A Content Model for the ICD-11 Revision

Samson W. Tu[1], Olivier Bodenreider[2], Can Çelik[3], Christopher G. Chute[4], Sam Heard[5], Robert Jakob[3], Guoquian Jiang[4], Sukil Kim[6], Eric Miller[7], Mark M. Musen[1], Jun Nakaya[8], Jon Patrick[9], Alan Rector[10], Guillermo Reynoso[11], Jean Marie Rodrigues[12], Harold Solbrig[4], Kent A Spackman[13], Tania Tudorache[1], Stefanie Weber[14], Tevfik Bedirhan Üstün[3]

[1]Stanford Univ., Stanford, CA, USA; [2]National Library of Medicine, Bethesda, MD, USA; [3]World Health Organization, Geneva, Switzerland; [4]Mayo Clinic College of Medicine, Rochester, MN, USA; [5]Ocean Informatics, Chatswood, NSW, Australia; [6]Catholic Univ. of Korea, Korea; [7]Zepheira, Fredricksburg, VA, USA; [8]Tokyo Medical and Dental Univ., Tokyo, Japan; [9]Univ. of Sydney, Sydney, NSW, Australia; [10]Univ. of Manchester, Manchester, UK; [11]Buenos Aires, Argentina;[12]Université de Saint Etienne, Saint Priest en Jarez, France; [13]IHTSDO, USA; [14]DIMDI - German Institute of Medical Documentation and Information, Köln, Germany

Abstract

The 11th revision of the International Classification of Diseases and Related Health Problems (ICD) will be developed as a collaborative effort supported by Web-based software. A key to this effort is the content model designed to support detailed description of the clinical characteristics of each category, clear relationships to other terminologies and classifications, especially SNOMED-CT, multi-lingual development, and sufficient content so that the adaptations for alternative uses cases for the ICD – particularly the standard backwards compatible hierarchical form – can be generated automatically. The content model forms the basis of an information infrastructure and of a web-based authoring tool for clinical and classification experts to create and curate the content of the new revision.

Introduction

ICD is the international de facto standard classification for most epidemiological and many healthcare and clinical uses. Originally designed to record causes of death, the usage of ICD has been extended to include morbidity classification, reimbursement, and several other specialty areas such as oncology and primary care. The current 10th edition of ICD was endorsed by the World Health Assembly in 1990 and has been updated periodically over the years.

ICD-10 is published in three volumes: a tabular listing of more than 155,000 codes, a reference on the usage rules of the codes, and an alphabetical index that maps linguistic terms (e.g., Dysacusis) to appropriate codes (e.g., H93.2). The tabular list is organized as a mono-hierarchy using letters for initial broad categorization (e.g., A and B for infectious diseases) and digits for each successive level of child codes. Sibling codes are required to be exhaustive and mutually exclusive, requiring the use of residual categories—"unspecified" and "other"—at each level.

A code may have associated *inclusions* and *exclusions* (Figure 1). *Inclusions* are exemplary terms or phrases that are synonymous with the title of the code or terms representing more specific conditions. *Exclusions* follow from the requirements for a mono-hierarchy of mutually exclusive siblings. Most *exclusions* are either conditions that might be thought to be children of a given condition but, because they occur elsewhere in the classification, must be excluded from appearing under it; others are codes representing possible co-occurring conditions that should be distinguished from the condition.

I21 Acute myocardial infarction
Includes: myocardial infarction specified as acute or with a stated duration of 4 weeks (28 days) or less from onset
Excludes: certain current complications following acute myocardial infarction (I23.-)
myocardial infarction:
- old (I25.2)
- specified as chronic or with a stated duration of more than 4 weeks (more than 28 days) from onset (I25.8)
- subsequent (I22.-) postmyocardial infarction syndrome (I24.1)

Figure 1. Example of ICD-10

The World Health Organization (WHO) aims to open ICD-11 development to a broad participatory Web-based process. Unlike previous revisions, which were performed manually using lists of codes for creating new drafts, the developmentof ICD-11 aims to create an information infrastructure and workflow processes that utilize knowledge engineering and management techniques supported by software [1]. In addition to the existing hierarchies of codes, titles, and supplementary

volumes of rules and indices, the information infrastructure will enable the definition of diseases and health conditions, encoding of the eotiology and the anatomical and physiological aspects of the disease, and mappings to other terminologies and ontologies. It hopes in this way to aid the review of best scientific evidences and support field trials of draft standards.

In terms of workflow, the information infrastructure will support the collaborative development of new content and proposed changes, review and approval processes, and the creation of draft classifications for field testing. Initially the work of Topic Advisory Groups (TAGs) for various specialty areas, the ICD-11 revision process will eventually be opened for comments and suggestions by interested parties in a social process on the Web. The Alpha Draft of ICD-11 will be completed by May 10, 2010. The Beta Draft should become available a year later.

The Content Model for ICD11 is the critical component that specifies the structure and details of the information that should be maintained for each ICD category in the revision process. In this paper we outline the requirements that the Content Model must satisfy , the basic structure of the model, and how it supports the software used to inspect, edit, and publish drafts of the ICD-11 revision.

Requirements

Backwards compatibility. As the most widely used standard coding system for diseases and related health conditions, this is a primary requirement. A code should not be retired unless there are compelling scientific reasons for doing so. Furthermore, the information infrastructure must support automated generation of the traditional morbidity and mortality classifications, with their inclusions, exclusions and indices, from the information curated on the basis of the Content Model.

Adaptations of ICD-10 modified ICD-10 for other use cases. The International Classification of Primary Care, Second Edition (ICPC-2),[1] for example, is a simplification of ICD-10 for encoding the reason for encounter, the diagnosis, and the treatment in episodes of primary care. It is the goal of the ICD-11 Revision to create the capability to generate the equivalent of ICPC-2 from its information infrastructure.

Coordination of the use of multiple classifications to specify the details of an ICD category. For example, the International Classification of Functioning, Disability, and Health (ICF),[2] another member of the WHO Family of International Classifications, should be used to describe the functional impacts of an ICD disease category. Similarly, the International Classification of External Causes of Injury (ICECI)[3] is another classification that should be coordinated with Chapter XX (External causes of morbidity and mortality) of ICD.

Clear relationships between ICD-11 and other de facto standard medical terminologies such as SNOMED-CT. There is inevitable overlap in the knowledge coverage of ICD-11 and other terminologies. Nevertheless duplication of effort to create competing and semantically non-interoperable terminologies is clearly undesirable.

Multingual and multicultural adaptation. As a coding system used around the world, WHO is mandated to support its official languages (Arabic, Chinese, English, French, Russian, and Spanish). Other national bodies have translated ICD to numerous languages. The Content Model and its supporting software must allow the incorporation of existing translations and facilitate the development and maintenance of ICD in multiple languages simultaneously.

Formalisation in a computer-interpretatable language. The Content Model must support software tools that enable content experts to view and curate the content, and that automate error checking and constraint enforcement.

The Content Model

The WHO ICD-11 Revision Steering Group (RSG) convened a Health Informatics and Modeling Topic Advisory Group (HIM-TAG) to develop the ICD-11 Content Model. The model is still evolving, but the main components have been specified (Figure 2). A detailed Guide document describes the expected content and usage of each component as seen by the user. It is the document that records the shared understanding of the Content Model.

This informal model is implemented in a three-layer model documented in UML: [4] a) The *Foundation layer* divided into (1a) the *Ontology layer* that is intended to be aligned with a subset of SNOMED, and (1b) the *Category layer* that contains the description of each ICD category; (2) the *Linearizations layer*—a generalization of the traditional ICD classifications that provides the backwards compabibility (including their inclusions, exclusions, and residual categories) and supports new use cases.

After Protégé was selected as the platform for supporting the curation of the Alpha Draft of ICD-11, the Content Model was implemented using the Web Ontology Language (OWL) supplemented by

1 http://www.who.int/classifications/icd/adaptations/icpc2/en/index.html

2 http://www.who.int/classifications/icf/en/index.html

3 http://www.rivm.nl/who-fic/ICECIeng.htm

4 http://informatics.mayo.edu/icd11model/v20090506/index.htm

metaclass constructs from Protégé. The OWL Content Model realizes the informal description in the Guide and formalizes the three-layer conceptualization originally documented in UML. It forms the basis for the ICD Collaborative Authoring Tool (iCAT) [2], a specialization of Web Protégé, that supports the web-enable workflow needed to produce the early drafts of ICD-11.

1. ICD Concept Title
2. Hierarchy, Type and Use
   - 2.1. Parents
   - 2.2. Type
   - 2.3. Use
3. Textual Definition(s)
4. Inclusion, Exclusion, and Index
   - 4.1. Base Index
     - 4.1.1. Synonyms
     - 4.1.2. Narrower Terms
   - 4.2. Base Exclusion
   - 4.3. Inclusions
5. Clinical Description
   - 5.1. Body System(s)
   - 5.2. Body Part(s) [Anatomical Site(s)]
   - 5.3. Morphologically Abnormal Structure
6. Manifestation Properties
   - 6.1. Signs & Symptoms
   - 6.2. Findings
7. Causal Properties
   - 7.1. Agents
   - 7.2. Mechanisms
   - 7.3. Risk Factors
   - 7.4. Genomic Characteristics
8. Temporal Properties
9. Severity Properties
10. Functional Properties
11. Specific Condition Properties
12. Treatment Properties
13. Diagnostic Criteria
14. External Causes

Figure 2. The components of the ICD-11 Content Model

In the Protégé implementation, an ICD category is represented as a class whose details are determined by a set of metaclasses. Each metaclass (e.g., a `ClinicalDescriptionSection` metaclass), groups a set of related properties (e.g., `body part`, `body system`, `signs and symptoms`, and `severity scale`) that an ICD category may have (Figure 3). By associating different metaclasses with an ICD category, we can flexibly specify different sets of properties with it. For example, external causes of injuries ICD categories do not have clinical description properties. Instead, they inherit descriptors—intent, mechanism of injury, place of occurrence, activity when injured, object or substance producing injury—that are more appropriate for them, and that are not relevant for disease-oriented ICD categories.

Each ICD category is related to *terms* that specify the detailed content of the category in the model. One subset of terms, including title, textual definition, inclusions, exclusions, and indices, are *linguistic terms*, which, in addition to their other attributes, have fields that are language-specific. A second subset of the terms—those from the Ontology Layer—are *reference terms* that must be specified using codes from external terminologies or ontologies.

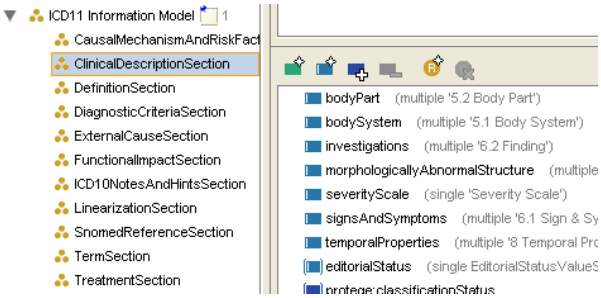

Figure 3. The ICD-11 Information Model. Metaclasses group related properties (e.g., body part, body system, signs and symptoms, and severity scale) into sections (e.g., `ClinicalDescriptionSection`).

Reference terms essentially represent coded information that expresses the meaning of a category in a computer-interpretable way. By contrast, *linguistic terms* are language-specific terms meant to help human users interpret the meanings of ICD categories. Each ICD category, therefore, will have multiple sets of linguistic terms. For example, separate titles and definitions—one for each of the supported languages—are needed.

Reference terms are also the means for coordinating related external classifications and terminologies with ICD-11. For example, the International Classification of Functioning, Disability and Health (ICF) divides the functional impact of a disease into domains such as cognition, mobility, self-care, interpersonal relations, and life activities. Correspondingly, the `FunctionImpactSection` metaclass stipulates that an ICD category may have an array of properties for encoding different aspects of a disease's functional impact. The array of properties are organized as a hierarchy of OWL object properties such that more specific properties have more specific ICF value sets than those of more general properties. For example, the `householdActivityImpact` property has as its value set ICF codes for household activities while the `lifedActivitiesImpact` property has as its range not only household activities, but also school and work activities.

For the external causes chapter we similarly define a set of descriptors as OWL object properties whose value sets correspond to different axes of ICECI.

For *reference terms*, it is intended that appropriate subsets of SNOMED-CT should be used as a source wherever possible, *e.g.* signs and symptoms, findings, possibly anatomical locations, and treatment.

The *Foundation layer*—the Category and Ontology layers taken together—is the locus of the clinical, scientific, and linguistic knowledge from which classifications that satisfy particular use cases will be derived. In the Foundation layer, an ICD category represents a health-related concept that is organized in a poly-hierarchy where the nodes are arranged by the principle of generalization and specialization. `Influenza due to identifiable influenza virus` (J10), for example, should be classified not only as a disease of the respiratory system, but also as a kind of infection in the Foundation layer.

Because the Foundation Layer is not constrained by the requirements of classification schemes, it is possible to map ICD categories to SNOMED-CT terms at a conceptual level and then to align the hierarchies of ICD categories and SNOMED-CT terms. In general, ICD categories are less granular than SNOMED-CT terms. An ICD category will often have one or more corresponding SNOMED-CT terms. With both ICD categories and SNOMED-CT expressible in OWL, it will be possible to develop more complex formal mappings between the two systems. Much detailed work remains to be done. WHO and the International Health Terminology Standards Development Organization (IHTSDO), the association in charge of developing SNOMED-CT, have signed an agreement to expedite coordinated development of and mapping between the two systems.

The *Linearization layer* is the locus of backwards compatibility with ICD classifications and adaptations of ICD11 to other user cases, *e.g.* Primary care. The requirement for the Linearization layer is that each classification (or “linearization”) should be able to be generated automatically from the information in the Foundation Layer. The basic information needed for deriving the mono-hierarchy of a linearization is described in the Use component of the Content Model, which specifies a triplet for each category: (1) the linearization in question (e.g., Mortality), (2) the linearization parent (when the category has multiple parents), and (3) a Boolean flag indicating whether the ICD category is in that linearization.

For each linearization, it's mandatory to indicate the linearization parent when a category has multiple parents, even if the category is not represent with a code in the linearization. The parent information is needed for computing the indices and exclusions of its ancestor that is part of the linearization.

The indices and exclusions may differ from linearization to linearization. For example, the codes in the primary care use case are much less granular than the morbidity or mortality codes. Therefore a specific index (e.g., influenza with pleural effusion, influenza virus identified) may be associated with different codes (e.g., J10.1 or J10) in different linearizations. The exclusions of an ICD category suggest other codes that should be used instead or in addition. The appropriate exclusions may also depend on linearizations.

Instead of associating linearization-dependent indices and exclusions with an ICD category, we want to compute them from some basic information in the Foundation layer. We define a *base index* of an ICD category as a linguistic term that is either an exact synonym or a narrower term that is not already an index to one of the ICD category's descendants. (The descendants of an ICD category in the Foundation layer are not required to cover exhaustively the meaning of the parent.) For exclusions, we want the them to be consistent across different linearizations. Therefore we define *base exclusions* if an ICD category as exclusions of the category in the most granular classification. The indices and exclusions of an ICD category in a particular linearization will be aggregated from the base indices and exclusions.

To compute the indices of categories in a linearization L, first consider a leaf node A of L. Find all "linearization descendants" of A that are not included in L (based on the linearization parent information). The union of their base indices and the base indices of A forms the indices of A in L.

For a non-leaf node B in L, create a residual child category R, collect all "linearization children" of B that are not included in L, and add their base indices as well as those of their descendants to the indices of the residual class R.

To compute the exclusion terms of an ICD category in L, use a similar algorithm, collecting exclusions from non-included descendants and adding them to leaf nodes or residual nodes. However, an additional step to adjust the codes associated with the exclusions is necessary. For example, `Essential Hypertension` (I10) has the exclusion "complicating pregnancy, childbirth and the puerperium (O10–O11, O13–O16)." If O10 and O14 are not part of the linearization, then we need to find the appropriate new codes to use in the exclusion. To do that, we follow the linearization parents of the categories not part of the linearization. If the most specific included linearization ancestor is a leaf node in the linearization (e.g., A), then use its code in the exclusion. If the most specific included linearization ancestor is a non-leaf node in the linearization (e.g., B), then use the code of residual class R that's below the non-leaf node.

The software for collaborative development, iCAT, is implemented on the basis of these methods and models,

using Collaborative Protégé and Web Protégé as foundation and the terminologies and ontologies available at the Bioportal [3] as the source of concept descriptors. The iCAT repository was initialized with a Start-Up List largely derived from ICD-10. During an intense two-week meeting in September/October 2009 ("iCAMP"), classification and informatics experts and representatives from each Top Advisory Group convened in Geneva to try out the ICD-11 alpha draft development process, to learn the ICD-11 Content Model and to evaluate the iCAT software. The results were highly positive in principle, but, as expected, the meeting generated a long list of desired refinements for both the Content Model and the iCAT software.

Discussion

The ICD-11 Content Model is very much a work in progress. Consensus formulation of several components, such as Temporal Properties, Severity Properties, and Diagnostic Criteria, are not yet available. New content elements, such as fully-specified names—a name that provides an unambiguous way to describe a concept and that is written in a grammatically and orthographically correct form suitable for natural language processing—are being proposed.

The ICD-11 Content Model is encoded in OWL supplemented by meta-constructs from Protege. Its usage of OWL is very different from that of other ontologies, such as those of the Open Biological and Biomedical Ontologies (OBO) Foundry [4]. It does not rely on an upper-level ontology such as the Basic Formal Ontology [5], nor does it attempts to characterize terms in the model using a set of fundamental relationships. This is delegated to the Ontology layer—expected to consist largely of subsets of SNOMED. The metaclasses that specify the properties of an ICD category define templates for entering data about an ICD category.

The use of metaclasses as templates for components of ICD categories makes the Foundation layer of ICD-11 an information model rather than an ontology, despite the use of OWL for some of its parts. The ICD category `Influenza due to identifiable influenza virus`, for example, has as its anatomical site a reference term that is associated with the SNOMED-CT term for lung. The reference term represents data about the ICD category, not a semantic restriction that a description logic reasoner can use as a logical axiom. Furthermore, as implemented in Protégé 3, the *values* of those properties that are inherited from a class's metaclasses are not inherited by subclasses of the class. Thus, for example, the reference term signifying *lung* as the anatomical site of the ICD influenza category is not inherited by its subclasses.

Nevertheless the ICD-11 Content Model provides the mechanism for a rudimentary form of post-coordination. This is most clearly seen in the description of the external causes of injury, where orthogonal descriptors, such as intent, mechanism of injury, place of occurrence, and object or substance producing injury, are needed to fully specify dimensions of the cause of injury. Pre-coordinating all possible combinations of the axes will produce a classification that is too unwieldy to be usable.

The use of the Content Model to revise ICD is the first try of the WHO methodology to develop and maintain international classifications. If successful, it will be used for revising classifications such as as ICF (International Classification of Function), ICHI (International Classification of Health Interventions) and ICPS (International Classification of Patient Safety).

Conclusion

The ICD-11 Content Model allows the creation of an ICD Foundation Layer where clinical, scientific, and linguistic knowledge about ICD categories are systematically represented. It clarifies the relationship of ICD to a number of classification systems and to SNOMED-CT. Using the information specified in the Content Model, we can generate alternative linearizations—specialized classifications adapted to satisfy different use cases. The ultimate goal is Web-based software that allows wide participation in an expanded and enriched revision of the ICD.

References

1. World Health Organization. Production of ICD-11: The overall revision process, 2007. http://www.who.int/classifications/icd/ICDRevision.pdf.
2. Tudorache, T, Falconer, S., Nyulas, C, Storey, M, Musen, MA, Supporting the collaborative authoring of ICD-11 with WebProtégé, Proc AMIA Symp, 2010. submitted.
3. Noy NF, Shah NH, Whetzel PL, Dai B, Dorf M, Griffith N, et al. BioPortal: ontologies and integrated data resources at the click of a mouse. Nucleic Acids Res 2009 Jul 1;37(Web Server issue):W170-3.
4. Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol 2007;25(11):1251-5.
5. Grenon P, Smith B, Goldberg L. Biodynamic Ontology: Applying BFO in the Biomedical Domain. In: Pisanelli DM, editor. Ontologies in Medicine. Amsterdam: IOS Press; 2004. p. 20-38.