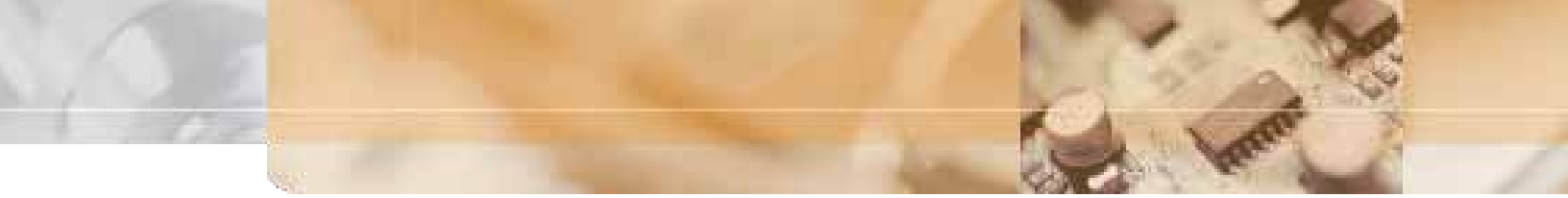


Using a Medical Data Dictionary to Comply with Vocabulary Standards and Exchange Clinical Data

What is needed is an integrated, structured terminology system that can mediate the differences between vocabularies and code sets and easily be incorporated into existing clinical information systems.

The Standards Final Rule

On July 13, 2010, the Department of Health and Human Services (HHS), through its Office of the National Coordinator for Health Information Technology (ONC), issued a final rule providing details on an initial set of standards, implementation specifications, and certification criteria for electronic health record (EHR) technology. The final rule lists the standards and criteria that EHR systems must meet in order to be considered "certified." Under the **Health Information Technology for Economic and Clinical Health (HITECH) Act**, in order for professionals and hospitals to qualify for certain incentive payments, they must demonstrate meaningful use of such certified technology.

ONC has adopted an initial set of vocabulary standards to support the proposed requirements for Stage 1 of meaningful use, which begins in 2011. The vocabulary standards that ONC has adopted and against which EHRs must be tested for certification include the following:

- ICD-9-CM
- SNOMED CT®
- Code sets by RxNorm data source providers
- CPT-4®
- LOINC®
- Vaccines administered (CVX)
- OMB standards for race/ethnicity

A Practical Approach to Implementing Vocabulary Standards

The current state of communication between health information technology (HIT) systems is fraught with uncertainty and ambiguity, which can compromise healthcare quality, efficiency, and safety. Vocabulary standards can be particularly challenging to effectively and efficiently implement across a healthcare organization or EHR system. Is there a "right" medical vocabulary that can describe, access, and make sense of all clinical data?

CPT is a registered trademark of the American Medical Association.

Unfortunately, no single healthcare vocabulary or terminology can meet all the needs of all the people who use healthcare information. Each terminology in use today has been designed for different purposes by different healthcare constituencies. What is needed is an integrated, structured terminology system that can mediate the differences between vocabularies and code sets and easily be incorporated into existing clinical information systems.

Over three decades ago, the concept of a **medical data dictionary** began as a means of supporting clinical information systems and the computerized patient record. A well-designed, intelligent medical data dictionary with terminology services can enable a healthcare organization or EHR system to comply with vocabulary standards and exchange clinical data, received in all of its various formats and terminologies from diverse systems and sources. When applied in real-time to an organization's information system infrastructure, such a dictionary can:

- Describe clinical data in all its possible forms, providing a road map to the content and structure of the patient database
- Support encoding of clinical data to remove ambiguities
- Support exchange and comparison of data between independent computer systems
- Provide structure and content for decision support across care encounters
- Enable users to effectively query and report on the database
- Support standardization of clinical data across enterprises by incorporating industry-standard **controlled medical vocabularies (CMVs)** as its foundation

The key to accomplishing these tasks is the incorporation of *de facto* CMVs (e.g., LOINC®) into the medical data dictionary itself. Adding and "mapping" (cross-referencing) industry-standard CMVs to a dictionary are labor-intensive efforts that many organizations cannot manage with existing resources. However, the resulting product supports standardization across enterprises and enables clinical data to be exchanged, compared, and aggregated at local, regional, national, and even international levels.

The vocabulary must be carefully monitored to avoid duplication, support synonyms, and completely describe terms from all areas of medicine—these characteristics lay at the heart of a **medical data dictionary.**

What's Not Making Sense in Clinical Data Today?

Computers easily process enormous volumes of data, but computers also "expect" unambiguous data in a specific form. Using computerized clinical data means more than simply storing numbers and words. Terms must be clearly defined and placed in a context.

What does the word "cold" mean?

- An accident victim brought into the ER tells the attending physician, "I feel cold."
- A pulmonologist tells a 58-year-old male patient that he is suffering from COLD (Chronic Obstructive Lung Disease).
- You call your family practitioner for an appointment and tell the receptionist, "I have a bad cold that's not getting better."

If the word "cold" is recorded on a patient's chart during any of these three scenarios, how does that word become an accurate and meaningful part of the EHR? How can the term be translated so that outcomes can be understood and managed and the medical events themselves accurately described and interpreted?

What is the "correct code" for a glucose result?

An organization currently receives laboratory test results from two different external vendors, a government lab, and several internal STAT labs. Each external system uses its own distinct set of test result codes, which differ from the internal set of codes used by the organization's STAT labs. Patient John Doe has glucose test results from one or more of these labs—how can the organization's order entry and results review applications "reconcile and reunite" all of these different test codes and results in the patient's record? How can clinicians "see" all of Doe's glucose results together for effective comparison and comprehensive analysis?

How can a healthcare organization make well-informed decisions regarding future expansion and resource deployment?

As part of the strategic planning process, a healthcare organization's team of risk managers, care providers, and financial officers wants to examine population studies by facility to determine how best to deploy an expensive special service like cardiology. How can the team get an accurate report of all cardiology-specific services and treatments delivered across the organization?

In all of these situations, people are defining the particular term, code, or information query according to a particular context. In the first example, "cold" can be a sensory perception, a pulmonary diagnosis, or an upper respiratory viral infection. The human mind easily resolves the ambiguity of a word like "cold," especially when the term is considered in its context. But how can a computer track and correctly interpret "cold," especially when it can appear in various forms and many different contexts?

Bottom line: The vocabulary must be carefully monitored to avoid duplication, support synonyms, and completely describe terms from all areas of medicine; these characteristics lay at the heart of a medical data dictionary.

What Is a Medical Data Dictionary?

A medical data dictionary is a database that describes the organization and logical structure of the medical data found in a clinical database. It contains "metadata"—or "data about data"—that describes the content, structure, and relationships between clinical data. In short, a medical data dictionary "translates," precisely defines, and effectively accesses the contents of the **EHR**.

What Information Does Today's EHR Contain?

A look at the information stored in today's typical EHR uncovers the Tower of Babel that exists in healthcare:

Data comes from and resides in many different systems and databases:

- Laboratory
- Radiology
- Pharmacy
- Hospital information systems
- Medical record coding
- Billing

Data from these diverse sources exists in many different *formats*:

- Coding and classification system formats
- Controlled medical vocabulary (CMV) formats (a typical patient EHR contains ICD-9-CM codes, CPT® codes, LOINC® codes, etc.)
- Proprietary formats from "home-grown" and commercial laboratory, hospital information, billing, and pharmacy systems and databases

Who Is Trying to Use the Information in the EHR?

The data in the EHR is valuable and needs to be available to many different people, ranging from clinicians and administrators to researchers and government regulatory agencies. The data should be retrieved and returned to these people in the appropriate format and with the correct degree of granularity for the audience. Data needs to be accessed in various ways and for many different purposes: patient clinical reports, statistical studies, *ad hoc* reporting, regulatory requirements, etc.

Not only must all of this data in the EHR be integrated, it must also be "normalized" into a form that can easily be shared by all audiences. For example, if users are compiling organization-wide reports on acute myocardial infarctions, they want quick access to all cases in the database, regardless of data format or terminology (e.g., "MI," "myocardial infarction," ICD code 410, etc.). By the same token, users should be able to retrieve all those cases—and even the associated patient records—for further study in their chosen data format.

A well-designed and intelligent medical data dictionary defines the connections between "MI," "myocardial infarction," and ICD code 410 within an EHR, and also makes it possible to retrieve and return the term most appropriate for a given audience.

Organizations can send messages to one another only to find that because the content of the message is not standardized, they still are not successfully "talking" to one another. This is the reason that HL7 version 3 is focusing on vocabulary standardization in its development and also why ONC considers vocabulary standards so important.

Other "Language" in the EHR: Interface Terminologies (Organization- and User-specific Terms)

A harsh fact of life for today's typical healthcare organization is that it has invested heavily in legacy information systems. Such information systems transmit data among themselves by means of interfaces, and each interface system has its own "language" or terminology. Many legacy systems use their own proprietary codes to describe clinical data, and these codes are not "understood" by any other system. The number and types of interfaces and proprietary coding systems are unique and specific to each healthcare organization.

The limits of Health Level 7 (HL7)

Interfaces do not evaluate the *content* of the messages they transmit. The healthcare industry-standard interface protocol, **HL7**, has standardized the message structure (syntax) of healthcare information transmissions. Thus, every organization that is HL7-compliant can send messages to other HL7-compliant organizations; as long as the message syntax is correct, HL7 considers the transmission successful. However, in HL7 version 2, while the syntax may be correct, the actual content of the transmission can be gibberish.

Organizations can send messages to one another only to find that because the content of the message is not standardized, they still are not successfully "talking" to one another. This is the reason that HL7 version 3 is focusing on vocabulary standardization in its development and also why ONC considers vocabulary standards so important.

Site-specific Terminologies

The people within a healthcare organization will continue using their accustomed "languages" and vocabularies. If they are familiar with a particular legacy system and its idiosyncrasies, they will continue to use that system. If site-specific pharmacy formularies are in place, those will continue to be used in addition to any other globally standard formulary. All of these "interface terminologies" and site-specific entities are legitimate components of a healthcare organization's "working vocabulary" and should be a part of an intelligent medical data dictionary.

How Does a Medical Data Dictionary Benefit Healthcare Organizations?

The power of the medical data dictionary is its ability to encompass and create the links between industry standard CMVs, coding sets, and organization-specific vocabularies. By integrating both global and site-specific vocabularies, the medical data dictionary allows **data in an EHR to be cross-referenced to standards everyone can understand, rather than locking the data up into yet another proprietary language that only a select few can speak.** When a dictionary can seamlessly translate universal and organization-specific data "behind the scenes," the healthcare organization can save time, money, and resources by avoiding the retooling or replacement of legacy systems and the re-training of its people.

Protecting an organization's investment. When a dictionary has been designed to be both extensible and flexible, it can encompass all of the diverse "languages" spoken by an organization and its legacy systems. Rather than replace legacy systems, the organization can use the dictionary with terminology services to translate between such systems and normalize the data they send into the EHR.

Accommodating future adoption of vocabulary standards. A well-designed medical data dictionary can seamlessly integrate classification systems such as ICD or CPT® along with other standard vocabularies and classification systems. Whatever HHS's future adopted standards turn out to be, they can be enveloped in a medical data dictionary as additional CMVs and consequently cross-referenced to the other CMVs still in use within a healthcare organization or vendor system.

Increasing quality. Two of the touchstones of an intelligent medical data dictionary are its insistence on unambiguous data and its use of industry-standard CMVs to help reduce the opportunities for misinterpreted, inaccurate, or imprecise data to become part of the patient's record. For example, to facilitate an automated pharmacy ordering system, a medical data dictionary can and should include a standardized pharmacy CMV (such as First Data Bank) that includes an accurate list of National Drug Codes (NDCs). When a dictionary can "enforce" accurate drug codes and positively identify the drug(s) being ordered, human errors and mistakes can be reduced.

Facilitating interoperability. The challenge here is for an organization to achieve comparable data with its business partners and move toward standardized code sets. A dictionary that adheres to sound vocabulary principles can identify a concept even when it is identified by numerous codes; these codes may even represent both past and present usage. For example, an ICD-9-CM code that is no longer acceptable to a claims payer must still be part of the dictionary to support historical data.

An Industry-tested Implementation Tool

The **3M™ Healthcare Data Dictionary (3M HDD)** was created over 15 years ago, and it has been continuously expanded and maintained throughout that time. It is the result of combining industry experience and professional and academic expertise to create a very practical implementation of informatics principles and industry-standard vocabularies. Supporting such applications as clinical information exchange, data warehousing, order entry, and results review, the 3M HDD is operational in the "real world," having been mapped in multiple commercial healthcare enterprises and all of the hospitals and clinics supported by more than one hundred Military Treatment Facilities (MTFs) worldwide in the

CPT is a registered trademark of the American Medical Association.

Department of Defense's AHLTA project. 3M's strategy has been to partner with healthcare organizations to help them make optimal use of their patient data, with the goal of improving quality of care, outcomes, costs, and competitiveness.

The strengths of 3M's dictionary lie in its:

- Structure and depth/breadth of content
- Architecture
- Ability to map "local extensions"
- Clinical foundations/expertise

3M Healthcare Data Dictionary's Structure: Depth and Breadth of Content

The 3M HDD comprises an **information model**, **vocabulary**, and **knowledge base**.

Information model

The 3M HDD uses an information model to accurately represent clinical data in the **3M™ Clinical Data Repository**. An information model describes how the vocabulary concepts should be used and how data can be combined to create meaningful database records that represent clinical events. It can be thought of as a set of "grammar rules" that show how data interacts with other data. The information model establishes temporal and spatial contexts for patient data, so that clinical observations can be attributed to the correct patient, clinical observer, and time sequence, and describes the appropriate information domains and types of values that should be present (see *Figure 1 below*).

A medication order consists of:	
Drug	What drug is being ordered?
Dose	What dosage is being ordered?
Route	How will the drug be given?
Frequency	How often will the drug be given?
Start time	When (date/time) will the drug first be given?
End time	When (date/time) will the drug stop being given?
Ordered by	Who ordered this drug?
Order #	What is the order number?

A medication order can be represented in the computer as:	
MedicationOrder ::=SET {	
drug	Drug,
dose	Decimal,
route	Route,
frequency	Frequency,
startTime	DateTime,
endTime	DateTime,
orderedBy	Clinician,
orderedNum	OrderNumber}

Figure 1. How concepts are placed in a 3M Healthcare Data Dictionary information model

Vocabulary: Industry-standard CMVs and coding and classification systems

The vocabulary component of the 3M HDD identifies and represents the various medical concepts found in clinical data, and it is organized to support synonyms and other lexical characteristics. As of January 2010, the 3M HDD contains over 2.6 million active concepts, over 17.9 million representations, and over 16 million relationships. The source vocabularies include:

- SNOMED CT®
- RxNorm
- Unified Medical Language System (UMLS)
- LOINC®
- National Drug Codes (NDCs) from the First Data Bank Pharmacy database
- ICD-9-CM
- Diagnostic Related Groups (DRGs)
- All Patient DRGs (AP-DRGs)
- All Patient Refined DRGs (APR-DRGs)
- CPT®
- HCPCS
- PTXT (from the 3M™ HELP System)
- Customer vocabularies (legacy systems, local and organization-specific terms)

In content, the 3M HDD encompasses:

- Encounter data and demographics
- Laboratory
- Microbiology
- Pharmacy
- Diagnostic and procedural coding
- Findings, signs, and symptoms
- Problem lists and diagnoses

Concepts and concept IDs

In the 3M HDD, a concept is a unique, definable idea or item that has a very specific, known meaning (e.g., cold, temperature, sensation, viral infection, infection, diagnosis) or a combination of concepts ("chest x-ray"). In the 3M HDD, each concept is defined by both a human-readable text description and an assigned, unique numerical identification, referred to as an "NCID" ("Numerical Concept Identifier," which has no intrinsic meaning or significance in itself). No redundant concepts are allowed, since they defeat the purpose of a controlled vocabulary. To demonstrate the need for concept IDs and a controlled vocabulary, consider how the word "cold" can be used in medical language, as shown below in *Figure 2*.

Concept ID	Definition
Cold, #123	A sensory perception ("patient complains of feeling cold")
Cold, #569	A pulmonary diagnosis (Chronic Obstructive Lung Disease)
Cold, #784	An upper respiratory viral infection ("common cold," "cold," "flu," etc.)

Figure 2. Sample of 3M Healthcare Data Dictionary's concepts and NCIDs

CPT is a registered trademark of the American Medical Association.

Synonyms—expanding the 3M HDD vocabulary

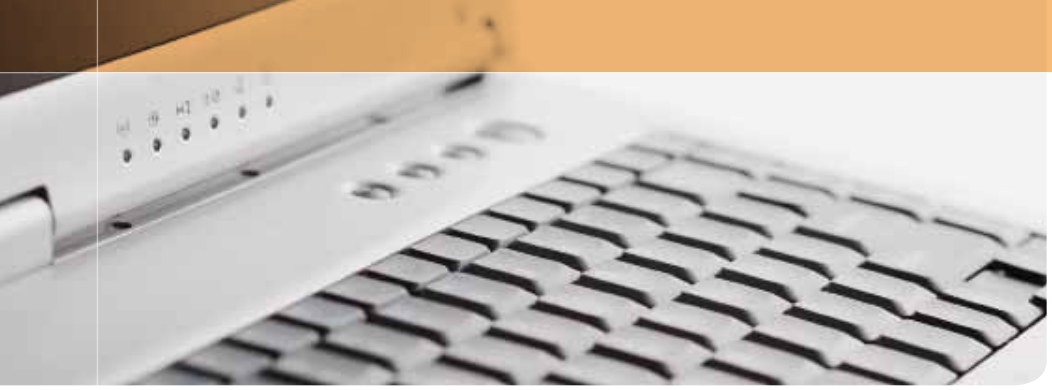
Much of the richness of the 3M HDD vocabulary comes from its use of synonyms (see *Figure 3*, below, for examples). In its use of synonymy, the vocabulary includes:

- Synonyms, homonyms, and eponyms (names derived from people or places)
- Different representations of the same concept, either in a natural language or other coded format
- Common variants of a term, such as acronyms or even common misspellings
- Foreign language equivalents (human languages—French, Spanish, Portuguese, etc.—can be added)
- The terms preferred in specific contexts (for example, "dyspnea" can be designated as the term for a cardiologist, while "shortness of breath" can be the preferred term for a lay person)

Synonym examples:
Acute Sinusitis
ACUTE SINUSITIS
Acute sinusitis, NOS
Sinusitis, acute
Acute infection of nasal sinus, NOS
Acute inflammation of nasal sinus, NOS
C0149512 (UMLS)
621850 (hospital-specific interface ID)

Possible contexts for a synonym—	
Synonym	**Context**
Acute Sinusitis	Problem list display
Acute infection of nasal sinus	Explanation
C0149512	UMLS code
621850	Interface code

Figure 3. Sample of how the 3M Healthcare Data Dictionary uses synonyms and can specify a context



The 3M™ Healthcare Data Dictionary's knowledge base

The 3M HDD's knowledge base consists of semantic networks and hierarchies that describe the complex relationships existing between concepts in the vocabulary. These relationships can be *hierarchical* (parent-child or "is-a") or *non-hierarchical* ("is-a-component-of"). *Figure 4* (below) is an example of how the knowledge base can describe the relationships between the components of a CHEM 4 laboratory test.

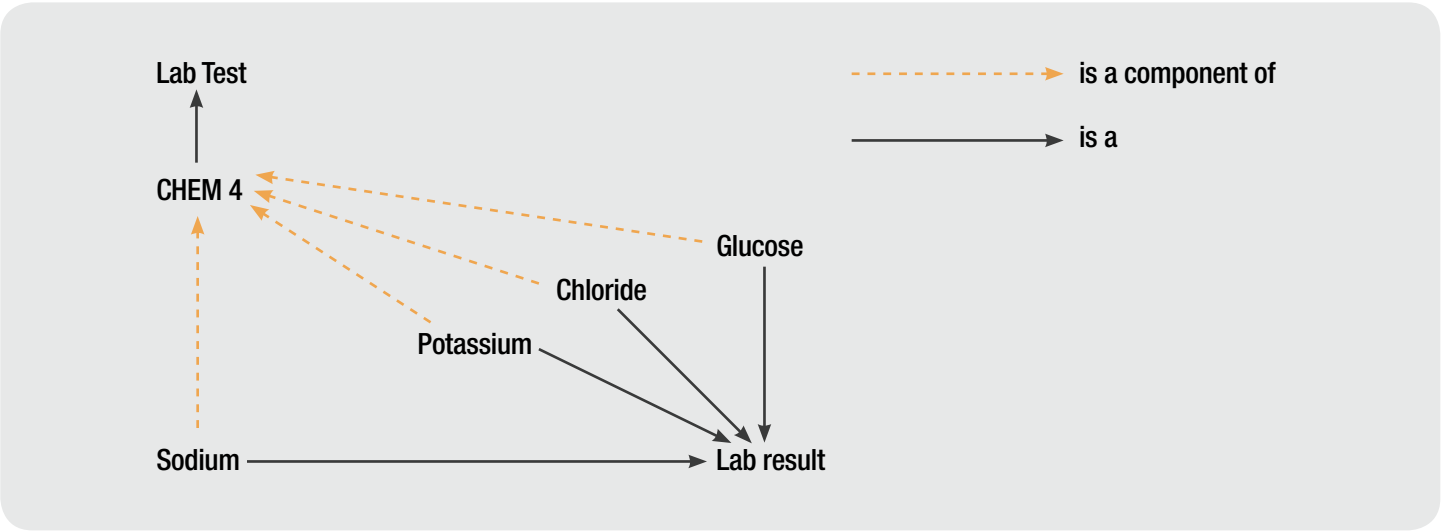


Figure 4. Sample of the 3M Healthcare Data Dictionary's knowledge base as applied to a CHEM 4 lab test

Figure 5 (below) shows how the 3M dictionary's **information model, vocabulary**, and **knowledge base** tie together to create a meaningful database record that represents a clinical event.

Information model		Database record	
MedicationOrder:	:=SET {	MedicationOrder {	
drug	Drug,	drug	Ampicillin,
dose	Decimal,	dose	500,
route	Route,	route	Oral,
frequency	Frequency,	frequency	Q6H,
startTime	DateTime,	startTime	08/01/09 10:01,
endTime	DateTime,	endTime	08/11/09 23:59,
orderedBy	Clinician,	orderedBy	John Doe MD,
orderedNum	OrderNumber }	orderedNum	A234567 }

Figure 5. Sample of a 3M Healthcare Data Dictionary information model (left) and the resultant database record (right) populated with vocabulary concepts and organized by the knowledge base

The 3M Dictionary's Ability to "Map" Source CMVs and "Local Vocabularies"

All industry-standard CMVs can coexist in the 3M HDD because of a process called "mapping," which cross-references elements in each CMV with a concrete, unambiguous concept in the 3M HDD.

Why mapping is necessary

- There is no single, universally accepted and applied standard vocabulary for the healthcare industry.
- Existing coding systems are incomplete.
- The HL7 version 2 standard specifies message *structure* only; it does not specify the actual data that is sent within the structure.
- Every healthcare organization uses different terms and codes.
- It is impractical and too expensive to replace all legacy systems.

Mapping "local extensions"

Because of the mapping process, an organization's "local" terminologies can also be integrated with the standard CMVs, along with such interface requirements as interface codes, billing codes, etc. Site-specific mapping is time and resource intensive, requiring highly trained and experienced clinical personnel who:

- *Understand* each CMV's inherent characteristics (e.g., "molecular" combinations)
- *Consider* each CMV's limitations (e.g., the lack of explicit relationships, reuses codes, etc.)
- *Design* a mapping strategy that:
  - Follows vocabulary principles
  - Meets user needs
  - Is flexible and extensible

The advantages of mapping

The ability to create local extensions means that the healthcare organization's users can continue using their own terms, while the dictionary seamlessly handles the "translation" of such terms behind the scenes. Mapping provides today's resource and time-strapped healthcare enterprise with several advantages:

- People do not need to be "retrained" in the language of another computer system. They continue using the terms they know and understand.
- Existing information systems do not need to be replaced or redesigned.
- Individual entities within an organization can retain specific information structures or sets, such as preferred formularies.
- A specific form of the concept can be displayed in a specific context.

Concept ID	Representation	Context
123	1234-5	LOINC code
123	NAS	Interface code for Site #1
123	Serum Sodium	User display for Site #1
123	CL357	Interface code for Site #2
123	S. Sodium	User display for Site #2

Figure 6. A sample of how the 3M HDD can represent a serum sodium laboratory result in different contexts

3M staff expertise and credentials

The 3M clinicians and informaticists who support the 3M HDD and provide mapping services belong to the following groups and associations:

- Laboratory and clinical committees of LOINC®
- American Medical Informatics Association (AMIA)
- HL7 Vocabulary Technical Committee
- Healthcare Information and Management Systems Society (HIMSS)
- American Health Information Management Association (AHIMA)

3M™ Healthcare Data Dictionary Architecture and "Platform Independence"

The 3M HDD was designed to meet open architecture standards, allow for platform independence, and conform to the following industry standards:

- Linux platform
- HL7 messaging
- HL7 Common Terminology Services (CTS)
- ASN.1 information model, translatable to XML
- Application Programming Interfaces (APIs) to access the 3M HDD

Finally, because the 3M dictionary is an independent entity, it can be updated with new medical knowledge without rewriting the application programs that use it.

Conclusion

Technology and industry drivers—not the least of which is HHS and ONC's final rule—are quickly making a medical data dictionary mandatory for any healthcare organization or EHR vendor that wishes to remain competitive. The potential power of a medical data dictionary lies in its ability to help healthcare organizations meet these challenges:

- Use the power of computers to **deliver decision support** to care providers in diagnostic, therapeutic, and management arenas.
- Significantly **improve outcomes and contribute to the quality and safety of patient care** by helping to reduce medical errors resulting from misinterpreted, inaccurate, or imprecise patient data.
- Effectively **integrate information systems** to increase the healthcare facility's competitive strength in the community and **reduce the costs associated with poor information infrastructure.**
- Meet the challenges posed by both **industry and governmental mandates** for uniform data standards and standardized clinical vocabularies, coding, and coding sets.
- Bridge communication gaps between healthcare providers, business partners, and consumers—*even as technology and software evolve*—and **allow expansion into Internet-enabled technologies, e-health relationships, and other communication vehicles.**

These challenges are best met through a working medical data dictionary that is an integral part of a clinical data repository. In design and architecture, the 3M Healthcare Data Dictionary can support existing technologies and continued use of legacy systems, providing a flexible and extensible "vocabulary server" that addresses regulatory pressures and serves the unique needs of both business partners and healthcare consumers.

Health Information Systems
575 West Murray Boulevard
Salt Lake City, UT 84123
800-367-2447
www.3Mhis.com

3M is a trademark of 3M Company. CPT is a registered trademark of the American Medical Association. LOINC is a registered U.S. trademark of Regenstrief Institute, Inc.

SNOMED and SNOMED CT are registered trademarks of the International Health Terminology Standards Development Organisation.

Please recycle. Printed in U.S.A.
© 3M 2010. All rights reserved.
Published 11/10
70-2009-9144-9