MEDINFO 2004
M. Fieschi et al. (Eds)
Amsterdam: IOS Press

Problems and Solutions with Integrating Terminologies into Evolving Knowledge Bases

Kurt L. Rickard, José L. V. Mejino Jr., Richard F. Martin, Augusto V. Agoncillo, Cornelius Rosse

Structural Informatics Group, Departments of Biological Structure, and Medical Education and Biomedical Informatics, University of Washington, Seattle, WA 98195

Abstract

We have merged two established anatomical terminologies with an evolving ontology of biological structure: the Foundational Model of Anatomy. We describe the problems we have encountered and the solutions we have developed. We believe that both the problems and solutions generalize to the integration of any legacy terminology with a disciplined ontology within the same domain.

Keywords:

Terminology integration, Knowledge representation, Ontology, Anatomy, Neuroanatomy.

Introduction

Emerging ontologies in the field of bioinformatics create abstractions for representing the relatively new disciplines of the biomedical sciences. This information, however, needs to be placed in the context of existing knowledge at higher levels of biological organization. In the discipline of anatomy, term lists have been in use for over a century and other computer-based term lists have been made available more recently. Since these traditional resources enjoy wide use, it poses a challenge to integrate them into a comprehensive ontology that has been implemented in accord with declared principles.

In this paper we explore the challenges and solutions encountered in the integration of *Terminologia Anatomica* [1] and *NeuroNames* [2] into the *Foundational Model of Anatomy* [3]. We start by describing these resources, give a brief account of the software we developed for merging the terminologies, describe the problems we needed to solve and explain the modifications we introduced into the FMA in order to accomplish the merge.

Anatomy Terminologies

Terminologia Anatomica (hereafter *Terminologia*) is universally accepted as the official anatomical terminology. Compiled by the international Federative Committee on Anatomical Terminology (FCAT), it is the revised form of *Nomina Anatomica*, first published in Latin over 100 years ago with the objective of standardizing the usage of anatomical terms. *Terminologia* is distributed in hard copy tabular form. Each of its Latin terms and its English equivalent is associated with a hard coded alphanumeric identifier. Inconsistencies in the semantic organization of its more than 10,000 terms relating to macroscopic anatomy have been discussed elsewhere [4].

NeuroNames is a structured neuroanatomical vocabulary pertaining to the brain. *Nomina Anatomica* was one of the principal sources for its more than 6500 neuroanatomical terms, 4000 of which are synonyms. It is implemented as a part hierarchy of brain subvolumes navigable only by computer.

The *Foundational Model of Anatomy* (FMA) is an evolving ontology that represents declarative knowledge about the physical organization of the human body [3]. Its concept domain encompasses anatomical entities ranging from the macroscopic to cellular, subcellular and macromolecular levels of biological organization. Key among the declared organizing principles of the FMA are a strictly structural context of modeling, a class subsumption hierarchy (Anatomy taxonomy) established on the basis of inheritable structural attributes, and the explicit definition of concepts and their relationships. The FMA currently contains approximately 70,000 concepts, which are associated with more than 110,000 terms and over 1.6 million instantiations of 170 kinds of relationships implemented in the frame-based Protégé-2000 knowledge acquisition system.

Our intent with the FMA is to make anatomical information available in a machine-understandable form that generalizes to all application domains of anatomy. Therefore, rather than attempting to standardize terminology, we want to include in the FMA all terms that currently designate anatomical concepts in order to facilitate navigation of the FMA by any user. This provides our motivation for the current project and for aligning the FMA with other ontologies [5].

Term Integration Software

Hard copy terminologies present a particular problem. We were, however, able to obtain a text file for most of the contents of *Terminologia* from its publisher and populated a MySQL database with the contents of the file. Some annotations of the hard copy text, however, were lost. The *NeuroNames* data were extracted from the compact disc accompanying the text [2] and placed in a separate database. These databases are accessed by the MySQL database server, which in turn is accessed by the Foundational Model of Anatomy Term Integration software. The FMA Term Integrator application is a computer program developed in-house for the incorporation of external data sets, such as *Terminologia* and *NeuroNames*, into the FMA. The software is

written in the Java™ programming language and uses the Java Database Connectivity (JDBC) API and SQL to access, query and update database entries as required.

The FMA Term Integrator semi-automates the data entry process allowing an author of the FMA to retrieve terms from the appropriate database. It then automatically instantiates new valid instances of the required anatomical concepts within the FMA and populates the appropriate slots in the concept frame with the values it reads from the database, while tracking which terms have already been processed.

Modeling Anatomical Terminologies in the FMA

Although there are numerous technical issues involved in the integration of terminologies into a knowledge base such as the FMA, the more interesting challenges are the conceptual difficulties. These difficulties must be resolved in terms of the FMA's semantic and implementation framework. Therefore we begin with summarizing the representation of terms in the FMA before addressing the challenges.

Representation of Terms in the FMA

Unlike *Terminologia* and *NeuroNames*, the FMA represents anatomical concepts rather than anatomical terms. Each concept has a randomly assigned unique numerical identifier (UW-DAID; University of Washington Digital Anatomist Identifier) and is associated with one or more terms. One of these terms is designated as the preferred name of the concept; other terms are synonyms. Each term is created as an instance of the class `Concept name`. Instances of `Concept name` have associated with them various metadata that describe the attributes of the term. For example, the source authority for each instance of `Concept name`, the date that the term was entered and by whom, the corresponding UMLS identification number (if it has one) and so on, are all contained in the frame of the term. Figure 1 shows a screen capture of the Protégé knowledge base API, including the `Concept name` dialog for 'Intercostal nerves', which displays all the information associated with this term. Each datum is contained in a slot and the value it takes is the slot's facet. In Figure 1, for the term 'Intercostal nerves', the 'Authority' is a slot which contains the string value 'Terminologia Anatomica 1998'.

The disciplined approach used by the Protégé knowledge representation system has allowed the FMA to be developed as a large-scale, robust and detailed model. Concepts entered into the FMA are to be explicitly defined. Unlike in *Terminologia*, no homonyms are allowed (i.e., each term must be unique), and a consistent naming convention is used throughout. This rigorous framework provides for the unambiguous modeling of concepts and complex relationships, in ways that less well principled terminologies do not support.

Merging Terminologies with the FMA

Reconciling the implied semantic structure of existing terminologies with a principled ontology presents challenges unique to each pair of sources. But when integration is limited to terms within the same domain of discourse (e.g., anatomy), we hypothesized that the problems and solutions will generalize to any pair of terminologies. We found this to be the case for both *Terminologia* and *NeuroNames* with respect to the FMA. Both contain anatomical terms that either match or do not match existing terms in the FMA; these terms are either in English or in a foreign language; they may be eponyms or abbreviations, they may be spelled differently; a code may or may not be associated with them; they may or may not refer to concepts already present in the FMA; some terms may be plural or conjunctions; other terms or codes may denote more than one concept; some terms may be outdated or inappropriate; and if accommodation of all terms in the ontology is a goal, the ontology may need to be modified. We believe, these are the problems likely to be encountered in matching any terminology to an ontology. Hence, the solutions we developed for integrating the terms of *Terminologia* and *NeuroNames* in the FMA will likely generalize to merging the terms of any terminology with a principled ontology provided both sources pertain to the same domain.

Matching Strings

In the FMA, English terms are used as preferred names of anatomical concepts. Therefore, the incorporation of the Terminologia and NeuroNames datasets into the FMA was done on the basis of a string match between English terms. Where there was a direct match between a term in the FMA and Terminologia, the Terminologia identification code for that term was inserted as a value in the 'TA ID' slot of the `Concept name` frame. NeuroNames does not include numerical codes. A perfect match also led to entering either 'Terminologia Anatomica 1998' or 'NeuroNames 2000' as a value of the 'Authority' slot of the term frame. If a term appeared in both terminologies, both terminologies were referenced as authority sources.

In cases where a term from *Terminologia* or *NeuroNames* did not exist in the FMA but the corresponding concept was present, the term was simply added to the 'Synonyms' slot for that concept. If the concept had not yet been modeled in the FMA, a new concept was created with the appropriate superclass and metaclass, and the *Terminologia* or *NeuroNames* English term was entered as the FMA preferred name. Plural terms presented a particular and challenging example in such situations.

Plural Terms

All of the terms in the FMA represent the singular form of an anatomical entity and conjunctions are not allowed. Even where more than one of a particular structure exists, the singular form is used. For example, although humans have two lungs, the term 'Lungs' does not appear in the FMA; all the relevant information pertaining to either lung is associated with the semantic concept `Lung`. `Lung` has two subclasses - `Left lung` and `Right lung`. In contrast, the term 'Lung' does not exist in *Terminologia*. Instead, the term 'Lungs' represents the generic form of Lung, with Left lung and Right lung as its hyponyms. Since the anatomical concepts `Lung` and `Lungs` are semantically distinct it would be incorrect to associate the *Terminologia* ID and Latin terms for 'Lungs' with `Lung` in the FMA. As the goal was to fully incorporate *Terminologia* into the FMA, a solution was required.

The FMA taxonomy includes two classes for accommodating such problematic concepts: 1. `Anatomical set`, and 2. `Anatomical cluster`. The concept `Lungs` satisfies the definition of `Anatomical set` - "Anatomical structure which consists

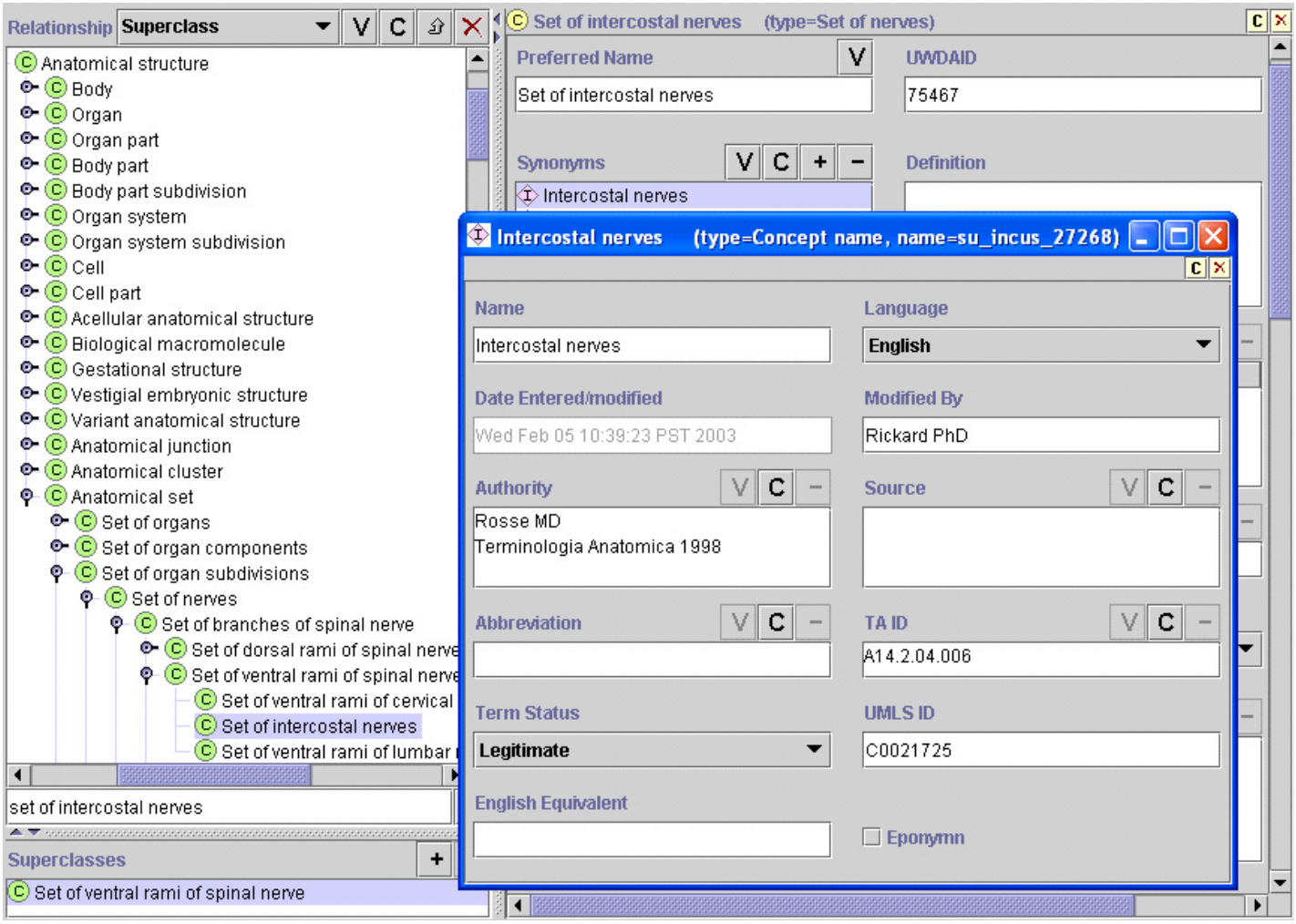

Figure 1 - A screen capture of the protégé-2000 knowledge acquisition tool interface demonstrating the way in which plural terms are modeled in the FMA as subclasses of anatomical set

of members that are of the same class and may or may not be continuous or interconnected with one other". The preferred name 'Set of lungs', designating this concept, was entered as a subclass of `Anatomical set`.

This solution satisfies the requirement for singular terms in the FMA. The plural term was then added as a synonym of this 'Set of' class, with the appropriate *Terminologia* ID and its Latin translation in the appropriate slots. Figure 1 illustrates how Intercostal nerves, another plural *Terminologia* term, is modeled. Although the Anatomy Taxonomy includes each of the 11 intercostal nerves on both right and left sides in the class `Intercostal nerve`, the concept designated by the plural term is distinct from those denoted by any one of the singular terms, which justifies the introduction of the new concept.

However, since *Terminologia* contains more than 1,500 plural terms, it soon became apparent that adding all of these as a direct subclass of `Anatomical set` would be unsatisfactory. Hence, a number of semantic subclasses were created under `Anatomical set` to represent meaningful aggregates of these plural terms (see Figure 1). In some cases the names of these new classes already existed in *Terminologia,* such as 'Muscles' and 'Bones' (which became `Set of muscles` and `Set of bones`, respectively). In other cases new classes had to be generated (for example, `Set of nerves` and `Set of glands`). Other situations in *Terminologia*, however, require the creation of two new concepts, instead of introducing one concept that represents a set of entities. Such cases include the designation of sexually dimorphic entities.

Sexually Dimorphic Terms

In *Terminologia,* terms that share an identification code are treated as synonyms. For example, the terms 'Marginal artery' and 'Juxtacolic artery' share the same identification code (A12.2.12.068) and are therefore assumed to be synonymous. This assumption, however, does not hold true for sexually dimorphic (male and female) anatomical parts such as the 'Ovarian artery' and the 'Testicular artery'. In *Terminologia* these two anatomical structures share the same *Terminologia* code (A12.2.12.086) and hence must be thought of as synonyms. This is clearly not the case. Despite their similar origin from the aorta and homologous embryonic derivation, the two arteries are anatomically distinct and have different connections and other spatial relationships. Consequently they have to be modeled as two

distinct concepts in the FMA. This requires, however, that the same value (A12.2.12.086) be entered in the 'TA ID' slot of both the terms `Ovarian artery` and `Testicular artery`. This solution remains faithful to *Terminologia*'s representation without violating the consistency of the FMA.

Foreign Language Terms

Terminologia includes Latin terms for all anatomical entities it names because Latin remains the language of anatomical discourse in non-English speaking countries. The inclusion of Latin in the FMA is, therefore, an important requirement. The modeling required to accomplish this goal is extendable to other foreign languages.

Regardless of language, all terms in the FMA are created as an instance of the class `Concept name`. We implemented several modifications in the frame of `Concept name` in order to capture fully the information relating to foreign language terms. First, a new slot, called 'Non-English equivalent' was created to distinguish English synonyms from foreign language synonyms, since not all users that access the FMA may wish to see non-English terms when synonyms are retrieved. Secondly, a new slot was added to all instances of `Concept name` to identify the language to which they pertain. Thirdly, another slot called 'English equivalent' was added to each instance of `Concept name`. This slot contains an English equivalent for all non-English terms that have one. In all other respects Latin terms are treated in an identical manner to English terms.

In several cases Latin terms have been incorporated into the English language and have become the accepted English term for a given anatomical structure. `Falx cerebri` and `Gingiva`, for instance, are commonly used English terms for the structures they represent and indeed are both used as preferred names in the FMA. However, these terms are first and foremost Latin terms and are therefore represented as such in the FMA, having their Language slot set to 'Latin' rather than 'English'.

Eponyms

An eponym is a person's name associated with a concept other than the person him or herself. In anatomy, eponyms are intended to honor the person who is credited with first describing an anatomical entity. For example, the Eustachian tube is named for the 16th century anatomist Bartolommeo Eustachio (1524–1574) who first identified what is referred to in the FMA as the `Pharyngotympanic tube`. Anatomical eponyms remain in wide use particularly in clinical medicine.

In each case where an eponym was to be incorporated in the FMA, it was treated as an additional English synonym and added to the 'Synonyms' slot of the anatomical concept to which it corresponded. An additional Boolean slot was added to the class `Concept Name` to identify it as an eponym.

In *Terminologia,* there are a number of eponyms that share identification codes with other terms that are not actually synonyms but rather members of a collection represented by that term. For example, `Santorini's muscles` has three *Terminologia* 'synonyms' – 'Procerus', 'Risorius' and 'Muscle of terminal notch of auricle'. None of these three terms are, however, synonymous with 'Santorini's muscles', or with each other for that matter, but rather are members of the *set* of Santorini's muscles. Hence, we entered 'Santorini's muscles' as a child of `Set of muscles` which contains in its 'Member' slot `Procerus`, `Risorius` and `Muscle of terminal notch of auricle`.

An interesting case is provided by two terms in *Terminologia's* eponyms index, which have identification codes but do not exist anywhere in the main term list. These are 'Meckel's diverticulum' (A05.6.03.003) and 'Sibson's fascia' (A06.6.02.018). We matched both of these terms manually to the concepts to which they correspond in the FMA; `Ileal diverticulum` and `Suprapleural membrane`, respectively. The example illustrates the need for domain knowledge in many instances.

Spelling Variations

In *Terminologia* the UK English form is used exclusively, whereas the FMA uses the American English spelling variant for the preferred name of all terms it contains. In all cases where a term had UK and American variants, the UK English terms were added as synonyms of the American English term. Hence, in the FMA the concept with preferred name 'Intercrural fibers' has a synonym 'Intercrural fibres'.

Abbreviations

The FMA contained no abbreviations prior to the incorporation of *Terminologia* and *NeuroNames*. *NeuroNames* contains abbreviations for a number of terms and we decided to incorporate them into the anatomy ontology. Associating abbreviations with their full name is particularly important when the abbreviation is non-intuitive or ambiguous. For example: 'AL' for 'Nucleus of ansa lenticularis', and 'CSp5' for 'Caudal part of spinal trigeminal nucleus'. To accommodate these abbreviations a new slot called 'Abbreviation' was added to each instance of `Concept name` and the abbreviated term entered as its value.

Inappropriate terms

Most vocabularies include terms that are antiquated, nonspecific or simply wrong. Many eponyms fit into this category. For example, there are several undesirable terms in *NeuroNames;* 'Lyra Davidis' and 'Psalterium' are obscure Latin synonyms of `Commissure of fornix`, yet they still appear in the literature. In the spirit of our intent to enter all known terms for an anatomical concept, we included such terms, but wanted to identify them as non-legitimate. Therefore, we attribute each instance of the class `Concept name` as 'Legitimate', 'Outdated' or 'Inappropriate'.

Discussion

We have incorporated a total of over 20,000 distinct anatomical terms from two distinct legacy terminologies into the frame-based ontology of FMA. The challenges we faced were distinct from those of another large merge between SNOMED RT and Clinical Terms Version 3 [6], in that one of our sources existed only in hard copy and in the other terms were aligned exclusively on the basis of part-whole relationships. Our intent was not to omit any term present in *Terminologia* and *NeuroNames*, even if we judged a term to be inappropriate. We solved the term integration problem primarily through enhancing the frame of the FMA class `Concept name` with slots that can capture diverse attributes of source terminology terms, which were initially thought to be inconsistent with the semantic and implementation structure of the FMA. The expressivity of the Protégé-2000

system allowed us to solve difficult problems such as the modeling of inappropriate terms and inconsistencies between terms and their associated code in the source terminology. As a result, rather than advocating the standardization of anatomical terms, we are able to associate in the FMA any known term that refers to an anatomical concept. The documentation associated with each term informs users of the FMA about the derivation and other attributes of alternative terms, which we believe, will allow them to judge the validity of the terms the FMA designates as the preferred name of an anatomical concept.

We plan to use the methodology developed for correlating the anatomy content of other vocabularies with the FMA. These include the Gene Ontology (GO), MeSH, GALEN, and SNOMED, as well as components of *Terminologia Anatomica* that are currently being developed for histology and embryology.

Acknowledgments

This work was supported by NLM grant LM06822. We thank Ian Whitmore and FCAT for permitting us to use *Terminologia*, and the publisher, Thieme, for providing the index files.

References

[1] Federative Committee on Anatomical Terminology (FCAT). Terminologia Anatomica. Stuttgart: Thieme, 1998.

[2] Martin RF, Bowden D. Primate Brain Maps. Elsevier: Oxford, 2000.

[3] Rosse C, Mejino JLV. A reference ontology for biomedical informatics: the Foundational Model of Anatomy. J Biomed Inform 2004;36:478-500.

[4] Rosse C. Terminologia Anatomica; Considered from the Perspective of Next-Generation Knowledge Sources. Clin Anat 2001;14:120-133.

[5] Zhang S, Bodenreider O. Aligning representation of anatomy using lexical and structural methods. Proc AMIA Symp 2003; 753-757.

[6] Wang AY, Barrett JW, Bentley T, Markwell D, Price C, Spackman KA, Stearns MQ. Mapping between SNOMED RT and Clinical Terms Version 3: a key component of the SNOMED CT development process. Proc AMIA Symp 2001;741-746.

Address for correspondence

Cornelius Rosse, M.D., D.Sc., Professor, Emeritus,
Department of Biological Structure, School of Medicine,
University of Washington, Seattle, WA 98195, Box 357420.
E-mail: rosse@u.washington.edu