

Aligning Anatomical Ontologies: The Role of Complex Structural Rules

Songmao Zhang¹, Ph.D., Olivier Bodenreider², M.D., Ph.D.

¹Institute of Mathematics, Academy of Mathematics and System Sciences
Chinese Academy of Sciences, Beijing, P. R. China

²U.S. National Library of Medicine, NIH, Bethesda, Maryland, USA

¹smzhang@math.ac.cn, ²olivier@nlm.nih.gov

An ontology is a formal representation of a domain supporting a variety of tasks. A given domain is often represented by multiple ontologies, providing overlapping, yet different coverage and possibly differing in their representation of the domain knowledge. There is a need for creating mappings among such ontologies in order to facilitate knowledge sharing and reuse. Anatomy is central to the biomedical domain and several anatomical ontologies have been created over the past fifteen years. This paper presents some of the techniques we developed for aligning two large anatomical ontologies: the Foundational Model of Anatomy (FMA) and the GALEN Common Reference Model (GALEN). Our approach first consists in aligning concepts across systems based on the lexical resemblance of their names and the structural similarity of their relations to other concepts. In addition, we created complex structural alignment rules for identifying mappings between groups of concepts and concepts that provably cannot have matches in the other system. Overall, about 44% of the FMA concepts and 69% GALEN concepts are characterized in the alignment, up from 4% and 13%, respectively, in our previous work. The advantages and limitations of the complex alignment rules are discussed.

Keywords: Ontology, ontology alignment, knowledge representation, anatomy, Semantic Web

1. Introduction

Ontology is a formal representation of a domain supporting a variety of tasks, including data integration, reasoning and the semantic annotation of resources in the Semantic Web. A given domain is often represented by multiple ontologies, providing overlapping, yet different coverage and possibly differing in their representation of the domain knowledge. There is a need for creating mappings among such ontologies in order to facilitate knowledge sharing and reuse. Ontology alignment aims at identifying correspondence among entities (i.e., concepts and relationships) across ontologies with overlapping content. (For a survey of alignment techniques, the interested reader is referred to (Noy, 2004)).

Anatomy is central to the biomedical domain and several anatomical ontologies have been created over the past fifteen years. We developed techniques for aligning two large anatomical ontologies: the Foundational Model of Anatomy (FMA) and the GALEN Common Reference Model (GALEN). This study extends previous work in which only one-to-one concept mappings were identified between FMA and GALEN, based on lexical resemblance between concept names and corroborated by shared hierarchical relations among concepts (Zhang and Bodenreider, 2003). In this study, we created complex structural alignment rules with the objective of identifying mappings between groups of concepts. Additionally, we identified concepts for which it can be demonstrated that no mapping to the other system can be found.

2. Materials

The Foundational Model of Anatomy¹ (FMA) [December 2, 2004 version] is an evolving ontology that has been under development at the University of Washington since 1994 (Rosse and Mejino, 2003). Its objective is to conceptualize the physical objects and spaces that constitute the human body. The underlying data model for FMA is a frame-based structure implemented with Protégé². 71,202 concepts cover the entire range of macroscopic, microscopic and subcellular canonical anatomy. In addition to preferred terms (one per concept), 52,713 synonyms are provided (up to 6 per concept). For example, there is a concept named *Uterine tube*, which has two synonyms: *Oviduct* and *Fallopian tube*.

The Generalized Architecture for Languages, Encyclopedias and Nomenclatures in medicine³ (GALEN) [v. 6] has been developed as a European Union AIM project led by the University of Manchester since 1991 (Rector, et al., 1997). The GALEN common reference model is a clinical terminology based on description logics. GALEN contains 25,322 concepts and intends to represent the biomedical domain, of which canonical anatomy is only one part. Only one name is provided for each non-anonymous concept (e.g., *Lobe of thyroid gland*). There are 3,170 anonymous concepts (e.g., *SolidStructure which < isPairedOrUnpaired leftRightPaired >*).

Both FMA and GALEN are modeled by *is-a* relationship. Additionally, FMA uses 7 kinds of partitive relationships (e.g., *part of* and *constitutional part of*) and GALEN 41 (e.g., *isStructuralComponentOf* and *IsDivisionOf*). Both systems have associative relationships.

3. Structural Alignment through Complex Structural Rules

Using the lexical alignment method followed by structural verification described in (Zhang and Bodenreider, 2003), 3,199 pairs of equivalent concepts were identified between FMA and GALEN, accounting for about 4% of FMA concepts and 13% of GALEN concept. The complex structural rules presented below allowed us to identify additional mappings and to identify concepts for which it can be demonstrated that no mapping to the other system can be found. Overall, about 44% of the FMA concepts and 69% of GALEN concepts were characterized in the alignment. In what follows, the term *anchor* refers to the 3,199 one-to-one matches obtained previously. Those are represented by double-lined boxes in figures. In contrast, the other concepts in the two systems are *non-anchors* (represented by single-lined boxes in figures). Finally, $des(X)$ denotes the set of all anchors in the descendants of concept X .

One-to-one matches. Two non-anchors X and Y across systems are likely to be a match if they reach the same nonempty anchor set in their descendants, i.e., $des(X)=des(Y)$. For example, as shown in Figure 1, non-anchor *Cuneiform* in GALEN (with three descendants) and non-anchor *Cuneiform bone* in FMA (with nine descendants) were identified as a match, because they both share the three anchors found in their descendants. 124 such one-to-one matches were found across systems.

One-to-group matches. For any two non-anchors X_1 and X_2 in one system and non-anchor Y in another system, if $des(X_1)$ and $des(X_2)$ are not subsets of each other, and $des(X_1) \cup des(X_2)=des(Y)$ holds, then it is possible that a single concept Y matches a group of concepts $\{X_1, X_2\}$. For example, as shown in Figure 2, there are four anchors in the descendants of *ExtremityLongPart* in

1 <http://fma.biostr.washington.edu/>

2 <http://protege.stanford.edu/>

3 <http://www.opengalen.org/>

GALEN: *Arm, Forearm, Leg and Thigh*. In FMA, *Proximal free limb segment* has two anchors in its descendants: *Arm* and *Thigh*, while *Middle free limb segment* has two anchors in its descendants: *Forearm* and *Leg*. Therefore, a one-to-group match was identified between *ExtremityLongPart* in GALEN and {*Proximal free limb segment, Middle free limb segment*} in FMA. 22 such one-to-group matches were found, involving 36 non-anchors in the FMA and 30 in GALEN.

Interestingly, in addition to the mappings between non-anchors presented above, one-to-group mappings can also occur between one non-anchor in one system and a group of anchors in the other system. This is often due to the use of different modeling principles in the two systems. For example, as illustrated in Figure 3, *Lobe of lung* in the FMA is first modeled by upper/middle/lower position (i.e., *Upper lobe of lung, Middle lobe of lung* and *Lower lobe of lung*) and then by laterality (e.g., for *Upper lobe of lung: Upper lobe of left lung* and *Upper lobe of right lung*). By contrast, in GALEN, *Lobe Of Lung* is first modeled by laterality and then by upper/middle/lower position. Our previous alignment identified six anchors under lobe of lung. In addition, we identified four one-to-group matches across systems. 49 such mappings between a non-anchor and a group of anchors were found, where 25 are one GALEN non-anchor matching FMA anchors, and 24 one FMA non-anchor matching GALEN anchors.

Group-to-group matches. For any pair {*X, Y*} of anchors across systems, if *X* and *Y* share exactly the same set of anchors (possibly empty) in their children, and *X* and *Y* have the same number of non-anchors in their children: {*X₁, ..., X_n*} and {*Y₁, ..., Y_n*}, respectively, then there is a possible mapping between the two groups of non-anchors, i.e., between {*X₁, ..., X_n*} and {*Y₁, ..., Y_n*}. For example, the anchor *Anterior intercostal artery* in FMA has eleven children and all of them are non-anchors: *First anterior intercostal artery* to *Eleventh anterior intercostal artery*. In contrast, the eleven non-anchor children of *AnteriorIntercostalArtery* in GALEN are anonymous: (*AnteriorIntercostalArtery which <isSpecificallyNonPartitivelyContainedIn First IntercostalSpace>*) to (*AnteriorIntercostalArtery which <isSpecifically NonPartitivelyContainedIn EleventhIntercostalSpace>*). These two groups of eleven non-anchors were mapped across systems. 49 such group-to-group matches were identified between the FMA and GALEN, involving 127 non-anchors in each system.

Concepts that provably cannot have matches in the other system. The total number of concepts in the FMA is about three times of that in GALEN. Intuitively, there should be a large number of FMA concepts either mapping to GALEN concepts group-to-one, or simply having no matches in GALEN. For example, the anchor *Submucosa* is a leaf node in GALEN, while it has 128 descendants in the FMA. All of its descendants are non-anchors and represent specialized concepts specific to the FMA, e.g., the submucosa of various organs. These 128 non-anchors were identified as having no matches in GALEN. Overall, 1,482 such cases were found, involving 11,189 FMA non-anchors and accounting for about 16% of all the FMA concepts.

On the other hand, some high-level concepts in GALEN represent non-canonical anatomical categories (e.g., *Non Normal Phenomenon*), clinical-related categories (e.g., *Process, Graft*), or non-anatomical categories (e.g., *Food, Risk Factor*). The concepts subsumed by these categories in GALEN are not expected to have matches in the FMA which is solely concerned with canonical (i.e., “normal”) anatomical entities. 13,626 such non-anchor concepts in GALEN were identified (2,051 of them are anonymous), accounting for 53.8% of all GALEN concepts. Examples include *Supernumerary Thumb* as a descendant of *Non Normal Phenomenon*, and the anonymous concept (*Alcohol which <playsPhysiologicalRole FoodRole>*) under *Food*.

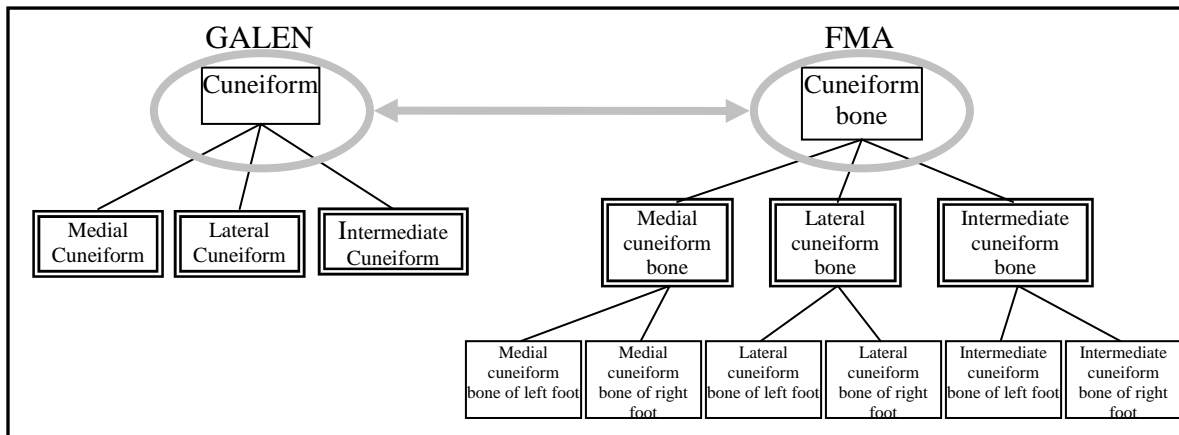


Figure 1 One-to-one match between Cuneiform in GALEN and Cuneiform bone in FMA

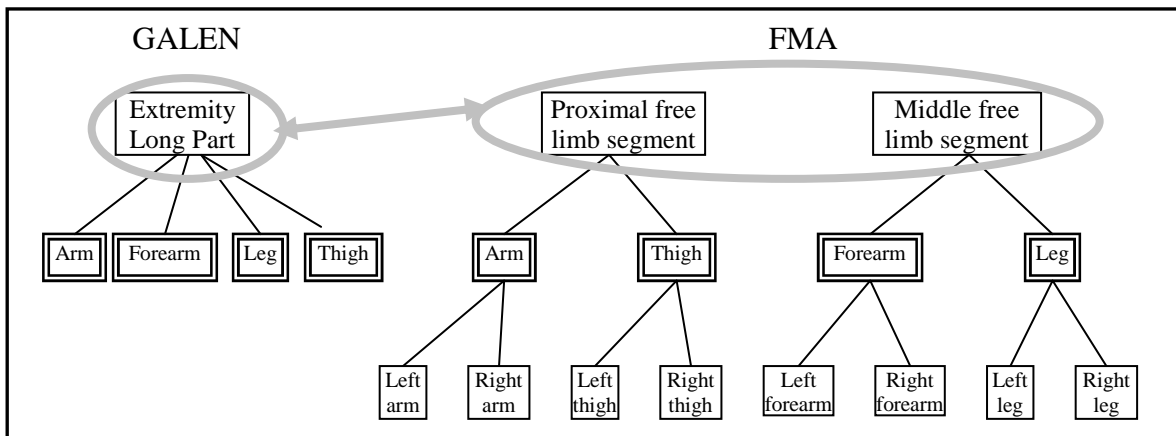


Figure 2 One-to-group match between ExtremityLongPart in GALEN and {Proximal free limb segment, Middle free limb segment} in FMA

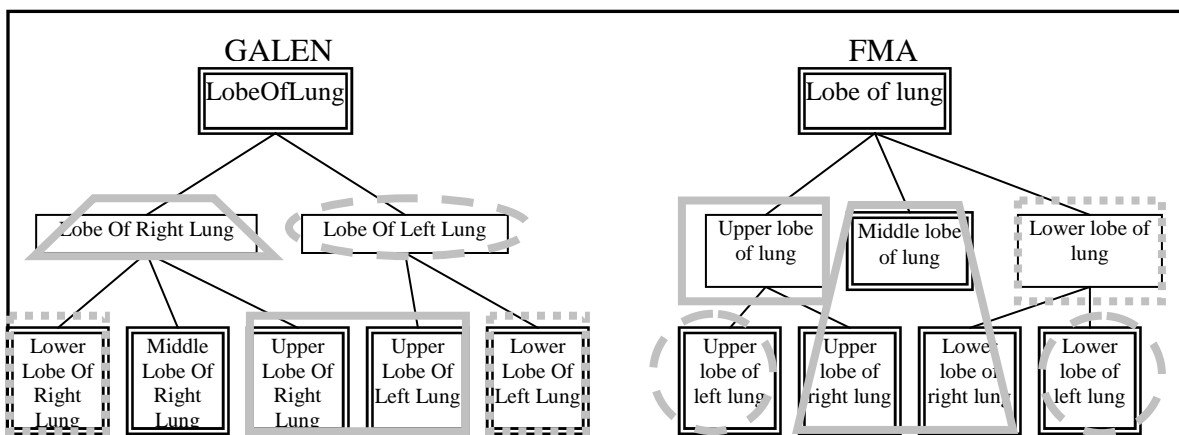


Figure 3 Four one-to-group matches under lobe of lung (those in the same type of outlines across systems are a match)

4. Discussion and Conclusions

Structural alignment techniques proved effective in identifying additional mappings, as well as in identifying concepts for which it can be demonstrated that no mapping can be found. Overall, compared to our previous work, the number of concepts characterized in the alignment increased from 4% to 44% for the FMA and from 13% to 69% for GALEN. Of note, while only named concepts could participate in the lexical alignment, the anonymous concepts in GALEN could also play a role in the structural alignment. The structural alignment rules exploited both commonalities and differences in the representation of knowledge across systems. The rules identifying concepts that provably cannot have matches in the other system were by far the most productive rules, for both the FMA and GALEN. In contrast to other aligning techniques, the use of domain knowledge has always played an important role in our approach (Zhang, et al., 2004).

A limited review of the matches shows that most of them are valid. However, the following is an example of invalid group-to-group match. The anchor *HeadOfRadius* in GALEN has two children: *DistalHeadOfRadius* and *ProximalHeadOfRadius*, while the two children of *Head of radius* in the FMA are: *Head of left radius* and *Head of right radius*. All four children concepts are non-anchors. The group-to-group match identified between {*DistalHeadOfRadius*, *ProximalHeadOfRadius*} in GALEN and {*Head of left radius*, *Head of right radius*} in the FMA is invalid, because the two groups of children result, once again, from differing modeling principles in the two systems (here, position from elbow vs. laterality). Domain expertise is required to validate every single match based on structural rules. Finally, 56% of all FMA concepts and 31% of all GALEN concepts are left uncharacterized after the alignment. Advanced aligning methods shall be explored for these concepts.

Acknowledgements

This research was supported in part by the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine (NLM) and by the Natural Science Foundation of China (No.60496324), the National Key Research and Development Program of China (Grant No. 2002CB312004), the Knowledge Innovation Program of the Chinese Academy of Sciences, and MADIS of the Chinese Academy of Sciences. Thanks for their support and encouragement to Cornelius Rosse and his team (FMA) and Alan Rector and his team (GALEN).

References

- Noy, N.F. (2004); Tools for mapping and merging ontologies; In Handbook on Ontologies, (S. Staab and R. Studer, eds). Springer-Verlag, (pp. 365-384).
- Zhang, S., and Bodenreider, O. (2003); Aligning representations of anatomy using lexical and structural methods; Proc AMIA Symp, (pp. 753-757)
- Rosse, C., and Mejino, J.L., Jr. (2003); A reference ontology for biomedical informatics: the Foundational Model of Anatomy; J Biomed Inform 36, (pp. 478-500)
- Rector, A.L., Bechhofer, S., Goble, C.A., Horrocks, I., Nowlan, W.A., and Solomon, W.D. (1997); The GRAIL concept modelling language for medical terminology; Artif Intell Med 9, (pp. 139-171)
- Zhang, S., Mork, P., and Bodenreider, O. (2004); Lessons learned from aligning two representations of anatomy; In Proceedings of the First International Workshop on Formal Biomedical Knowledge Representation (KR-MED 2004), (U. Hahn, S. Schulz and R. Cornet, eds). (pp. 102-108).