Lessons Learned from the Design, Development and Implementation of HDD Access

An Open Source Terminology Server with Publicly Available Terminology Content

OSEHRA Architecture Working Group
February 4, 2014

Senthil K. Nachimuthu, MD PhD
3M Health Information Systems, Inc., Murray, Utah
snachimuthu@mmm.com
A Variety of Lessons!

- Bridging the gap between publishers and users of standard terminologies
- Helping clinicians and researchers use standard and local terminologies together
- Applying medical informatics, software engineering, medicine, computer science, bioinformatics, mathematics, open source and intellectual property law
Affiliations and Conflict of Interest Disclosure

- 3M Health Information Systems, Inc.
  - Medical Informaticist – research and design lead of HDD Access
  - Relevant elected/standards development positions:
    - AMIA Open Source Working Group (past chair)
    - HL7 Vocabulary WG (co-author of CTS2 standard)
    - IHTSDO Technical Committee (US representative)
Introduction and Background
Terminologies
Terminology Mapping
What led to HDD Access
Terminology, Vocabulary, Code System

- The set of all concepts, their designations and relationships used in a specific field or application

- A “controlled terminology” has a consistent governance process, and follows good practices such as those described by Cimino’s desiderata (1998)

- There are specialized subtypes of terminologies, such as taxonomies, nomenclatures and formal ontologies
How Many Terminologies Does an Organization Need?

- **Standard terminologies**
  - Clinical: SNOMED CT
  - Lab: LOINC
  - Pharmacy: RxNorm, other proprietary pharmacy terminologies
  - Billing/reporting: ICD-9-CM, ICD-10-CM, ICD-10-PCS, CPT, HCPCS, DRG, and many others

- **Local terminologies**
  - Those created by vendors of EMRs, pharmacy and lab systems
  - Those created by hospitals themselves
  - Often “uncontrolled” without strict governance processes
What If There Are Many Terminologies?

- If each application in a hospital were compared to a healthcare provider
  - Each doctor, nurse, pharmacist or pathologist speaks a different language
  - They don’t understand what the other person says
- If the applications cannot understand data from each other, then we cannot use computers effectively to treat patients
- We want to enable the use of multiple terminologies together and to enable translations between them
Barriers to Terminology Adoption

- Standard terminologies may be available free of cost or for a fee, but it is still difficult to implement them.
- Each terminology has its own terminology model, distribution format, release frequency and governance process.
- Difficult to make applications use multiple terminologies seamlessly.
- Difficult to make multiple standard and local terminologies work together.
Terminology Mapping – Two Schools of Thought

Sometimes, both are used together

- **Centralized mappings**
  - For $n$ terminologies, we need $O(n)$ map sets, a minimum of $n-1$ as in the above example.

- **Point-to-point mappings**
  - For $n$ terminologies, we need $O(n^2)$ map sets, a minimum of $n(n-1)/2$ as in the above example.
Terminology Formats and Integration

- Different terminologies have different logical and physical designs.
- It is hard to keep them in their native formats while integrating them in a single terminology server.
- We maintain their logical structure as much as possible while transforming their physical structure to fit within the terminology server.
- We start by importing only the parts that we (our users) want, and add as needed.
Our Answer to the Terminology Management Question

- The 3M Healthcare Data Dictionary (HDD) was created in the early 90s through an SBIR grant.
- Two of the early users were the DoD EHR and Intermountain Healthcare.
- The 3M HDD is a concept-based terminology server that integrates the current version of many standard and local terminologies used by hospitals and other healthcare organizations.
- Partitioning capabilities are included to support terminology content from different organizations.
DoD, VA and HDD Access

- The Department of Defense (DoD) provides healthcare to active duty military personnel – DoD AHLTA EHR system uses the HDD as its data dictionary

- The Department of Veterans Affairs (VA) provides healthcare to veterans and their families – the VA VistA EHR system uses its own terminology

- In 2009, the DoD and the VA announced the joint iEHR (Integrated Electronic Health Record) project

- The DoD and VA gave 3M a contract in 2012 to make the 3M HDD publicly available as HDD Access, and they will use this public version for interoperability between their EHR systems
HDD Access
What is HDD Access?

- An open source terminology engine to enable
  - Implementation of various standard terminologies
  - Interoperability between multiple local and standard terminologies

- Contains various standard terminologies and a subset of the 3M HDD to enable interoperability between them

- HDD Access supports standard terminologies by bridging the gap between SDOs and users – it doesn’t compete with standard terminologies

- Supported by DoD and VA; published by 3M since May 2012
HDD Access Components

- Includes terminologies that we have permission to release publicly
- User can create local extensions without modifying the core content
- Database (Oracle, SQL Server, MySQL)
- Terminology browser, authoring tool and search engine
- API (HL7 CTS, 3M)
  - Runs on Windows and Linux
- Import/export tools
- 1200+ users
- Discussion forum, blog, FAQ
- Open Exchange – users can share their contributions
HDD Access Terminology Design

- Concept-based terminology, with concept permanence and graceful evolution
- Multi-hierarchical terminology
- Supports semantic relationships
- Supports and integrates terminology models and content of multiple standard and local terminologies
- Supports mappings between equivalent and non-equivalent concepts
HDD Access Content

- Concept-based terminology (medical vocabulary + knowledge base)
- Information model
- Integrates multiple standard and local terminologies
- Supports multilingual content
- Includes an example terminology to demonstrate mappings and how to create user’s own terminologies
Concept

- A concept is a unique, definable idea or item that has a specific, known meaning.

**Concept Representation**

**COLD**
- a sensory perception
- “I’m feeling cold”
- NCID 68215

**COLD**
- a pulmonary diagnosis
- Chronic Obstructive Lung Disease
- NCID 1005480

**COLD**
- an upper respiratory viral infection
- “I have a cold”
- NCID 1005313
Concept Representation Cross Reference

Chickenpox

Patient

Varicella

Physician & Nurse

3M Preferred Representation

Infection

Coder/ Bill (ICD-9-CM) 052.9

Unified Medical Language System CUI

Customer A Interface Code 1350

© 3M 2014. All rights reserved
Knowledge Base Relationships

Lab Test

Chem 4

Sodium

Is a

Has component

Potassium

Chloride

Glucose

Lab Result
Local Extensions

- A local namespace where the user can author their own terminology content, without modifying the core content.
- We provide a range of identifiers for the user’s local extension, to avoid collision with other users.
- Documentation describes how to author content in local extensions.
- HDD Access Content License allows the user to use or share their local extension with others.
- HDD Access tools allow export and import of local extensions.
HDD Access v.18 – Content Domains

Published in January 2013

- Allergies
- Laboratory tests, results, specimens, microbiology, units
- Demographics, encounter (ADT)
- Problems
  - Diagnoses (cardiovascular, digestive, musculoskeletal, endocrine, neuro, urinary, eye, ear, genital, reproductive, respiratory, …)
  - Clinical findings (vitals, CV, digestive, eye, ear, neuro, resp., …)
- Procedures
  - Musculoskeletal
- Medication
  - Drug, ingredient, form, route, dose/units, adm. site, frequency
- HIPAA standard code sets
- Radiology
- Provider taxonomy
- Qualifiers
HDD Access Content v.18 – Standard Terminologies

Terminologies that we have permission to include in HDD Access are supported and duly acknowledged

- ICD-9-CM Diagnoses
- ICD-9-CM Procedures
- ICD-10-CM
- ICD-10-PCS (codes but not attributes/hierarchies)
- HCPCS Level II
- HCPCS Modifiers

- MS-DRG
- TRICARE/CHAMPUS MS-DRG
- APC
- MS-DRG MDC
- TRICARE/CHAMPUS MS-DRG MDC
- RxNorm (since Nov 2013)
HDD Access Content – Data Row Counts

- Concepts
- Relationships
- Representations
- Representation Contexts

- 0
- 500,000
- 1,000,000
- 1,500,000
- 2,000,000
- 2,500,000
- 3,000,000
- 3,500,000
- 4,000,000

- Aug-12
- Oct-12
- Dec-12
- Feb-13
- Apr-13
- Jun-13
- Aug-13
- Oct-13
- Dec-13
HDD Access Information Models

- 3M Medical Information Models are successfully used by the DoD and other EHR systems
- ASN.1 format
- Includes terminology bindings to the HDD
- Can be used with the HDD Access terminology
- Covers many different clinical and administrative domains
- Helps to compose complex pieces of information from discrete pieces of clinical data (e.g. vital signs observation)
- Open source, Apache License v2
HDD Access Software

- Supports H2 (embedded), Oracle, Microsoft SQL Server, MySQL databases
- Runs on Windows and Linux operating systems
- Web-based search/browse tool included in the download
- Web services API includes a partial set of HL7 CTS v1.2 functions, and selected 3M functions – more are being added
- Installer (windows and linux), binary and source releases
New Since Our Last Presentation in November 2012

- Support for Microsoft SQL Server 2008
- Content authoring API (look under 3M API in documentation)
- RxNorm content published in November 2013
HDD Access Community

- More than 1,200 users, from government, industry (HIS vendors, consulting), academia, and international users
- Active discussion forums, blog, FAQ
HDD Access Download Counts

- Installer
- Binary
- Source
- DB Source
- Content
- Documentation

Total monthly download counts from August 2012 to December 2013.

© 3M 2014. All rights reserved
HDD Access Software License

- HDD Access software is open source under the Apache License v2 – source code is published in addition to the binary installers

- The license allows you to use, modify, and release your modifications
  - You are also allowed to release your derivative work under a proprietary license

- Your modified works should be clearly identified, and should not be identified as that of the original author

- No warranties
HDD Access Content License

- HDD Access content is “publicly available” under the HDD Access Content License
- You cannot modify or re-release the core HDD Access content
- But you can extend it and release your extensions as you please
- This is done to protect the integrity of the terminology content, and in turn, patient safety
- No warranties
Current Research Areas
Namespace Dependencies
Search Engine
Namespace Dependencies

- HDD Access core content and local extensions are made of “namespaces”, which are logical partitions of the content.

- Namespaces may have dependencies among them, since we integrate multiple terminologies using a central concept-based terminology.

- Example 1: The SNOMED CT concept of “Aspirin (substance)” and the RxNorm concept of “Aspirin” (ingredient) are mapped to the HDD concept “Aspirin” – the SNOMED CT namespace and the RxNorm namespace depend on the core namespace.
Namespace Dependencies

- Example 2: If we load the “SNOMED CT to ICD-10-CM Mappings from NLM” into HDD Access, the map set will be in a separate namespace which depends on the SNOMED CT namespace and the ICD-10-CM namespace.

- Soon, we will have a complex network of namespaces.

- The dependencies are between specific versions of namespaces, rather than the unversioned namespaces.

- The “versioned” namespace dependencies form a directed acyclic graph.
Namespace Dependencies – Our Approach

- Users define the namespaces that their namespace depends on.
- The “versioned” namespace dependencies are automatically calculated by the software.
- Content can only be moved from the “leaf” towards the “root” of the dependency tree, not in the other direction (which would break concept permanence).
HyperSearch – Search Engine

- Users want to search the terminology using the text representations of a concept
- The search engine needs to support synonyms, acronyms, acronym expansions, inflectional variants, different word orders, spelling variants and typos
- We created HyperSearch, a search algorithm that has the above capabilities
- Described in AMIA 2005 paper by Nachimuthu and Lau
- Part of the HDD Access terminology browser
HyperSearch – Search Engine

- Uses UMLS Lexical Tools (LVG) and Apache Lucene
- We created a permutation algorithm that has a similar approach as the DNA shotgun sequencing algorithm
- This has been internally used by our mapping tools since 2005
- We are working on a newer version to improve accuracy, performance and add more features
In Progress

- We are working on the following to meet user requests
  - Support for other ONC-recommended standard terminologies pending permission from SDOs: LOINC, SNOMED CT
  - Support for CTS2 services

- Please send us your requests, questions and feedback through our discussion forums or the ‘Contact Us’ page
HDD Access – References

- HDD Access website: http://www.hddaccess.com
- Online browser: http://search.hddaccess.com
- FAQ: http://www.hddaccess.com/home/faqs
- Discussion forums: http://www.hddaccess.com/forum
- Contact us: http://www.hddaccess.com/contact
- Apache License v2: http://www.apache.org/licenses/LICENSE-2.0.html
- HDD Access Content License: http://www.hddaccess.com/hdd-content-download-license
Demo and Questions

- Demonstration of HDD Access
- Questions?
- Thank you!

Senthil K. Nachimuthu, MD PhD
3M Health Information Systems, Inc.
snachimuthu@mmm.com