Terminology Services: Standard Terminologies to Control Health Vocabulary
Experience at the Hospital Italiano de Buenos Aires

Fernán González Bernaldo de Quirós, Carlos Otero, Daniel Luna
Department of Health Informatics, Hospital Italiano de Buenos Aires, Buenos Aires, Argentina

Summary
Healthcare Information Systems should capture clinical data in a structured and preferably coded format. This is crucial for data exchange between health information systems, epidemiological analysis, quality and research, clinical decision support systems, administrative functions, among others. Structured data entry is an obstacle for the usability of electronic health record (EHR) applications and their acceptance by physicians who prefer to document patient EHRs using “free text”. Natural language allows for rich expressiveness but at the same time is ambiguous; it has great dependence on context and uses jargon and acronyms. Although much progress has been made in knowledge and natural language processing techniques, the result is not yet satisfactory enough for the use of free text in all dimensions of clinical documentation. In order to address the trade-off between capturing data with free text and at the same time coding data for computer processing, numerous terminological systems for the systematic recording of clinical data have been developed. The purpose of terminology services consists of representing facts that happen in the real world through database management in order to allow for semantic interoperability and computerized applications. These systems interrelate concepts of a particular domain and provide references to related terms with standards codes. In this way, standard terminologies allow the creation of a controlled medical vocabulary, making terminology services a fundamental component for health data management in the healthcare environment. The Hospital Italiano de Buenos Aires has been working in the development of its own terminology server. This work describes its experience in the field.

Introduction
In the last few decades, health care stakeholders and policy makers around the world have emphasized the importance of establishing electronic health records (EHRs) for all health care institutions. Their goals for doing so include increasing patient safety, reducing medical errors, improving efficiency, and reducing costs [1, 2]. Medical knowledge is at the core of the EHR, and is central to meeting these goals. To represent medical knowledge, it is necessary to represent patient’s data from different sources including, among others, problem lists, progress notes, procedures, medication list, laboratory and complementary test results, social determinants of health, environmental information, people’s decisions about health and medical treatments, and genomics and proteomics data. From the healthcare provider’s point of view, data entry is an obstacle to the adoption and effective use of EHRs. Providers prefer to document healthcare findings, processes, and outcomes using free text in natural language [3]. A narrative format allows them to share complex ideas in an efficient and effortless manner.

As a result, ambiguities must be resolved and vocabulary must be standardized. To reach this goal, an EHR should capture clinical data in a structured and preferably coded format.

Codify is the process to reduce a concept into a code [5]. Codes are usually numeric or alphanumeric. In order to represent facts that happen in the real world to be managed in a database the need of a standard codification system (SCS) arise. Evans et al. stated that the medical community required a “common, uniform, and comprehensive approach to the representation of medical information” [6]. Codification should be one-to-one: one term should only exist for a given object. Each term should describe only one object. The aim is to avoid ambiguity through polysemy or homonymy [5]. This, in health context, must be performed using international standards like classifications (ICD 10, CPC, LOINC, etc.) or terminologies (SNOMED, SNOMED-CT) [7, 8].

After the publication of Cimino’s Decisional Data [9], the difference between terminology systems like SNOMED-CT and classification systems like ICD-10-CM/PCS became clearer. Both coding schemes provide the data structure needed to support healthcare clinical and administrative processes. Clinical terminology systems as well as clinical classification systems were originally designed to serve different purposes and different users’ requirements [10]. ICD-10 is a classification system designed as an output for general reporting purposes such as, for instance, public health surveillance. On the contrary, SNOMED-...
CT is a clinical terminology developed as a standard data infrastructure for clinical application. For this reason, it requires a higher degree of specificity [11].

According to the International Standards Organization (ISO), terminologies should be made of formal aggregations of language-independent concepts; those concepts being represented by one favored term and appropriate synonymous terms, and of explicitly represented relationships among the concepts [12, 13]. The use of standard terminologies in an EHR is useful to enable decision support systems [8, 14, 15], to exchange data between health information systems, for epidemiological analysis, for research to support health services research, and to manage administrative tasks, among others. While many terminologies have been developed, no single terminology has been accepted as a universal standard for the representation of clinical concepts. By contrast, individual terminologies or components have been identified by standards organizations as candidates for specific uses [16].

Terminologies

The ISO specification stated that terminologies must define their purpose and scope, quantify the extent of their domain coverage, provide mappings with external terminologies designed for classification, and support administrative functions [12, 13]. The ISO also highlighted the value of mapping different terminologies designed to meet different needs. This would allow, for example, a physician to choose a concept from a clinically-oriented terminology to construct a patient’s problem list and automatically get a mapped concept from a classification (like ICD-9-CM) for billing purposes [12, 13]. In 1998, Cimino summarized the work of different research teams toward the definition of the precise attributes of a multipurpose and shareable terminology [9, 14]. He stressed the value of “concept orientation” pending terminology construction. Concept orientation implies “...to use concepts as basic building blocks ahead words, terms, or phrases”.

It allows a terminology to be useful in several situations, represented in different languages, and easily accessed for quality [17]. According to Cimino, the aim was to have a universal single clinical terminology that would completely cover a specialty domain’s concepts at multiple levels of detail. Non-specific phrases such as “not elsewhere classified” must be avoided [9, 14]. It is important to point out the need for complete and comprehensive domain coverage using non-ambiguous, non-overlapping concepts. In the absence of complete domain coverage, terminologies should integrate other terminologies. Terminologies need to support synonymy and compositionality [18]. “High quality vocabulary” has been defined as the vocabulary that approaches completeness. It is well organized and made of terms which meaning is clear [9, 14].

Clinical Reference Terminologies

A reference terminology is defined as a set of concepts and relationships that provide a common reference point for comparison and aggregation of data about the entire health-care process, recorded by multiple different individuals, systems, or institutions [19]. Cornet et al. [20] defined it as “...a system of concepts with assigned identifiers and human language terms, typically involving some kind of semantic hierarchy. Some systems may support the assignment of multiple terms, or synonyms, to a given concept...”.

SNOMED-CT as a Clinical Reference Terminology

SNOMED-CT was developed to serve as a standard data infrastructure for clinical applications which require a greater degree of specificity [21-23]. In order to achieve “domain coverage”, terminology developers have created new concepts by the utilization of two methods: pre-coordination and post-coordination. With pre-coordination, also named enumeration, it is possible to model suitable levels of detail with distinct concepts, derived from real world. Generally, only clinically meaningful concepts are pre-coordinated [24]. By contrast, with post-coordination, also called compositionality, complex concepts can be composed from simple concepts [16]. Pre-coordination and post-coordination can complement each other, with pre-coordination providing logics and complexity and post-coordination allowing for expressivity and more complete domain coverage. Existing terminologies that allow post-coordination are more capable to represent phrases and concepts extracted from clinical documents as compare to pre-coordinated terminologies [25]. Such terminologies may improve terminology domain coverage because users can both access existing concepts and dynamically compose new concepts according to their needs.

SNOMED-CT provides a unified language: it may be used as a standard for communication among healthcare providers. It also highly promotes semantic interoperability in healthcare information systems [26-28]. Its standardized logical structure and its wide acceptance make it more appropriate for high-level information exchange at national and also international levels [26-28]. SNOMED-CT not only supports pre-coordination and post-coordination but it also includes several descriptions that can be used as an “entry terminology”. Finally, SNOMED-CT has a standard cross mapping model. The official distribution includes data for ICD-9 mapping. ICD-10 cross mapping has also been developed. These mappings provide the aggregate terminology features to SNOMED-CT [26].

Interface Terminology

Interface terminologies, also called colloquial terminologies, application terminologies and entry terminologies, have been defined as systematic collections of healthcare-related phrases (terms) that support clinicians’ entry of patient-related information into computer programs [16]. But how does it work? When healthcare providers enter some patient information into the EHR, the interface terminology links free text patient descriptors to structured, coded internal data elements used by specific clinical computer programs. Interface terminologies also facilitate the display of recorded patient information to clinical users as simple human readable text [16]. These terminologies generally embody a rich set of flexible, user-friendly
phrases displayed in the graphical or textual interfaces of specific computer programs.

These entry terminologies allow users to interact easily with concepts through common colloquial terms and synonyms. Entered terms can then be mapped to concepts explicitly defined in a more formal terminology, such as a reference terminology, which can then define relationships among concepts [29]. EHRs rely on interface terminologies for successful implementation in clinical settings since such terminologies provide the translation from clinicians’ own natural language expressions into the more structured representations required by computer applications [16]. Interface terminologies are crucial to foment direct categorical data entry by physicians in EHRs.

Among the aims of interface terminology, we can mention that they: 1) provide an institutional vocabulary for all user interfaces allowing users to interact with known terms, including local jargon and preferences and 2) provide concept lookup functions with loose lexical matches and options to be used to enter new items in a problems list or similar user interfaces. It is also important to provide short pick-lists of definitions for more structured data entry in specific use templates, with a short list of valid entries and different preferred terms for the same concept in different settings. It should include the ability to accept new terms from the user, in case a concept or description is not represented, and to detect inappropriate terms because they are too general or not valid in a subset [30].

The usability of an interface terminology refers to the ease with which its users can accomplish their intended tasks when using it. In addition, it has been demonstrated that interface terminology usability correlates with the presence of attributes that enhance the efficiency of term selection and composition [31, 32]. The usability of a clinical interface terminology correlates with the presence of relevant insertional medical knowledge, adequacy of synonyms, balance between pre-coordination and post-coordination, and mapping with terminologies by means of formal concept representations. Synonymy refers to the number of individual terms that can correctly represent a unique concept. Synonym types may include alternate phrases, acronyms, definitional phrases, and eponyms [33]. Interface terminologies also enhance their usability by decreasing the number of steps required for users to find or compose the terms needed for a given task [24, 33].

### Terminology Services

Many definitions exist for terminology services. We define terminology services as complex systems offering a conceptual representation of medical knowledge with relationships between concepts, external representations of concepts in lists of standard terms (classifications), and lexical tools that facilitate the search for terms [34].

A terminology server (TS) is a software platform made up of a local interface vocabulary modeled with a reference terminology which is trans-codified with clinical classifications (Figure 1). The TS should also provide interactive information for refining concepts. This feature of the TS is achieved using semantic information included in SNOMED-CT while navigating the sub-types/super-types hierarchies [35]. On the desiderata for TS, Chute et al. [29] attempted to articulate the functional needs of a terminology server oriented toward the clinical needs of care providers using applications in an operational environment. The desirable characteristics for a terminology server include word normalization, word completion, target terminology specification, spelling correction, lexical matching, term completion, semantic locality, term composition, and term decomposition.

### The Hospital Italiano de Buenos Aires: Terminology Services Experience

Our aim was to design a new terminology system, which objectives can be related to the functions of the terminology system previously described (entry, reference, and aggregate terminology). In 1998, the terminology work team started a centralized
secondary coding scheme, where a small number of trained persons coded the narrative text recorded by physicians while taking care of patients. The coding scheme included problem lists, diagnoses, and procedures [36]. We stored all free text documents for five years and processed them from problem lists in order to build the first thesaurus.

In 2004, we achieved one million secondary coded narrative texts. This gave us the chance to optimize the recognition of terms when users looked up for a concept in our thesaurus. As an example, the server contains wide recognition of synonyms (there are for example 140 ways to describe arterial hypertension) and more than 8,000 terms in Spanish language that do not appear in the Royal Academy of the Spanish Language (only in the problem list domain). Today, Hospital Italiano’s terminology server has 6,692,916 unique descriptions and 546,522 health concepts and it “learns” every day while users interact with the service; 9.72% of those descriptions are non-valid terms (for example, clinical appointment as a problem in the problem list). Recognized non-valid concepts block the input of that kind of data in the system and improve the overall quality. We are still working on enriching thesaurus descriptions and adding social determinants of health, social activities for promoting and preventing health, information on behavioral changes and people decisions and preferences [37], familiar and other social relationships that are important for health, social assets, and relevant genetic information.

In 2010, we started providing remote terminology services (RTSs) through a transnational and inter-institutional implementation [38]. In 2011, we started with the extraction of the largest amount of clinical information from the system existing in a national provider network (mostly in free text), and added this information after it was coded to the new clinical data repository. For this purpose, extracted data were processed by RTSs and coded when possible. Data included allergies, reason(s) for clinical encounters, habits, risk factors, symptoms and diagnosis entered by physicians in a free text form, and coded diagnoses when physicians felt it was particularly necessary. Using a batch processing of these data, RTSs initially recognized and auto-coded 11,118,760 (78.74%) texts (including valid and not valid text), and did not recognized 3,001,991 (21.26%) of the original data [38]. In 2012, we began developing natural language processing tools and extended terminological services to drugs domain, diagnosis tests, and medical procedures. In 2014, in the context of an accreditation process conducted by the Joint Commission International (JCI), we implemented an in-house developed software tool for the synchronous disambiguation of acronyms in EHRs [39, 40]. After all these enrichments and accumulated experiences, RTSs currently recognize up to 90% of new uploaded text.

We currently provide terminology services to several healthcare organizations in countries of Latin America, such as Argentina, Chile, and Uruguay. These include a thesaurus tailored to the local needs and the jargon of the professionals who interact with EHRs, SNOMED-CT as the reference standard for interoperability and clinical decision support system (CDSS) implementation, cross maps to ICD-9CM (diagnosis & procedures), ICD-10 (diagnosis); ICPC-2 (diagnosis), ATC (Anatomical Therapeutic Chemical Classification) (drugs), the creation of different types of refsets according to the needs of the organization and the drug composition service modeled after the UK’s dm+d model. The interface terminology is based on the use of SNOMED-CT, which is used as the reference terminology. In this sense SNOMED-CT serves as a uniform back-end representation allowing our interface terminology to adapt to the local needs of the organization for which we provide services [40-42].

A very frequently asked question is why using interface terminologies instead of only SNOMED-CT? Among the reasons why we made this choice, we can name:

- It is simpler for end users;
- When a single concept is not sufficient to define the information, it is possible to build a new concept using post-coordination, this new concept being understood as the combination of two or more SNOMED-CT concept identifiers under SNOMED rules;
- The thesaurus allows the management of synonyms (different descriptions related to a concept), lists of valid and not recognized terms (typos, …etc.), validated jargon and acronyms, list of “Not Valid” terms;
- The thesaurus is expanding in a continuous learning process and allows for drug composition information (commercial products) [30].

The Hospital Italiano’s Reference Terminology: Functions and System Description

Regarding reference terminology functions, our TS allows the entry terminology to be represented in the reference terminology (SNOMED Spanish Language Version). New concepts can be created for institutional terms that cannot be represented by standard SNOMED-CT codes. The system also provides tools to take advantage of the knowledge stored in SNOMED-CT relationships, like obtaining more refined or more general terms, and updating new versions of SNOMED-CT without losing information. We used SNOMED Spanish Language Version as the reference terminology, but it is important to note that all different language versions of SNOMED-CT share the same concepts and relationships. During the translation process only, new descriptions may be added. Both entry and reference terminologies were stored following the SNOMED data model and using SNOMED tools to represent the concepts of the entry terminology.

SNOMED-CT defines concepts by their relationships with other concepts, so we created new relationships as part of our SNOMED-CT extension. SNOMED-CT has around 300,000 concepts, but in a clinical setting, health professionals usually use very detailed expressions, adding modifiers to general concepts, like “mild ankle sprain”. Any new concept can be represented using this post-coordination technique, creating more detailed subtypes of existing SNOMED-CT concepts. Around 33% of the concepts included in the Problems List subset could be directly mapped with existing SNOMED-CT concepts; the other 67% needed the addition of one or more modifiers (post-coordination) in order to fully represent the meaning of the entry.
terminology concept. This rate of post-coordination was dictated by a very permissive policy allowing the use of any term requested by the users, often very specific or personalized. This decision facilitates user acceptability but adds complexity in semantic interoperability. In each subset, professionals usually try to enter terms that are not valid for later use. We would rather have clinicians record the proper diagnosis or the reason for encounter. In order to reject these terms in the context of the administration of invalid terms, we tag them and add textual information so that the professional may understand the coding guidelines of the institution. This module provides the tools for tagging terms and editing the information.

The Hospital Italiano’s Aggregate Terminology: Functions and System Description

Our TS provides outputs to several standard classifications (ICD-9, ICD-10, LOINC, ICPC-2, ATC), local billing nomenclatures, and data aggregated according to SNOMED-CT hierarchies and DRG (diagnosis related group) grouping. All these functions run on a centralized software platform and data structure. The TS provides these functions to all existing applications in the Health Information System in the form of Web Services. A terminology maintenance software application has now to be developed to administer the institutional terminology, its relationships with SNOMED-CT, and the mappings. The official SNOMED cross maps model is implemented, a multi-classification interface has been created as part of the terminology maintenance software to visualize, test, and modify mappings from SNOMED-CT to different classifications. An SQL query was designed and implemented in the relational database management system Oracle to aggregate concepts according to knowledge stored in SNOMED-CT relationships, like all kinds of diabetes, including diabetes complications but excluding maternal and neonatal diseases. Queries are maintained by a specific module of the terminology maintenance software.

The coding application requests that the appropriate classification to be used to code the terms in the EHR is selected from the list of classifications available. The system then assigns the code to each term. Using this mechanism, it is possible to select ICPC-2 for the epidemiological analysis of a problem list of outpatient EHRs, ICD-9 and ICD-10 for the discharge summary of inpatient EHRs. This mapping is possible because we use the official cross-match offered by our reference terminology (SNOMED-CT) or because our specific terminology team created our own mapping.

We found that SNOMED-CT cross mapping to ICD-9 is still not adequate for clinical use in our setting and requires additional manual work. This may be caused by a different use of the classification in Argentina as compared to the United States. Our clinical data extraction process, using rules based on SNOMED-CT knowledge data, is very effective. However, these rules should be revised for each new SNOMED-CT version as changes in hierarchies and models may affect their effectiveness.

Conclusion

A terminology server is useful to help healthcare providers represent what “happens” with their patients in “real time” using standard terminologies and to allow the health information system manage information for several purposes in smooth interoperability with other systems. Developing our own terminology server gave us the chance to offer a growing dataset of coded information based on international standards in a friendly interface for our users, thus improving data documentation quality. Creating and maintaining a sharable Spanish interface vocabulary database between different countries is a big task as medical Spanish language is a rich vocabulary and there are different ways of naming the same clinical entities (polysemy), and there are different acronyms and synonyms between countries. There are as many expressions in Spanish languages as there are countries and regions of a given country. There is still much work to do in order to control the representation all the needed medical knowledge. In addition,
dimensions such as social determinants of health, social activities for promoting and preventing health, behavioral changes, people decisions and preferences, family and other social relationships that are important for personal health, social assets, and relevant genetic information, among others, need to be better represented and codified. We know that offering semantic control with a terminology server improves the granularity and quality of information, the adaptability to local culture, lexical variants and priorities, and thus increases end users acceptability. Build it as a service, the terminology server allows good scalability and sustainability by reusing the efforts of the system knowledge in a “continuous learning process”. Because of all that, we consider that providing services to other organizations in a regional approach is of great value.

References


5. Health Informatics. Hannah KJ, Ball MJ, editors (Formerly Computers in Health Care).


18. Chute CG, Cohn SP, Campbell KE, Oliver DE, Campbell JR. The Content Coverage of Clinical Terminology to a Classification. AHIMA’s 75th Anniversary National Conference; 2015; Available from: http://library.ahima.org/PB/ControlledHealthTerminology-StructureandHighLevelIndicators worldviewp.pdf?


40. Williams M del PA, Perez D. Building an Snomed CT Content Extension with abbreviation support. SNOMED-CT Expo 2015.


Correspondence to:
Fernán González Bernaldo de Quirós
Hospital Italiano de Buenos Aires
Gascon 450 (1181) CABA
Argentina
E-mail: fernan.quiros@hospitalitaliano.org.ar