Desiderata for Representing Anatomical Knowledge

Robert H Baud, Christian Lovis, Paul Fabry, Antoine Geissbuhler

Medical Informatics Division. University Hospitals of Geneva, Switzerland

Abstract:
The general problem of knowledge representation for gross anatomy supporting both computers and human is rarely globally solved. Partial solutions are flourishing, but the actual and potential users are left with a lack of satisfaction and uncomfortable feeling of incompleteness. Moreover, these solutions are not ready for a sound evolution and are at risk to disappear at any moment by default of adequate maintenance. In addition, the problem is complicated by the fact that any solutions should be relevant for Natural Language Processing applications in a multilingual environment.

This paper tackles with this problem and defines the basic steps for a proper knowledge representation scheme. Taking the subdomain of gross anatomy, it shows how each step has been solved and what performances and benefits are expected by such a solution. A discussion is done on the way to interface from a common source for both computers and humans.

Keywords: Knowledge representation, Natural Language Processing, ontology, anatomy

1. Desiderata

Considering a medical subdomain like gross human anatomy, one is faced to the global problem of knowledge representation of all relevant objects. The objective is a proper usage of this knowledge by computer programs as well as adequate understanding by human beings and their pedagogical support. The main desiderata consist in a formal model of this domain of anatomy, organizing all the objects and encompassing their relationships. The model should be accessible by computer programs as well as supporting intelligent browsers of anatomy atlases [1].

Two aspects are superposed to this first drawing of the situation: first, the model shall support reasoning on objects of the domain with global coherence and satisfying the needs of sound formal logics. Second, the model shall be extended for NLP processing in multiple languages. Any solution shall maintain the uniqueness of the source (therefore its coherence) when facing so many different needs.

All these prerequisites are rarely found altogether in any implementation, but hopefully the situation of gross anatomy is privileged and tends to this satisfactory solution. This paper aims at introducing and discussing all aspects of this situation and will enhance the expected advantages and benefits. Moreover, it traces the way of an exemplary solution, having the potential of being reproduced in another domain.
2. Methods: the basic modelling steps

In a companion paper, the first author has developed with some details what are the minimal procedural steps for acquiring and developing a model covering a specific domain [2] and only a short presentation is given here. The basic steps for a successful development of a model of a domain are the following:

- **Enumeration of the valid terms**: this initial task is basically the making of an inventory. At this stage, any object may be represented by multiple words or terms: this is part of the next step of detecting the synonyms.
- **Unique identification**: the task is to provide one identifier for any single object, and only one. The difficulty is that the very same object may be recognized by more than one term. In other words, there are less identifiers than terms.
- **Taxonomic organization**: A taxonomic hierarchy is by definition based on *isa* links. It is the backbone of any consistent model, because this type of links preserves the inheritance of properties and attributes. If the father object has a specific attribute value, the child object, which is of the same kind as the father, necessarily has this attribute value.
- **Other hierarchies (like meronomy)**: taxonomy is fundamental, but it is not necessarily natural or convivial. For this reason, other hierarchies like a meronomy based on the part_of attribute of any object are important to define and to make explicit in a model.
- **Horizontal links**: if a hierarchy is usually associated to the vertical direction, horizontal links are the expression of relationships between objects in different branches. Such relationships are as essential in a model as the hierarchical links.
- **Definitions**: when modelling objects of a domain and when each object has been given a unique identifier, the need to be very precise about this object is present. When communicating about an object, one wants to be sure to be understood! This raises the problem of definition of the objects.
- **Evolution and maintenance**: This point is related to the methodology used for modelling and its repercussions on evolution and maintenance of the model, where numerous pitfalls are awaiting the newcomers.

3. The modelling of gross anatomy

In this section, the most relevant above steps will be reviewed when applied to the domain of gross anatomy. It will be shown that the convergence of actions from several groups working independently brings an excellent solution to the scientific community. Of course, the process of final integration remains to be done, but this paper wants to emphasize a possible realistic solution not so far from being concrete.

Each step as given in the preceding section should be developed in turn. In all cases, the following aspects of the model shall be considered: adequacy for computer programmes, including the multilingual constraints, adequacy for human beings, logical soundness and coherence and uniqueness of the source. In order to concentrate on the major aspects, the points on meronomy, horizontal links and evolution will be skipped. They will certainly be developed in another paper.

---

1 Not everybody is really sure of what is meant by the TA term A02.6.02.010: hidden part of duodenum!
Naming the objects of the domain

For centuries now, human beings have compiled lists of terms relevant for the description of the human body. Facing a plethoric vocabulary emerging from uncontrolled sources, the Federative Committee on Anatomical Terminology edited in 1998 the first version of the TA [3], which has the statutes of a universal agreement. For a predefined degree of detail, this compilation is representative with more than 7400 different entries of the whole body.

In addition, there are thousands of alternative terms naming the same objects, either from historical sources, national variations and eponyms. Whatever the recommendation for their usage, all these alternate terms do exist and do occur elsewhere in the patient medical record if not in the scientific literature. Therefore, the proper naming of this domain requires an extension of the TA naming by existing sources. This task is partially language dependent; hopefully large similarities between languages exist. The authors are preparing now a database for French, Latin and English starting from the original TA and compiling several lexicons [4] and atlases. Currently more than two terms per object and per language are present as a mean value for the TA.

This effort is only partially adequate for computer applications, because of the versatility of languages: there are tens of different ways to speak about the same objects, as it has been demonstrated in a recent article [5]. The current list of anatomical terms is representative but it is not exhaustive. This means that some intelligent processing should take place elsewhere for a close to perfect recognition of body parts entities spread into free texts. On the contrary, this effort is satisfactory for human beings: the reader usually retrieves more names for a specified object than he knows and he is able to solve more or less trivially the language variations.

Logical soundness is a major difficulty when naming anatomical objects, because there is always a need or a temptation to document this object at the same time: what is the formally sound name for deltoid tubercle of spine of posterior surface of scapula or anything shorter? There is certainly a need to make a term self explanatory when seen outside of its context, but this need should not spill out the naming task (a quite common error). The solution to this problem, as implemented by the author [6], is to perform automatically the expansion of a term and to store only the minimal distinctive part (deltoid tubercle only). A few specific database attributes are driving the expansion on an individual term and language basis. Basic terms as well as their expansion are later available for display or for other programs. The minimal distinctive part sometimes generates polysemies, but the expansion is here for solving this problem.

The existence in a single database of all variants in multiple languages and the fact that this database may be updated with new terms at any time is part of a guaranty that no other source is necessary. This reference database may de facto act as a unique source, saving the underlying principle.

Identification or code

The main quality of an identifier is its uniqueness: one object of the domain gets one and only one identifier. Setting a unique identifier is often complicated by the fact that one is embarrassed by taxonomy considerations: it is certainly a quite common error to mix identification and hierarchy representation. Nothing is better than a sequential number.

---

2 Eponyms concern the usage of Proper names when naming body parts (like Meckel’s diverticule).
3 This is the case for ICD10, where the code represents the chapter and the category, with a huge embarrass when more than 10 subdivisions happen, because they are not accepted!
Anyway, there are several situations where the position in the hierarchy is reflected in the identifier or what is often called a code.

The former naming task provides in general more than one term for each existing object. In the domain of anatomy, it is not rare to have five or more terms in one language for a single object. The goal of the identification task is to assign an identifier to each set of terms pointing to the same object. As a consequence each synonym in the set is linked to the same identifier.

The difficulty is due to the fact that it may not be known precisely what is the meaning of a given term. In addition, it will be seen below that the problem of definitions is not yet fully resolved. In these conditions there is always a risk of attributing a term to the wrong object.

The easiest solution is the attribution of a code systematically by program, often using sequential numbers not reusable after deletion. The solution is totally adequate for computer usage, but not convivial nor error prone for human beings, who are not good when handling long numbers. The anonymous aspect of such an identifier makes it logically sound. The computer generation of identifier guaranty the uniqueness of the identifiers.

In the case of anatomy, the design of the taxonomy is realized under the form of the FMA model as shown in the next subsection. This system implemented as a computer application generates sequential numbers having the necessary qualities. Moreover, such an internal identifier is substituted to the TA code, this later becoming simply an attribute of each object. This is the way to escape numerous difficulties arising from the badly shaped TA code [7].

**Taxonomy**

Building the taxonomy of objects in a domain is the principle part of any knowledge representation schema. For a domain as complex as the gross anatomy, this task should be counted in months if not in years of work. Hopefully, the work has already been done by Cornelius Rosse and his team at University of Washington, giving birth to the Foundational Model of Anatomy, and incorporating the objects of the TA [8]. This project is a long-term development and has not yet reach a full coverage of all attributes of the objects of the domain, but its 80’000 existing objects including totally the TA is a solid ground. Copy of the model is available from the developer for free if the goal is research and development.

The taxonomy has been developed on a sound logical basis, without compromise, by a team competent in anatomy and with solid background on knowledge representation with computers, knowing the last refinements of the science of ontology. The isa taxonomy is pure from top to bottom including all objects. A Protégé implementation is an easy way to play with this model and to understand its intricacies and goodies. In addition, the model allows seeing the objects in several side hierarchies like a part-of hierarchy, but nowhere the different hierarchies are intermixed. The original TA hierarchy is ignored in this model; only the TA objects are considered and matched to the FMA.

A computer program based on this model receives the benefit of the sound isa hierarchy: inheritance of properties and possibility to generalize through the hierarchy. This last feature is typically necessary to match a query against a factual data when they are not given at the exact same level of detail.

However, such taxonomy is not really convivial for human beings: it lacks the natural feeling of the more traditional organization by system for example. The answer to this criticism is coming from the possibility to browse several hierarchies in parallel. Typically, a meronomy (part-of links) is more natural when teaching anatomy than the given
taxonomy: a renal papilla is part of the inner zone, which is part of the renal medulla, which is part of the kidney, which is part of the urinary system. Being built on formal basis, the model is perfectly coherent and satisfactory to this point of view.

The uniqueness of this model is guaranteed for at least three reasons: the model is good; to develop another model is too expensive and necessitates a large know-how; this model is adequately available to any interested scientific user.

Definitions

In a companion paper [6], the author has made the distinction between 5 approaches to the problem of definition on the basis of a research on actual publications in this matter. The 5 solutions are the following: encyclopedic definitions, formal taxonomic definitions, other hierarchical solutions (meronomy), multimedia definition and semantic stemming. It soon appears that any solution is either favorable to human beings or to computers, but generally not to both of them.

The encyclopedic definitions are the kind of definitions we find in a regular dictionary. They are made of natural language and therefore, despite they are very useful for human beings, computers rarely use them. A number of good dictionaries are made available in electronic format.

Formal definitions are built on the basis of a recognized taxonomy, by application of the principle of genus et differentia. Other hierarchical definitions may be constructed following links like part_of or branch_of. In theory, when appending successive extracted sentences through several consecutive levels in the hierarchy, a concrete definition is prepared. With the help of a few syntactic clues, it is possible to arrange such definition to something acceptable for a human reader. But the final result is somewhat artificial for human beings. Work in this direction is to be mentioned [9].

Multimedia definitions may be important and significant in several domains, in particular anatomy: an image with a pointer may act as an extremely precise definition.

Semantic stemming allows extracting the meaning of compound words on the basis of known rules validated for a specific language. A recent development on automatic extraction of definitions [10] based on morphosemantic decomposition technique is valid for neoclassical compound words quite frequent in the medical domain. An example is achloropsy, which automatically extracted meaning is pathology characterized by the lack of vision of green.

Computer programs are satisfied either by the formal definition, the hierarchical solutions or partly by the semantic stemming solution. NLP applications may take advantage of these definitions, but to the expense of delicate supplementary processing: it is indeed not easy to take advantage of a formal definition to disambiguate a term.

Human beings will be happy with the encyclopedic solution, the hierarchical solution and the multimedia definition. The best intersection between the two modes is the hierarchical solution. This means that a particular effort should be agreed with good priority on this topic.

A typical and prevalent example of hierarchical definition is the meronomy or part_of hierarchy. It is pleasant to human because it is natural and has been adopted for decades by the teachers of anatomy: it is direct to sketch an anatomical site with all its parts and subparts in a unique drawing. But the same technique of presentation is also convenient for programs if it is issued automatically from a formal model. This is actually the case with

---

4 The carpal bones are usually presented in a vertical view of the hand, but the carpal tunnel is invisible there and a transverse view is necessary, in order to act as a definition.
the FMA. It comes with the logical soundness of the model. It is particularly effective in the subdomain of anatomy. The uniqueness of source is preserved with this solution.

4. Discussion and conclusion

Is the coverage of the FMA model sufficient for NLP applications? This implies that besides the structure of the model, each object should have attached several attributes relevant for linguistic processing. The authors of the FMA know about this aspect. Second the author of this paper is preparing new terms with more extensive coverage in English and in French to be possibly made available with some future version of the FMA. Extension to other languages is expected to take place in the future.

We have shown on all the developed points above that the FMA is a satisfactory if not an excellent solution. This is also true for the other points, which have been skipped. This is enough to conclude that anatomy is a favoured terrain, because such a complete solution is generally not available in other subdomain of medicine. This is due to the fact that anatomy is one of the oldest basic medical sciences and that it has been extensively documented for centuries now. But this also means that current formal representations and state of the art for knowledge representation using extensively computer programs have reach a productive status. Last but not least, it indicates that knowledge representation techniques have to be married with NLP techniques in order to reach such a degree of satisfaction. In the future, it will be difficult to trust a totally ontological perspective as well as a purely terminological approach: only mixed solutions have a chance to convince the potential users.

5. Acknowledgements

This work has been partly funded by the Swiss National Science Foundation (SNF 632-066041).

References