>
<th>c_dimcode</th>
</tr>
</thead>
<tbody>
<tr>
<td>concept_cd</td>
<td>concept_dimension</td>
<td>concept_path</td>
<td>T</td>
<td>LIKE</td>
<td>\RPDR\Demographics\Gender</td>
</tr>
<tr>
<td>concept_cd</td>
<td>concept_dimension</td>
<td>concept_path</td>
<td>T</td>
<td>LIKE</td>
<td>\RPDR\Demographics\Gender\Female</td>
</tr>
<tr>
<td>concept_cd</td>
<td>concept_dimension</td>
<td>concept_path</td>
<td>T</td>
<td>LIKE</td>
<td>\RPDR\Demographics\Gender\Male</td>
</tr>
<tr>
<td>concept_cd</td>
<td>concept_dimension</td>
<td>concept_path</td>
<td>T</td>
<td>LIKE</td>
<td>\RPDR\Demographics\Gender\Unknown</td>
</tr>
<tr>
<td>concept_cd</td>
<td>concept_dimension</td>
<td>concept_path</td>
<td>T</td>
<td>LIKE</td>
<td>\RPDR\Demographics\Gender\Unknown\Unknown-Female</td>
</tr>
<tr>
<td>concept_cd</td>
<td>concept_dimension</td>
<td>concept_path</td>
<td>T</td>
<td>LIKE</td>
<td>\RPDR\Demographics\Gender\Unknown\Unknown-Male</td>
</tr>
<tr>
<td>concept_cd</td>
<td>concept_dimension</td>
<td>concept_path</td>
<td>T</td>
<td>LIKE</td>
<td>\RPDR\Demographics\Gender\Unknown\Unknown-U</td>
</tr>
</tbody>
</table>


Import tool

- Intended for small sets of local data
- Enter data into staging area or into table directly.
- Upload to mapping table when complete.
- Entries will appear in unmapped terms table.
patient_dim v concept_dim approaches

- HISTORY: early versions of i2b2 did not support the patient_dimension approach.
  - Demographic data was inserted as entry in observation_fact table and as a result appeared as a tick mark on the timeline.

- Patient_dimension approach does not require an entry in the observation_fact table for that concept.

- Patient_dimension approach will not result in a tick mark in the timeline for that concept.

- Patient_dimension approach is easier to maintain.
Ethnicity, two choices: concept_dimension query

```
select concept_cd from concept_dimension
where concept_path like '%PCORI\DEMOGRAPHIC\HISPANIC\Y\%'
```

Edit c_basecode to match your ethnicity codes.
Ethnicity, second choice: patient_dimension query

1. Add Ethnicity_cd column to patient_dimension table

2. Edit entry to a patient_dimension query; modify c_dimcode to match your ethnicity codes.

```sql
select patient_num from patient_dimension
where ethnicity_cd IN (‘HISPANIC’)
```

(select [c_facttablecolumnname] from [c_tablename] where [c_columnname] [c_operator] [c_dimcode])