

## **HIT Standards Committee – Clinical Operations Workgroups – Task Force on Vocabulary**

Panel 3: Best Practices & Lessons Learned: Vocabulary Infrastructure

March 23, 2010

Testimony Submitted by Ken Buetow, Ph.D.

Director, Center for Bioinformatics and Information Technology

National Cancer Institute, National Institutes of Health

### **Background**

Vocabulary subsets and value sets play a vital role in electronic health records (EHRs), as well as in many other clinical, research, and public health activities. The National Cancer Institute (NCI) of the National Institutes of Health has developed a number of operational and technical approaches that have proved successful for NCI and its partners, and is actively pursuing new approaches to make them part of a comprehensive health IT infrastructure that supports personalized medicine and integrated care.

NCI terminology support is concentrated in NCI Enterprise Vocabulary Services (EVS) (see <https://wiki.nci.nih.gov/display/EVS/EVS+Wiki>). EVS relies on the National Library of Medicine's (NLM) Unified Medical Language System (UMLS) as an important source of licensed terminology content. NCI is a longtime licensee of UMLS. The NCI Metathesaurus, which is built on top of (and also extends) the UMLS Metathesaurus, is one long standing example of the use NCI makes of the UMLS. Another is that NCI relies on UMLS to obtain the sanctioned US distribution of SNOMED CT as well as several other terminologies. NCI is also a contributor to the UMLS, for example by providing monthly builds of NCI Thesaurus to NLM for inclusion as a source terminology in UMLS.

### **1) What vocabulary subset or value set creation and distribution services do you provide?**

NCI provides support for the creation and distribution of vocabulary subsets and value sets through both its terminology and metadata services.

Terminology services support the creation of named, tagged subsets and value sets within NCI maintained terminologies, most notably our NCI Thesaurus (NCIt) reference terminology (see <http://ncit.nci.nih.gov/>). The Cancer Community, FDA, CDISC, NCPDP and other partners are able to identify or create NCIt concepts using shared meanings and define their own subsets and value sets within that environment, sharing a common coding scheme while also including their own specific terms, codes, definitions, and other data. NCI terminology servers provide access to 20 separate terminologies, and more than 80 terminologies integrated in the NCI Metathesaurus, that can be used to define subsets and value sets through services built into the terminology server, through direct reference to is-a hierarchies or features, or through NCI metadata services (see <http://ncim.nci.nih.gov/>).

Metadata services support the creation of subsets and value sets associated with data element concepts. NCI's ISO/IEC 11179 metadata registry helps users identify or create concepts and sets of values that they can share. Where there are differences in values or even concepts, the registry facilitates identification of shared or similar meanings in the sets being used.

Creation and maintenance of subsets and value sets is supported in both environments by several important types of services:

1. Specialized editing, user feedback, quality assurance and other tools.
2. Editorial curation and support teams combining subject matter and technical expertise.
3. Training to promote user understanding of best practices and technical process.

Active engagement with and involvement of NCI, The cancer Biomedical Informatics Grid (caBIG) community, and other user communities in developing both content and technical aspects of subsets and value sets is a core feature of NCI services. Community-based resources such as caBIG's LexEVS terminology servers (see [https://cabig-kc.nci.nih.gov/Vocab/KC/index.php/LexBig\\_and\\_LexEVS](https://cabig-kc.nci.nih.gov/Vocab/KC/index.php/LexBig_and_LexEVS)) – which support open APIs (see [https://cabig-kc.nci.nih.gov/Vocab/KC/index.php/LexEVS\\_Version\\_5.1](https://cabig-kc.nci.nih.gov/Vocab/KC/index.php/LexEVS_Version_5.1)) and various end user tools (see for example <http://ncit.nci.nih.gov/>) – as well as collaborative terminology development resources such as LexWiki (see <https://cabig-kc.nci.nih.gov/Vocab/KC/index.php/LexWiki>), the caBIG Knowledge Centers (see [https://cabig-kc.nci.nih.gov/MediaWiki/index.php/Main\\_Page](https://cabig-kc.nci.nih.gov/MediaWiki/index.php/Main_Page)), and the various caBIG subject matter and cross-cutting workspaces (see <https://cabig.nci.nih.gov/concepts/workspaces/index.html>) provide forums in which cancer survivors, researchers and clinical care providers interact to identify the need for code sets, value sets and subsets are surfaced, and work to draft, review and support subsequent to deployment.

Distribution services address the wide range of practice and preference within the community including:

1. Subset and value set files are freely downloadable in delimited text, XML, Excel, and other formats as required.
2. Public browsers and wikis make subset and value set content available in their broader terminology and metadata contexts, with links to applications for user input and term suggestions.
3. NCI terminology and metadata servers are freely distributed for use elsewhere, and provide programmatic access via several types of application programming interface (API).

NCI and caBIG partners leverage the tight integration of our subsets and value sets with each other and with all layers of the NCI semantic infrastructure. At the same time, many others are primarily interested in one or a few of these services, and find it useful that they are separately available.

## **2) Who uses your services and what is the level of use?**

More than a hundred named, tagged NCIt subsets and value sets with some 20,000 values are maintained in collaboration with a variety of partners, most notably:

- U.S. Food and Drug Administration (FDA): Many FDA subsets are maintained in NCIt and required for regulatory reporting and other purposes. These include: 16 subsets used by Structured Product Labeling (SPL) for submission of proposed labeling by all manufacturers using electronic formats; Device Event Problem Codes subsets used for the reporting of medical device problems to FDA (roughly 3,000 different reporting locations used these subsets in 2009); and Individual Case Safety Report (ICSR) subsets used for adverse event reporting (proposed regulations for electronic submissions will create similar levels of use for these subsets).
- Clinical Data Interchange Standards Consortium (CDISC): All CDISC controlled terminology is maintained and published as NCIt subsets, including the Study Data Tabulation Model (SDTM), which is an approved standard for FDA submissions and has been downloaded more than 14,000 times in over 60 countries, primarily for institutional use.

- National Council of Prescription Drug Providers (NCPDP): Three NCI subsets have been adopted as part of the SCRIPT (10.5) and Telecommunication (D.3) standards employed by some 200 vendors serving approximately 15,000 pharmacies nationwide.

NCI subsets are used extensively within NCI and caBIG systems. The extensive commitment that NCI has made to create active collaborations uniting large segments of the cancer and biomedical community has facilitated wide-spread adoption and reuse of value sets and code sets. For example, community-based resources such as the caBIG Knowledge Centers and the various caBIG subject matter workspaces are forums in which patients, researchers and clinical care providers interact to identify the need for value sets and subsets, as well as working together on drafting, review, deployment and maintenance. Community-based resources such as the caBIG Knowledge Centers and Support Service Providers supplement and extend ability of the NCI to encourage adoption and proper use of subsets and value sets across the community.

NCI metadata services are one of the primary means by which these and other vocabulary subsets and value sets are used in NCI and other systems and applications. There are over 135 information models represented as ISO 11179 metadata in caDSR, recording the data semantics in software applications and in services on NCI's caGrid. This metadata encompasses some 20,000 data elements referencing subsets and value sets drawn from NCI terminology services, many shared between users. Significant users of metadata-based subsets and value sets are:

- NCI internal programs: Many NCI divisions, centers, and programs make use of metadata-defined subsets and value sets. The Cancer Therapy Evaluation Program (CTEP) alone uses more than 10,000.
- Other NIH institutes: Some 5,000 data elements are used by the National Heart, Lung and Blood Institute (NHLBI), the National Institute of Child Health and Development (NICHD), and the National Institute of Dental and Craniofacial Research (NIDCR).
- caBIG: Virtually all projects use metadata-supported subsets and value sets in their models, interfaces and information content.
- Biomedical Research Integrated Domain Group (BRIDG): Over 1,600 data elements are used in this domain analysis model of clinical and pre-clinical protocol-driven research created in collaboration with the Clinical Data Interchange Standards Consortium (CDISC), the HL7 Regulated Clinical Research Information Management Technical Committee (RCRIM TC), and the US Food and Drug Administration (FDA).
- CancerGrid: UK deployments of NCI terminology and metadata services are supporting a variety of projects at both the UK and EU levels.

### **3) What, if any, additional services and capabilities are in active development?**

NCI is enhancing its Terminology Services to fully support the Health Level 7 (HL7) Common Terminology Services II Specification (CTS2) (see <http://wiki.hl7.org/index.php?title=CTS2>) and its companion Object Management Group (OMG) Platform Independent Model for CTS2. Beginning in the Fall of 2010, NCI terminology services will be extended to fully support inter-terminology mapping, local extensions to standard terminologies, and value sets, as called for by HL7 and OMG standards.

Value set support under the CTS2-compliant services will include services to create, manage, organize, search and retrieve value set specifications and to record extensive metadata about them, including the reason the set was created, its intended purpose, responsible parties and their contact

information, version history, and terminologies and terminology versions from which the value set is drawn.

The CTS2-compliant services will also enable NCI to support the full range of needs for value sets, as called for in HL7 clinical communications strategies, including Version 3 Messaging, Common Document Architecture (CDA), and Services Oriented Architecture (SOA) implementations.

The CTS2-compliant services enhancement is one of an integrated series of service development projects intended to greatly extend NCI's ability to provide semantics support for national initiatives such as the ambulatory oncology extensions for electronic health records (caEHR). The other semantics initiatives include a Knowledge Management Service and a Rules Management Service. The Knowledge Management Service will provide human and machine interpretable information about the nature and use of all services and other information resources, including the purposes for which the services are intended. The Rules Management Service will provide human and machine interpretable information about the meaning of events and outcomes of interactions among services and certain other automated clinical and bioinformatics information resources. The Knowledge Management and Rules Management Services depend on the Terminology Service for standard terminologies, ontologies, value sets, terminology maps, and terminology extensions required to enable the services to operate across clinical and research systems and information coded in multiple, diverse terminologies. Both the Knowledge Management and Rules Management Services will employ natural language interpretation, machine reasoning and decision support functionalities that will depend on the Terminology Services.

#### **4) If applicable, what process is used to establish and revise any subsets or value sets that you distribute?**

Creation and revision of vocabulary subsets and value sets starts with stakeholder input and feedback. The caBIG program has provided an important framework for broadening NCI's internal processes to include Cancer Centers, Cooperative Groups, and a broad range of other government, academic, professional and private stakeholders. Many NCI-maintained subsets and value sets involve intensive ongoing interaction with specific partners such as FDA, CDISC, and NCPDP.

Expert curators compare requests and requirements with existing subsets, value sets, and other content in EVS terminologies and, when appropriate, the metadata repository. New or revised sets are often circulated for internal and external review before implementation, but requests cleanly aligned with existing content and policies can sometimes be released after basic internal QA.

Internal QA involves a variety of human and computer processes, as describe in a recent publication (The NCI Thesaurus Quality Assurance Life Cycle. *Journal of Biomedical Informatics* 2009 June;42(3):530-539.)

Given the extensive use of NCI subsets and value sets, we often get user feedback on possible extensions or modifications. NCI then consults partners and known stakeholders in the existing sets, and will either design and implement agreed upon changes or suggest other approaches (possibly including a separate extended or new subset or value set) to meet user needs.

Distribution is also a vital part of any process dealing with vocabulary subsets and value sets. While it is important that all such sets be accessible in file, API and browsable forms standardized for all sets, individual maintainers and users often have their own distinctive requirements that are

crucial to the success of the set.

The FDA subsets are an instructive example of matching process to specific use cases. NCI support of FDA terminology is conducted under a Memorandum of Understanding. Extension or modification of the underlying terminology is driven by FDA regulatory requirements, as well as ongoing quality assurance operations within NCI Enterprise Vocabulary Services (EVS). Subsets are defined to meet FDA operational needs, and are then reviewed and revised jointly by FDA, NCI, and other stakeholders during pre-release development. Distribution mechanisms are tailored to meet FDA and regulated industry requirements, as are online support and end user documentation. Publication and distribution of the subsets are mission critical responsibilities of NCI EVS and technical operations groups. Post-release revision of the subsets is an ongoing responsibility of the NCI EVS content curation group. The revisions are driven by the needs of the FDA and regulated industry, which surface their needs to the FDA and NCI project officers who oversee the revision and release process. Broadly similar arrangements are in place for the CDISC and NCPDP subsets.

A recent example of broad-based community collaborative development of a heavily used value set is the 4<sup>th</sup> edition of the Common Terminology Criteria for Adverse Events (see <https://cabig-kc.nci.nih.gov/Vocab/KC/index.php/CTCAE>), which was produced using the LexWiki service. The CTCAE v4 effort was sponsored by the caBIG Vocabulary and Common Data Elements Workspace (VCDE) (see <https://wiki.nci.nih.gov/display/VCDE/VCDE+Wiki+Home+Page>), which conducts an ongoing program to review and certify terminological products for use in caBIG and establishes the criteria the community will follow to achieve and sustain semantic consistency across clinical and research resources. Community-based resources such as the caBIG VCDE, Knowledge Centers and Support Service Providers supplement and extend NCI's ability to encourage adoption across the community and support proper use of terminology and metadata artifacts.

NCI's metadata environment has formal governance mechanisms that provide direction and oversight of the creation, deployment and re-use of Value Domains, Valid Value Lists and other components of the Common Data Elements that compose the bulk of NCI's metadata. Value sets and subsets drawn from the terminology services are bound within the metadata domain to specific business uses and are provided with situation-specific contextual representations. These business-specific representations are critical to understanding the nuanced meaning of data described by the metadata, and so have received careful curation from the beginning of the services.

## **5) Based on your experience, what advice would you offer regarding best practices and pitfalls to avoid?**

NCI has devoted many years' effort to standardizing and sharing its vocabulary subsets and value sets, as an integral part of efforts to build comprehensive content and technology standards that can create interoperability and synergies in our clinical, research, and public health systems. We have learned many lessons and worked to create best practices, involving process and institutional lessons at least as much as technical ones:

- Work closely with stakeholders to identify content, operational, and technical requirements. Existing terminologies, subsets and value sets reflect specific purposes and use cases, including important institutional, professional and regulatory constraints. For example, MedDRA is the internationally accepted standard for adverse event reporting, and NCI (with extensive community input) redesigned its CTCAE adverse event terminology to be a harmonized intersection with a subset of MedDRA, even while recognizing MedDRA's limitations and preferring other terminologies for most coding of diagnoses and findings.

- Subsets and value sets, as well as terminologies, terminology maps and extensions generally, are built for specific operational uses. It is important that information be provided about the purpose of a subset or value set, its provenance, currency, intellectual property limits, and other information pertaining to suitability for use, ideally as an organic part of the subset or value set.
- Especially with regulatory terminology or terminology that is to be adopted and used by large diverse communities, such as regulated device manufacturers or pharmacies, technical documentation about the structure and proper deployment of subsets and value sets must be provided. In many cases, online or telephone support for adopters is required to ensure proper use.
- For dynamic, flexible, reusable infrastructure, it is important to decouple value set management from concept definitions. For example, not all subsets and value sets are intended for use by regulated or large groups; some are conveniences intended for the use of a few people or perhaps a pair of cooperating groups. In the past, support for such small-scale uses has had to be limited, but with the advent of interactive services such as the CTS2 Terminology Services, such uses are expected to grow rapidly. On-line training and end user support for such users will be an important role for organizations providing terminology services.
- The infrastructure needed to ensure best practices in the creation, maintenance, and distribution of vocabulary subsets and value sets is complex and evolving. Some centralization is important to cost-effective provision of high quality services. At the same time, the vital role of interaction with both creators and users seems likely to put limits on centralization given current technologies.
- Use of available content and technical standards, and contributing to the development of such standards, is important to both implementation and adoption of subsets and value sets.
- Vocabulary subsets and value sets need robust and transparent mechanisms for input by affected communities. This will mean providing methods for public input for publicly defined sets.
- Workflow support is vital to the creation and maintenance process for any robust set of vocabulary subsets and value sets.
- Distribution formats often need to cover a very broad spectrum of users and implementations, ranging from very simple text files of terms and codes through to complex representations of full underlying vocabulary data. Failure to analyze and meet the specific requirements of target user communities can greatly impair use.
- The pace and patterns of versioning vary enormously with the purpose. Cutting edge activities often require 24-hour turnaround to support coding needs, while in other areas it is important to establish much longer revision cycles. History, change tracking, and version labeling and description require a uniform framework but also sufficient flexibility to address the full range of requirements.
- High quality, usable content requires an unusual blend domain and technical expertise. Sometimes a significant part of this can come from outside contributors, but both will be required of maintenance organizations to ensure quality and consistency.
- Diversity in underlying source terminologies is a fact of life for the foreseeable future. Improvements in mapping between them, and between subsets and value sets based on them, will be crucial to interoperable health information in the years to come.