Developing Subsets and Reference Sets

Expo 2019 Tutorial

David Markwell and Anne Randorff Højen
SNOMED International

Overview

- Part 1. Introduction to subsets and reference sets
  - Subsets, value sets and reference sets
  - Meeting subset requirements with reference sets
  - Enhancing subsets with reference sets

- Part 2. Subset creation principles
  - Purpose
  - Principles
    - Context
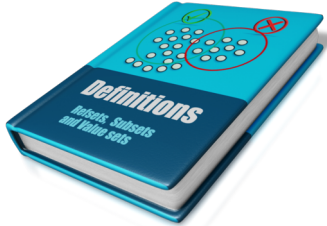   - Consistency
    - Check-it-again
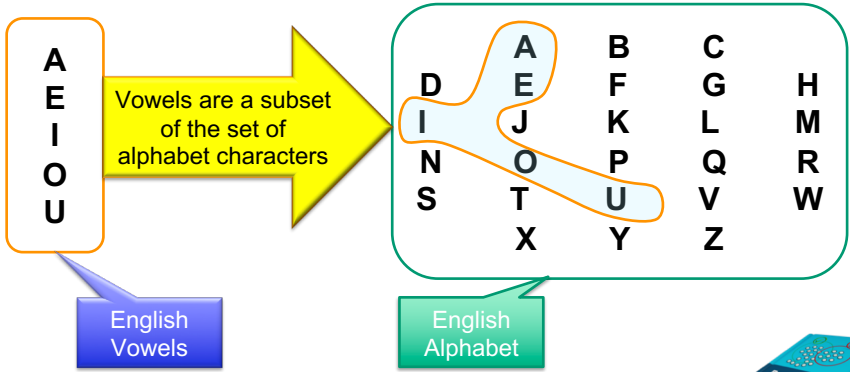 - Demo

What are …
‣ **Subsets?**
‣ **Value Sets?**
‣ **Reference Sets?**

How are they connected?

What is a Subset?

- A set whose members are all contained in another set
  - This is a general definition not specific to SNOMED CT
- Example

English Vowels

A E I O U

Vowels are a subset of the set of alphabet characters

English Alphabet

A B C
D E F G H
I J K L M
N O P Q R
S T U V W
X Y Z

What is a Value Set?

- A set of concept representations used to represent values in a particular data item
  - This definition is not specific to SNOMED CT
  - May include codes from different code systems, coded refinements or predefined text strings
- Example
  - A specification for a problem list might define a single value-set including:
    - SNOMED CT disorder concepts
    - SNOMED CT expressions that are subtypes of disorder
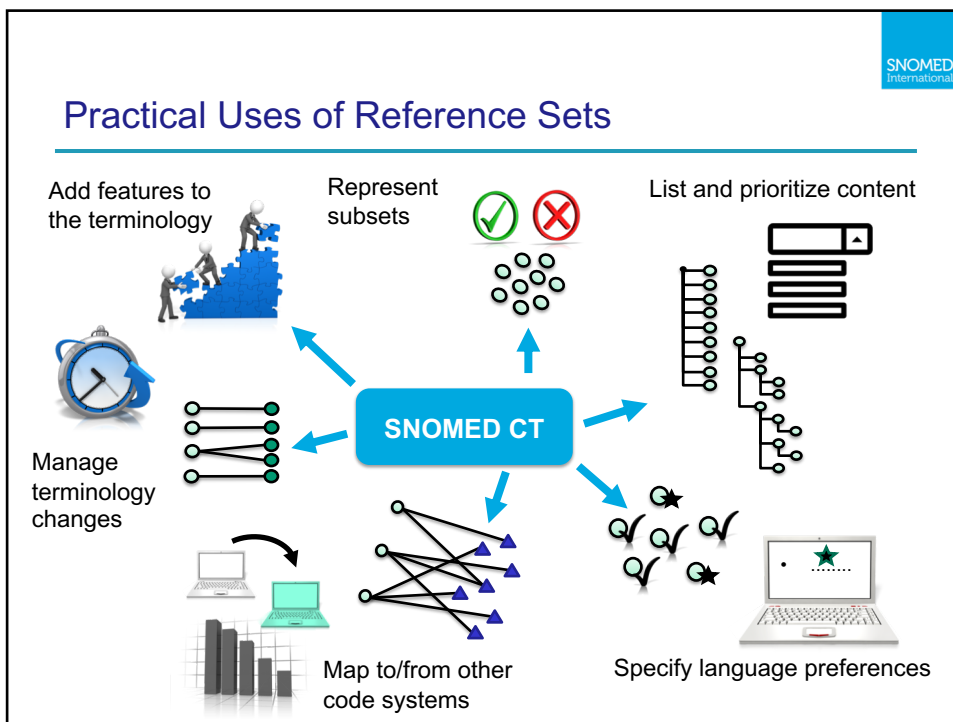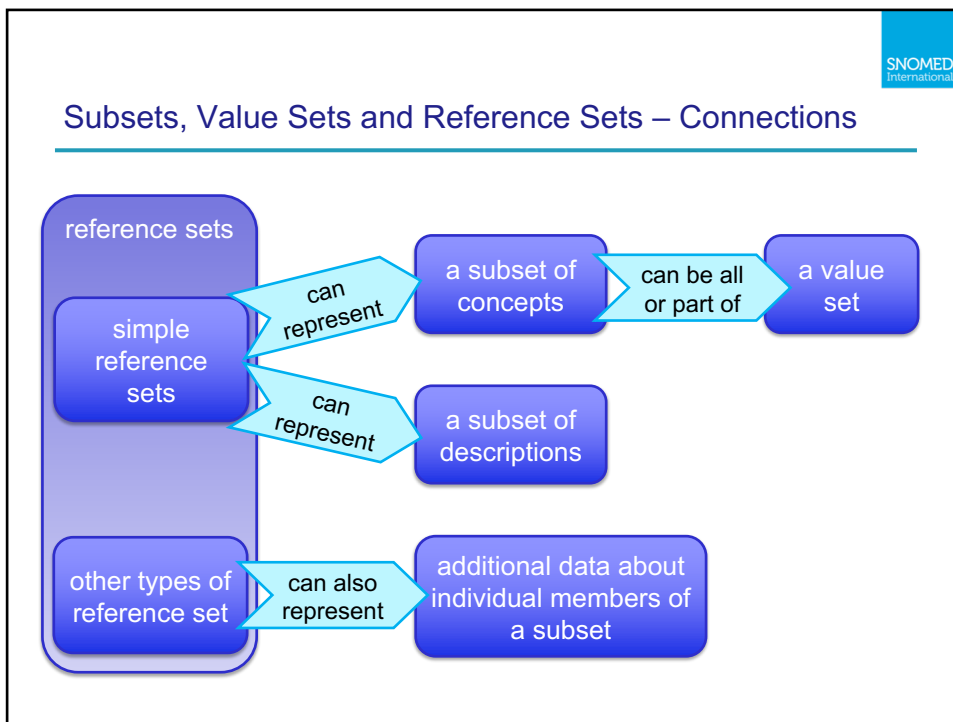   - ICD-10 classification codes representing diseases

What is a Reference Set?

- A standard SNOMED CT release file format
- A set of references to SNOMED CT components
- A way to add optional data to referenced components
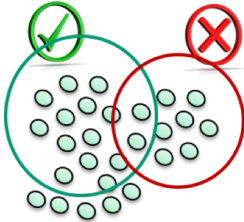
id	Identification and versioning and modularization information
effectiveTime	
active	
moduleId	
refsetId	**An identifier of the reference set**
referencedComponentId	**An identifier of a referenced component**
<attribute-1 … attribute-n>	Optional additional data items to meet specific requirements

Subsets, Value Sets and Reference Sets – Connections

reference sets

simple reference sets

can represent → a subset of concepts → can be all or part of → a value set

can represent → a subset of descriptions

other types of reference set → can also represent → additional data about individual members of a subset

Practical Uses of Reference Sets

Add features to the terminology

Represent subsets

List and prioritize content

Manage terminology changes

SNOMED CT

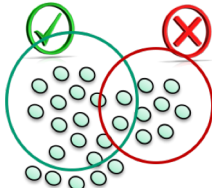Map to/from other code systems

Specify language preferences

Meeting Subset Requirements with Reference Sets

Practical Requirements for Subsets

- Searches and data entry
    - Restricting searches to a set of concepts or descriptions
    - Specifying descriptions to appear in a list of options
    - Constraining data entry to a specified set of concepts
- Information model and communications
    - Specifying value sets for particular data items
- Data retrieval and analysis
    - Specifying query criteria
- Other uses
    - Subsets can be used for any purpose that requires selective inclusion or exclusion of specified sets of components

Requirements for Representing Subsets

- **A simple list of identifiers can represent a subset of SNOMED CT components**
  - For example:
    - 82272006
    - 6142004
    - 55604004
    - … etc. …

In a SNOMED CT reference set, the list of component identifiers is represented by the
- **referencedComponentId** column

Subset Represented by a Simple Reference Set

Subset

Id	Term
82272006	Common cold
6142004	Influenza
55604004	Avian influenza

Reference Set

id	effective Time	active	moduleId	refsetId	referenced ComponentId
0000d6608041-005b-5108-b608-07891b210365	20140731	1	19999999103	49999999102	82272006
00008774f374-5f54-5234-8af1-8aca6e73b5c2	20140731	1	19999999103	49999999102	6142004
00006ac9-2f97-5ed4-b37e-84cb158c388d	20140731	1	19999999103	49999999102	55604004

Requirements for Representing Subsets

1. A simple list identifiers can represent a subset of SNOMED CT components
2. **For practical use, a subset needs to be identified and named so it can be referred to unambiguously**

When a subset is represented by a reference set:
- The subset is identified by the **refsetId** column
- The **refsetId** refers to a concept that has
  - A description that names the reference set
  - A relationship that refers to the reference set type
    - For example |is a| |simple type reference set|

Subset Represented by a Simple Reference Set

Reference Set

id	effective Time	active	moduleId	refsetId	referenced ComponentId
0000d6608041-005b-5108-b608-07891b210365	20140731	1	19999999103	49999999102	82272006
00008774f374-5f54-5234-8af1-8aca6e73b5c2	20140731	1	19999999103	49999999102	6142004
00006ac9-2f97-5ed4-b37e-84cb158c388d	20140731	1	19999999103	49999999102	55604004

49999999102
Infectious disease simple reference set (EXAMPLE ONLY)

|is a|
subtype of

446609009
Simple type reference set

Requirements for Representing Subsets

1. A simple list identifiers can represent a subset of SNOMED CT components
2. For practical use, a subset needs to be identified and named so it can be referred to unambiguously
3. **Subset membership may need to change with future releases of SNOMED CT or due to evolving requirements for inclusion of different content**

When a subset is represented by a reference set:
- The standard SNOMED CT approach to versioning and modularization allows full tracking of changes.
- This uses the following columns
  - id, effectiveTime, active, moduleId

Subset Membership Changes Represented by a Simple Reference Set

Remove (inactivate)

Changes in July 2016

Add

Id	Term
82272006	Common cold
6142004	Influenza
55604004	Avian influenza
3928002	Zika fever

Reference Set

id	effective Time	active	moduleId	refsetId	referenced ComponentId
0000d6608041-005b-5108-b608-07891b210365	20140731	1	19999999103	49999999102	82272006
00008774f374-5f54-5234-8af1-8aca6e73b5c2	20140731	1	19999999103	49999999102	6142004
00006ac9-2f97-5ed4-b37e-84cb158c388d	20140731	1	19999999103	49999999102	55604004
00006ac9-2f97-5ed4-b37e-84cb158c388d	20160731	0	19999999103	49999999102	55604004
0000c5c4-50bb-5937-9141-1ec3e2d517e8	20160731	1	19999999103	49999999102	3928002

Requirements for Subsets

1. A simple list identifiers can represent a subset of SNOMED CT components
2. For practical use, a subset needs to be identified and named so it can be referred to unambiguously
3. Subset membership may need to change with future releases of SNOMED CT or due to evolving requirements for inclusion of different content
4. **It may be useful to define the membership of a subset using rules rather than a list of identifiers**
   - **This is called an *intensional subset definition***

Not a misspelling !
Inten**s**ional is not the
same as inten**t**ional

Intensional Subset Definitions: Substrate and Expansion

Substrate
- The set to which an intensional subset definition is applied

Intensional definition
- A set of rules that defines whether a member of the substrate is included in the subset

Expansion
- The result of applying an intensional definition to a given substrate

Substrate

Intensional definition

Expansion

Intensional Subset Definitions: Substrate and Expansion (Example)

Substrate
- All the letters in the English alphabet

Intensional definition
- A rule that defines which letters are included in the subset

Expansion
- The result of applying the intensional definition to a given substrate

English Alphabet

Letters between **L** and **S**

M, N, O, P, Q, R

Extensional and Intensional Subset Definitions

- Extensional subset definitions
  - Subset membership defined by enumeration
    - Identifying each of the members individually

SNOMED CT supports extensional subset definitions:
- **Simple type reference sets** with members identified by component identifiers (referencedComponentId)

- Intensional subset definitions
  - Subset membership is defined by a set of rules
    - The rules are expressed as a query that computes the membership of a subset

SNOMED CT supports intensional subset definitions:
- **SNOMED CT expression constraints** can represent rules that determine which concepts are in a subset

Intensional Subset Definitions: Substrate and Expansion (SNOMED CT Example)

Substrate
- A specified version of a SNOMED CT edition

Intensional definition
- A SNOMED CT expression constraint
- For example subtypes of infectious disease

Expansion
- Concepts in the specified version of the edition that comply with the constraint

SNOMED CT International Edition 2019-07-31

< |infectious disease|

Concepts that are subtypes of |infectious disease|

Enhancing Subsets with Reference Sets

Requirements for Ordered Lists

- Requirements for ordered lists of descriptions
    - Presenting terms in an order that is rational or helpful for a particular purpose in user interface controls including:
        - Simple lists
        - Drop down lists
        - Popup menus
- Requirements for ordered lists of concepts
    - Presenting concepts in an order that is rational or helpful for a particular purpose irrespective of the term displayed
    - Examples
    - Body parts that have a natural order
        - Fingers, cranial nerves and vertebrae
    - Enumerated values
        - Mild, moderate, severe

Requirements for Prioritization

- Prioritization is similar to order but multiple components may have the same rank
- Requirements for prioritization of concepts
    - Making it easier to find concepts that are most commonly used in a particular specialty, department or data entry scenario
    - Highlighting concepts that are preferred options for a particular purpose without preventing access to a wider selection of concepts
- Priorities may be implemented in various ways
    - Highlight high priority items
    - Only show prioritized items initially with a display more option

Ordered Component and Reference Sets

Requirement
Sort members of
the subset in a
specified order

Requirement
Organize subset
members into
priority groups

Integer

Ordered
Component

▪ Ordered component reference set has an additional
integer column to specify either an order or a priority
group for members of the subset

Ordered Component Type Reference Set Representing Display Order or Priority

Fingers sorted A-Z	
21356012	**Fifth finger**
136021011	**Fourth finger**
138873019	**Second finger**
108884010	**Third finger**
127053016	Thumb

The **order** column is an integer that specifies:
• Display order
• Priority rank: 1 (first) highest rank

id	effective Time	active	moduleId	refsetId	referenced ComponentId	order
…	20160731	1	19999999103	809999999103	127053016	1
…	20160731	1	19999999103	809999999103	138873019	2
…	20160731	1	19999999103	809999999103	108884010	3
…	20160731	1	19999999103	809999999103	136021011	4
…	20160731	1	19999999103	809999999103	21356012	5

• **Thumb**
• **Second finger**
• **Third finger**
• **Fourth finger**
• **Fifth finger**

Requirements for Adding Information to a Referenced Component
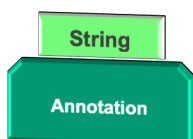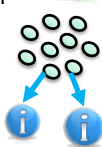
- Displaying a textual note or comment about a listed or selected concept
  - For example, a guidance note on requesting a particular procedure or service
- Marking descriptions with indications of whether the terms they contain are acceptable or preferred in a specified language or dialects
  - For example, distinguishing term usage between different languages, dialects, local or specialty groups
- Marking particular concepts with specific values to provide processable and/or displayable information
  - For example, marking inactive concepts with indicators of the reasons for inactivating them

Annotation and Attribute Value Reference Sets

Requirement
Add text annotation to members of a subset

String

Annotation

Requirement
Add enumerated data to members of a subset

conceptId	term
405738005	blue
371254008	brown
371246006	green
371240000	red

SCTID

Attribute Value

- Annotation reference set has an additional text column to add notes to subset members
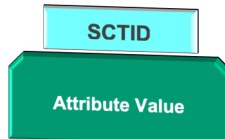
- Attribute value reference set has an additional concept id column to refer to structured data

Annotation Type Reference Set
Adding Unstructured Information to a Concept

Id	Term
82272006	Common cold
6142004	Influenza
55604004	Avian influenza
3928002	Zika fever

The **annotation** column is an string that contains a free text note about the referenced component.

id	effective Time	active	moduleId	refsetId	referenced ComponentId	annotation
…	20160731	1	19999999103	119999999106	82272006	Trivial just treat symptoms.
…	20160731	1	19999999103	119999999106	6142004	Risk in elderly. Advise immunization in future years.
…	20160731	1	19999999103	119999999106	55604004	Report to public health is suspected. Admit if any respiratory distress.
…	20160731	1	19999999103	119999999106	3928002	Indicate if pregnant or potentially pregnant. Note travel in affected area.

Attribute Value Type Reference Set
Adding Structured Data to a Concept

The **valueId** column refers to a concept that represent the additional data applicable to the concept

id	effective Time	active	moduleId	refsetId	referenced ComponentId	valueId
…	20160731	1	900000000000207008	900000000000489007	1720006	900000000000482003
…	20160731	1	900000000000207008	900000000000489007	4515009	900000000000484002
…	20160731	1	900000000000207008	900000000000489007	4961001	900000000000485001
…	20160731	1	900000000000207008	900000000000489007	5381002	900000000000483008

1720006 |Dextroposition of aorta| |Duplicate|
4515009 |Keratoderma punctata| |Ambiguous|
4961001 |Psychiatric function| |Erroneous|
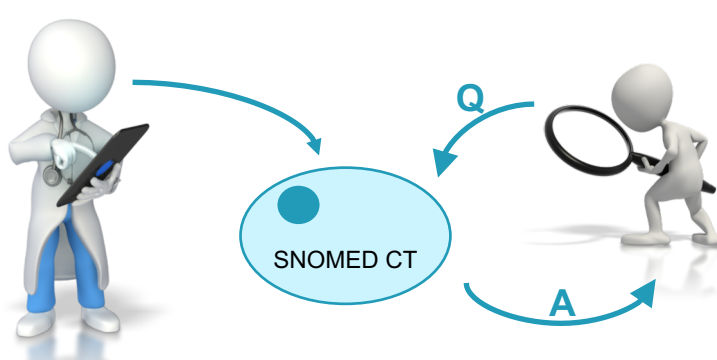5381002 |Enterobacter gergoviae| |Outdated|

Overview

- Part 1. Introduction to subsets and reference sets
  - Subsets, value sets and reference sets
  - Meeting subset requirements with reference sets
  - Enhancing subsets with reference sets
- Part 2. Subset creation principles
  - Purpose
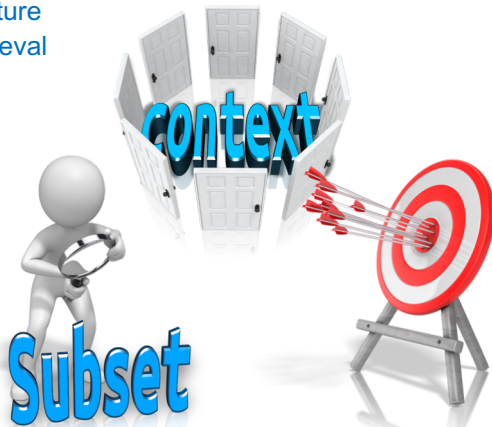 - Principles
    - Context
    - Consistency
    - Check-it-again
  - Demo

Purpose

- Ensure that the patient data captured using your subset, results in good data quality in clinical records
  - Good quality data capture
  - Good quality data retrieval

SNOMED CT

Q

A

Purpose

- Support that the patient data captured using your subset, results in good data quality in clinical records
  - Good quality data capture
  - Good quality data retrieval
- Dependent on
  - Context
  - Consistency
  - Checking

Context Principles
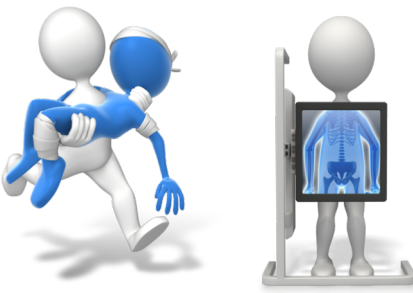
- Ensure that the subset reflects the intended clinical data
- Ensure that the subset reflects the surrounding information model
- Choose a strategy for when no concept is available
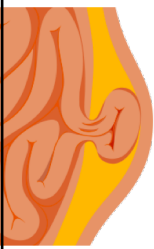
Correct Clinical Content

- Ensure that the subset reflects the intended clinical data
  - Understand the clinical details of the data element in which your subset should work
  - Incorrect assumptions about the clinical workflow of a data entry form can lead to poor quality data
  - Avoid using concepts which are used to group specific clinical meanings

Correct Clinical Content – Example 1

- ☰ Hernia
- ○ Hernia

- ○ Edema
- ☰ Edema

Correct Clinical Content – Example 2

Information Model Aware

- Ensure that the subset reflects the surrounding information model
  - A subset often populates a specific data element in an information model
  - Take into account the surrounding data elements
  - Consider the relationship between interface model and storage model

Information Model Aware - Example

Fracture

Type (Coded text)

Body site (Coded text)

Fracture type subset

Concept Id	Term
439987009	Open fracture of bone (disorder)
359817006	Closed fracture of hip (disorder)
16114001	Fracture of ankle (disorder)

Fracture body site subset

Concept Id	Term
71341001	Bone structure of femur (body structure)
33696004	Bone structure of ankle (body structure)

Information Model Awareness - Example

Fracture

Type (Coded text)

Body site (Coded text)

No finding site

< 125605004 |Fracture of bone |:
[0..0] 363698007 |Finding site| = 272673000|Bone structure|

<272673000 |Bone structure (body structure)|

Fracture type subset

Concept Id	Term
439987009	Open fracture of bone (disorder)
423125000	Closed fracture of bone (disorder)
443395009	Compression fracture (disorder)

Information Model Awareness - Example

Fracture

Type
(Coded text)

Fracture with location subset

Concept Id	Term
71620000	Fracture of femur (disorder)
16114001	Fracture of ankle (disorder)

Strategy for no Concept

- Choose a strategy for when no concept is available
  - Subsets often require clinical concepts not in your edition
  - Options
    - Split the meaning across multiple data elements
    - Allow postcoordinated expressions
    - Create a new extension concept
    - Use a more general concept

Choose up front!

Strategy for no Concept - Example

> Body weight in the morning

1. Split the meaning across multiple data elements
   - Observation: coded text, 27113001 |Body weight|
   - Time frame: coded text, 73775008 |Morning|
2. Allow postcoordinated expressions

27113001 |Body weight|:370134009 |Time aspect |= 73775008 |Morning|

3. Create a new extension concept
   - xxxx |Body weight in morning|
4. Use a more general concept:
   - 27113001 |Body weight|

Consistency Principles

- Be consistent in hierarchy use
- Avoid overlapping content

Consistent Hierarchy

- Be consistent in hierarchy use
- Typically include concepts from only one hierarchy
  - E.g. Cardiology diagnoses and cardiology procedures are both in a cardiology patient journal, but mixing is uncommon
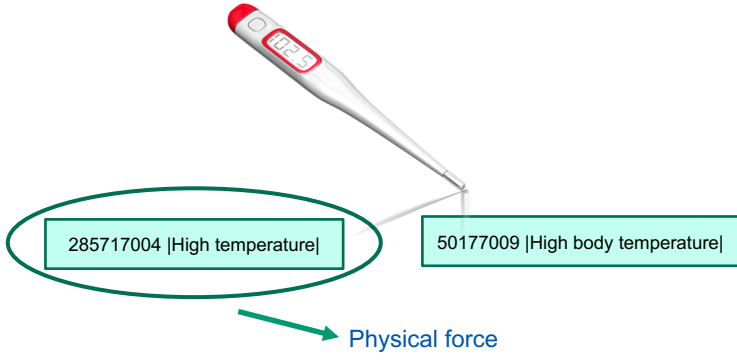- Exceptions do exist
  - E.g. Reason for admission subset

Concept Id	Term
233604007	Pneumonia (disorder)
86591008	Fall from ladder (event)
177141003	Elective cesarean section (procedure)

Consistent Hierarchy - Example

- Similar sounding concepts

285717004 |High temperature| 50177009 |High body temperature|

Physical force

- Not term search only. Consider meaning and hierarchy

Consistent hierarchy - Using ECL

1. Define an intentional subset using ECL
2. Adapt it manually

<<118797008 |Procedure on heart (procedure)|

Concept	Id
~~Positron emission tomography with computed tomography of heart (procedure)~~	~~16554411000119104~~
Positron emission tomography myocardial metabolism study (procedure)	16545811000119100
Single photon emission computed tomography of myocardial perfusion with computed tomography (procedure)	16545381000119101
~~Radionuclide myocardial perfusion stress study (procedure)~~	~~16545101000119100~~
Computed tomography of heart with contrast for cardiac structure and morphology (procedure)	16439801000119108
Magnetic resonance imaging for cardiac morphology and function without contrast (procedure)	16439481000119102
Magnetic resonance imaging for cardiac velocity flow mapping (procedure)	16439401000119105
~~Computed tomography of heart without contrast with calcium scoring (procedure)~~	~~5010841000119100~~
Two dimensional echocardiography of fetus (procedure)	422851000119101
Transcatheter aortic valve implantation (procedure)	773996000

Avoid Overlaps

- Difficult for user to select between overlapping concepts
  - Leads to inconsistent selections by different users
  - This can lead to poor data quality
- Overlapping meaning can be caused by
  - Including both a subtype and supertype concept
  - Including two concepts with a common ancestor that represent two different aspects of the same idea
- Sometimes it may be appropriate to include both a subtype and a supertype
  - Consideration should be given to make it clear to the users which one to chose in different situations

Avoid Overlaps - Example

Medical problems

Concept Id	Term
195967001	Asthma
86049000	Cancer
46635009	Diabetes mellitus type 1
54329005	Acute myocardial infarction of anterior wall
57054005	Acute myocardial infarction

Avoid Overlaps - Example

- Two concepts with a common ancestor represent two different aspects of the same idea

Pain subset

Concept Id	Name
25064002	Headache (finding)
247373008	Ankle pain (finding)
274663001	Acute pain (finding)
55145008	Stabbing pain (finding)

Acute headache

- Only subtypes of pain, but each concept in the subset refers to different characteristics of the pain

Check-it-again principles

- Check inclusions and exclusions
- Avoid careless mistakes
- Use reviewers with multiple perspectives

Inclusions and Exclusions

- Check inclusions and exclusions
  - Check inclusions
    - Logical definitions and descriptions
    - Display terms should be valid synonyms
  - Check exclusions
    - Concepts missing from the subset

Check Definitions - Example

- Check that the concept belongs to the intended hierarchy
- Check explicit rules about the attribute relationships

Pain quality subset

Concept Id	Term
279093005	Cramping pain
8708008	Sharp pain
36349006	Burning pain
55145008	Stabbing pain

- Check expression constraint
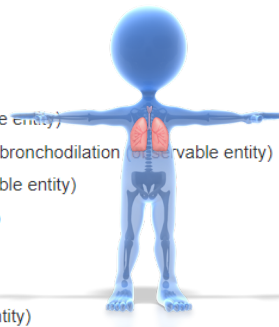 - < |Clinical finding|: |Finding site| =*

Check Descriptions - Example

- Especially for primitive concepts, we rely on the fully specified names
  - Choosing between primitive sibling concepts may be challenging

  <165041004 |Forced expired volume (observable entity)|

  - Forced expired volume in 0.75 seconds (observable entity)
  - Forced expired volume in 1 second (observable entity)
    - Expected forced expired volume in 1 second (observable entity)
    - Forced expired volume in 1 second after bronchodilation (observable entity)
      - Percentage predicted forced expiratory volume in 1 second after bronchodilation (observable entity)
    - Forced expired volume in 1 second before bronchodilation (observable entity)
    - Forced expired volume in 1 second post steroids (observable entity)
    - Forced expired volume in 1 second pre steroids (observable entity)
    - Forced expired volume in 1 second reversibility (observable entity)
  - Percent predicted forced expired volume in one second (observable entity)

Avoid Mistakes

- Avoid careless mistakes
  - Many subset design processes are still to some degree manual
    - They are prone to human mistakes
  - Careless mistake examples
    - Concept goes into the wrong subset
    - Digit goes missing
    - Inactive concept
  - Do not assume that this will never happen to you
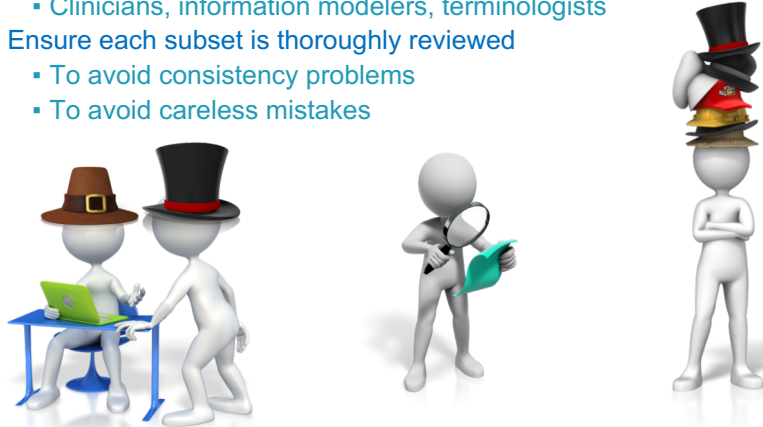
Avoiding Mistakes - Solutions

- Create automatic tests
  - Check concept ids in your subset are active
  - Check each concept's fully specified name and preferred term
- SNOMED-specific subset development tool
  - Speed up the process of finding appropriate concepts
  - Avoid copy-paste errors
  - Use existing subsets or ECL as starting point
  - Browse for additional concepts
  - Validation checks and review processes

Multiple Perspectives

- Use reviewers with multiple perspectives
  - Ensure team has necessary capabilities
    - Clinicians, information modelers, terminologists
  - Ensure each subset is thoroughly reviewed
    - To avoid consistency problems
    - To avoid careless mistakes

Demonstration

Demonstration

- SNOMED International Refset Management Tool
  - http://snomed.org/tools
  - https://refset.ihtsdotools.org
- Refset Management Tool
  - Enables the management & creation of reference sets against the International Edition of SNOMED CT and Member extensions
  - Provide a directory of existing reference sets that can be searched and downloaded to be used by others

Extensional Refset Example

- Rare Diseases concepts
  7199000, 9014002, 13213009, 16631009, 22053006, 23238000, 24700007, 30188007, 31323000, 44785005, 51615001, 58606001, 62067003, 63702009, 65389002, 65880007, 74911008, 75053002, 76670001, 77128003, 80651009, 82275008, 86044005, 128241005, 190794006, 195353004, 230418006, 230791000, 234542004, 236403004, 239928004, 252246005, 387732009, 387759001, 396338004, 410795001, 417357006, 423590009, 699310000, 716997004

- Concept to add
  190905008 | Cystic fibrosis (disorder) |

Intensional Subset Example

Pain Screening

Body site (coded text)	^ |Body sites|
Pain quality (coded text)	^ |Pain quality subset|
Severity (coded text)	< 272141005 |Severities (qualifier value)|
Frequency (coded text)	^ |Pain frequency subset|

Questions?