MEDINFO 2007
HISA Ltd Australia
© 2007 The authors. All rights reserved.

Semantic Annotation of Patient Data in a Commercial Health Information System

Isabel Barth[a]; Sven Tiffe[a]; Michael Dahlweid[b]

[a] *Agfa HealthCare, Health Care Application Platform, Metatools, Germany,*
[b] *Agfa HealthCare, Product Management, Germany*

Abstract and objective

To facilitate reusability of information in a complex Health Information System, a semantic component has been created that allows for formal descriptions of clinical data at a very granular level. Through an extension to the clinical database, semantic concepts selected from controlled vocabularies residing on a terminology server can be directly linked to the patient data. The use of specifications such as LexGrid and CTS allows for cross-terminological information retrieval and interinstitutional exchange of clinical data.

Keywords:

terminology, semantics,
clinical decision support systems

Introduction

Understanding the semantics of a healthcare provider's data is crucial for use cases where this information has to be interpreted, such as for applying clinical decision support systems (CDSS). Such systems need to operate on formal information from controlled vocabularies whenever reusability of the knowledge base is required [1].

We added a semantic extension that operates on controlled vocabularies to Orbis, a hospital information system with more than 1200 installations in Europe. In contrast to other systems that are based on a central data model, Orbis allows its users to define an arbitrary amount of data structures during runtime by providing a visual tool for the definition of forms. Such forms have a graphical representation and are stored in a generic, entity-attribute-value (EAV) based data model. Through this mechanism, users can add implicit data structures that meet their needs in terms of documentation. The classification of such customized data structures necessary for use by CDSS is a major challenge that can be addressed by applying formal codes to the EAV model.

Methods

Large parts of Orbis' electronic patient record (EPR) consist of EAV structures in a generic data model [2]. These structures are termed 'forms' and are designed by administrative users. Later on, these form definitions can be instantiated by clinical users of the system during runtime ("form instances"). The flexibility of this approach empowers clinicians to define arbitrary forms which meet their specific documentation needs, instead of being limited to a central data model. The set of input fields that constitute a form represent a data structure which can be accessed within Orbis like a database table. Thus, in this context, a form is more than a user interface representation, since it also corresponds to a virtual data structure.

This openness implies that there are multiple ways to document the same clinical fact. For example, the administrative user who is required to offer the documentation of a patient's childhood illnesses in the scope of an anamnesis form is free to define a text field expecting free text or a select box offering a set of predefined values. Alternatively, the information request might be modeled with checkboxes or radio buttons. At runtime, the clinical user may document a childhood illness by selecting a checkbox (figure 1) or by filling a free text field (figure 2).

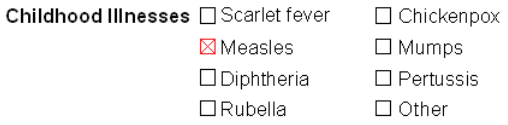

Figure 1 - Representation of infantile diseases as a set of checkboxes

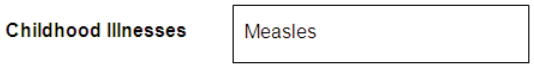

Figure 2 - Representation of infantile diseases as a free text field

To describe the semantics of such a form structure, each of its elements can be annotated by one or more formal semantic concepts. During design time, this annotation is performed manually by the form author or a coding expert and describes the attribute of the element, such as *'systolic blood pressure'*, *'allergy'* or *'infantile disease'*.

In the previous example, there are several possibilities to annotate the "measles" input fields of the form definition with a SNOMED code. One possibility is:

[14189004 Measles – disorder – Clinical findings].

Since the larger context is the patient's medical history, this code is incorrect, because the SNOMED code denotes an acute illness. Considering this, the following SNOMED code could be more appropriate:

[161419000 History of Measles – situation – Context dependent categories]

One important factor to reinforce valid coding is to consider the context clinical information is embedded in. To the clinical user, much of this context is obvious and presents itself in form of the document's graphical structuring. If 'context clues', such as the semantics of a field's label or of the embedding form, are not available during the annotation process, it can become impossible to deduce the correct code.

All formal concepts that describe these semantics are stored in the same relational database as Orbis' EPR and are linked to the attribute and value entries through foreign keys. Based on these links, queries have been added to Orbis that retrieve all patient information for a specific medical case whose semantic annotation equals the search criterion that is also a formal semantic concept. For example, querying 'allergy' for a specific patient, retrieves all form instance data whose semantic annotation corresponds to allergies. In the last example, the result would be a Boolean and a free text from a technical point of view.

Thus, during runtime, *values* entered by the clinical user should also be formalized if possible. Both user support through semi-automatic machine-aided selection of appropriate terminology items and fully automatic semantic tagging of medical terms in free text by natural language processing algorithms are possible. By this, the result of the query not only consists of the technical data types, such as "Boolean" or "text" but also of their semantic meaning as a formal code. Another special kind of information are numeric values, such as *'blood pressure'*. Queries on these elements retrieve both the stored values and their semantic type, such as *80* and *'systolic blood pressure (mmHg)'*. To specify units of numeric values, post-coordinated concepts are used.

The underlying data model for semantic annotation is open for all kinds of vocabularies and terminologies. For maintaining them and searching adequate concepts during the annotation process, we used an external terminology server. For this purpose, LexGrid [3], a data model and software developed by Mayo Clinics which provides a common data format has been has been connected using the HL7 Common Terminology Services interface. LexGrid offers tools to convert terminologies of different formats to the LexGrid database structure. For example, an attribute of an EPR annotated by a LOINC code could be retrieved, although the query searches for a synonymous SNOMED code.

Such queries have been applied to one of Orbis' rule based decision support systems that is founded on Arden Syntax for Medical Logic Modules (MLMs). Here, extensions in the data retrieval part of the rules allow to access clinical information identified by formal codes. The current implementation is limited to using formal codes in the rules that exactly match those used in the annotation of the EAV structures, but is open to future cross-terminological queries. In this way, information stored in arbitrary customized data structures can be retrieved by a single rule.

Results

By linking formal concepts to atomic data elements of Orbis' EPR we have been able to describe patient data semantics at a very granular level.

The combination of LexGrid and the NCI Meta Thesaurus empowered us to carry out cross-terminological and cross-lingual searches.

Instead of defining queries directly on the EPR, data retrieval works on an abstract, semantic level that focuses on individual concepts and follows their links to the EAV entries.

Discussion

However, the application of this method does have certain limitations when applied to compound concepts such as *'blood pressure'*, which consist of a set of individual EAV entries. Whether this aggregation can best be achieved by post-coordination or through an additional abstract layer that defines compounds or clinical archetypes [4], requires further evaluation.

Conclusion

Our bottom-up approach to semantic annotation ensures maximum reuse of the database's inherent knowledge by avoiding premature reduction of primary data as a side-effect of summarization. Thus, through the separation of annotation process from abstraction, the knowledge base is open for interpretation from various perspectives arising form different requirements. Further layers of semantic abstraction may allow the interpretation of the same knowledge from different views.

Semantic annotation has to be carried out in a way that ensures the retrievability of the documented data. The success of unambiguous annotation is dependent on the form element's data type and the form items' larger context.

Using the generic LexGrid data model for the administration of clinical vocabularies and a standardized terminology server interface, the gap caused by a proliferation of terminologies in different data formats can not

only be bridged, but the terminologies can also be shared among different institutions.

References

[1] Jenders, Robert A. and Sailors, M. (2004). Convergence on a Standard for Representing Clinical Guidelines. MEDINFO 2004: 130-134.

[2] Lenz, Richard (2001). A Practical Approach to Process Support in Health Information Systems. University of Marburg.

[3] http://informatics.mayo.edu/LexGrid. Accessed May 21, 2007.

[4] T. Beale (2001). Archetypes, Constraint-based Domain Models for Future-proof Information Systems. Available: http://www.deepthought.com.au/it/archetypes/archetypes.pdf

Address for correspondence

Isabel Barth
Agfa HealthCare
Monaiser Str. 11, D-54294 Trier, Germany
Email: Isabel.barth@agfa.com