Aligning Representations of Anatomy using Lexical and Structural Methods

Songmao Zhang, Ph.D., Olivier Bodenreider, M.D., Ph.D.

U.S. National Library of Medicine, Bethesda, Maryland
National Institutes of Health, Department of Health & Human Services

{szhang|olivier}@nlm.nih.gov

Objective: *The objective of this experiment is to develop methods for aligning two representations of anatomy (the Foundational Model of Anatomy and GALEN) at the lexical and structural level.* ***Methods:*** *The alignment consists of the following four steps: 1) acquiring terms, 2) identifying anchors (i.e., shared concepts) lexically, 3) acquiring explicit and implicit semantic relations, and 4) identifying anchors structurally.* ***Results:*** *2,353 anchors were identified by lexical methods, of which 91% were supported by structural evidence. No evidence was found for 7.5% of the anchors and 1.5% received negative evidence.* ***Discussion:*** *The importance of taking advantage of implicit domain knowledge acquired through complementation, augmentation, and inference is discussed.*

INTRODUCTION

Although various medical knowledge representation systems claim to support reasoning, there is little work done on investigating their respective reasoning capabilities. In this regard, the objective of this study is to analyze the shared characteristics and differences of multiple representations of a given domain. With the emergence of the Semantic Web, agents relying on different ontologies needed to communicate and exchange reliable information and techniques for comparing ontologies could play a central role in this process [1, 2].

Several approaches have been explored to compare ontologies, resulting in several byproducts: *merging*, where a new ontology is created [3, 4], *transformation* from one to another [5], and *alignment*, where a mapping is built between two ontologies [6]. In this study, alignment is the technique of choice because we do not intent to create any new system or to alter existing ones. The methods we use for aligning ontologies act not only at the lexical level, but also, more importantly, at the structural level.

In order to address the large size of the biomedical domain, we limited this study to one subdomain central to biomedicine: anatomy. We selected two representations of anatomy available in electronic format that are comprehensive enough to support clinical applications: the Foundational Model of Anatomy[1] (FMA) [August 30, 2002 version] and the GALEN[2] common reference model [v. 5].

In parallel to our effort, but unrelated to it, another group also took up the challenge of aligning these two representations of anatomy. Common to both approaches is the use of lexical and structural techniques in the alignment process. However, while their approach relies on generic schema matching [7] and would therefore be immediately applicable to other domains, ours takes advantage of domain knowledge to maximize the chances of finding similarities when the representations differ, as well as differences when the representations appear similar.

MATERIALS

The Foundational Model of Anatomy (FMA) is an evolving ontology that has been under development at the University of Washington since 1994 [8, 9]. 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 Protégé-2000. With 59,422 concepts, FMA claims to cover the entire canonical anatomy. Concept names in FMA are pre-coordinated, and, in addition to preferred terms (one per concept), 28,686 synonyms are provided (up to 6 per concept). For example, there is a concept named *Uterine tube* and its synonym is *Oviduct.*

The Generalized Architecture for Languages, Encyclopedias and Nomenclatures in medicine (GALEN) has been developed as a European Union AIM project led by the University of Manchester since 1991 [10, 11]. The GALEN common reference model is a clinical terminology represented using GRAIL, a formal language based on description logics. GALEN contains 25,192 concepts and intends to represent the biomedical domain, of which canonical anatomy is only one part. Unlike FMA, GALEN is compositional and generative. Concept names in GALEN are post-coordinated, and only one name is provided for each concept[3]. There are 3,176 anonymous concepts de-

---

[1] http://sig.biostr.washington.edu/projects/fm/AboutFM.html
[2] http://www.opengalen.org/
[3] By design, no synonyms are provided in the Common Reference Model

fined by expressions such as *(SolidStucture which <isPairedOrUnpaired leftRightPaired>)* (representing a bilateral solid structure).

Both FMA and GALEN are modeled by *IS-A* and *PART-OF* relationships and allow multiple inheritance. Relationships in GALEN are finer-grained than in FMA. For the purpose of this study, we considered as only one *PART-OF* relationship the various kinds of partitive relationships present in FMA (e.g., *part of*, *general part of*) and in GALEN (e.g., *isStructuralComponentOf*, *isDivisionOf*).

METHODS

Aligning FMA and GALEN consists of the following four steps: 1) acquiring terms, 2) identifying anchors (i.e., shared concepts) lexically, 3) acquiring (explicit and implicit) semantic relations, and 4) identifying anchors structurally. The first two steps constitute the lexical component of our method, the last two, its structural component. These four steps are presented in detail below. From a technical perspective, we have developed a unifying Java-based interface to access both FMA and GALEN data. The Java API provided by Protégé-2000 is used to communicate with FMA. The GALEN server is accessed through a COM interface and we use JACOB[4], a Java-COM bridge, to interact with GALEN in Java.

Step 1: Acquiring terms

Acquiring terms consists of extracting the concept names. For FMA, both preferred names and synonyms are extracted. For GALEN, non-anonymous concept names are acquired. Terms in GALEN take the form of words concatenated with the first letters capitalized, such as *RoundLigamentOfUterus*. In order to compare GALEN terms to FMA terms by standard matching techniques, we first segment GALEN terms into words separated by space, such as *Round Ligament Of Uterus*.

Step2: Identifying anchors lexically

Lexical alignment compares the two systems at the term level, by exact match and after normalization. This process makes the source and target terms potentially compatible by eliminating such inessential differences as inflection, case, hyphen, and word-order variation. Both preferred terms and synonyms, when available, are used in the alignment process. Multiple mappings are identified where more than one GALEN term matches one FMA term or vice versa. Some of them may be later disambiguated by the structural alignment, while the others will require validation by a domain expert. Concepts exhibiting similarity at the lexical level across systems are called anchors, as they are going to be used as reference concepts in the structural alignment. For example, the concepts *Fibularis tertius* (synonym: *Peroneus tertius*) in FMA and *Peroneus Tertius* in GALEN are identified as anchor concepts.

Step 3: Acquiring semantic relations

Inter-concept relationships – either hierarchical or associative – are generally represented by semantic relations <*concept*$_1$, *relationship*, *concept*$_2$>, where *concept*$_1$ links to *concept*$_2$ through *relationship*. This experiment focuses on hierarchical relations (*IS-A* and *PART-OF* relationships). Acquiring semantic relations, here, consists of extracting the relations explicitly represented. Then, subsequent processing techniques are applied to generate the relations that are not represented explicitly (i.e., implicit knowledge). We define three such techniques: complementing, augmenting and inferring.

Complementing. As partial ordering relationships, hierarchical relationships are anti-symmetric. However, *IS-A* and *PART-OF* have inverse relationships, *INVERSE-IS-A* and *HAS-PART*. Except for *IS-A*, a minority of relations is represented bidirectionally. For example, <*Arm*, *HAS-PART*, *Humerus*> is explicitly represented in FMA but its inverse relation is missing. In canonical anatomy, the inverse relations are essentially always valid, although this is not necessarily the case in the real world [12]. In order to perform simple comparisons of semantic relations across systems, we complemented FMA and GALEN with the inverse relations that were not explicitly represented. For example, we generated the relation <*Humerus*, *PART-OF*, *Arm*>.

Augmenting goes one step beyond and attempts to represent with relations knowledge that is otherwise embedded in the concept names, for example through reification or other linguistic phenomena.

Knowledge representation systems sometimes **reify** the *PART-OF* relationship by creating concepts such as *Subdivision of X*. Instead of using a *PART-OF* relationship between the concept *P* (the part) and *W* (the whole), a system may create a concept *Part of X* and represent the *PART-OF* relationship using the *IS-A* hierarchy (<*P*, *IS-A*, *Part of W*> instead of <*P*, *PART-OF*, *W*>). These two representations are equivalent for most purposes and few *PART-OF* relations are represented by both methods (7% in FMA). We augmented the concepts representing reified *PART-OF* relationships encountered in FMA and GALEN with their *PART-OF* relation counterpart in order to facilitate comparisons across systems. For example, in FMA, the *PART-OF* relationship between *Cardiac chamber* and *Heart* is represented as <*Cardiac chamber*, *IS-A*, *Subdivision of heart*>, but there is no explicit (direct or indirect) *PART-OF* relationship

---

[4] http://danadler.com/jacob/

between *Cardiac chamber* and *Heart* while such a relationship is defined in GALEN[5]. In addition to *Subdivision of X*, we applied this augmentation technique to other reified part-whole relationships including *Organ component of X* (in FMA) and *Component of X* (in GALEN).

Relations are also captured in various **other linguistic phenomena** such as nominal modification and prepositional attachment. The former often represents a hyponymic relation involving the head of the noun phrase. For example, a *Sweat gland* is a kind of *Gland*. Therefore, a <*X Y*, *IS-A*, *Y*> relation can be tentatively extracted from the term *X Y*. In anatomical terms, prepositional attachment using "of" (*X of Y*) often denotes a partitive relation between *X* and *Y*. For example, we generated the relation <*Upper lobe of lung*, *PART-OF*, *Lung*> from the term *Upper lobe of lung*.

Inferring consists of generating new inter-concept relationships by applying inference rules. These inference rules combine *IS-A* and *PART-OF* relationships among concepts to generate additional *PART-OF* relations. More precisely, rules R1 and R2 assign to a more generic concept the relationships common to all its leaf descendants, *PART-OF* and *HAS-PART*, respectively[6]. For example, unlike GALEN, there is no *PART-OF* relationship between *Nasal bone* and *Musculoskeletal system* in FMA. However, since the two descendants of *Nasal bone* (*Left nasal bone* and *Right nasal bone*) are in *PART-OF* relationship with *Musculoskeletal system*, we add the relation <*Nasal bone*, *PART-OF*, *Musculoskeletal system*> to FMA. Rules R3 and R4 represent limited reasoning along the *PART-OF* hierarchy, generating a partitive relation between a specialized part and the whole (R3) or between a part and a more generic whole (R4). For example, we add the relation <*Left eyeball*, *PART-OF*, *Eye*> from <*Left eyeball*, *IS-A*, *Eyeball*> and <*Eyeball*, *PART-OF*, *Eye*>.

Step 4: Identifying anchors structurally

All relations – explicitly represented or generated during Step 3, including the relations to anonymous concepts in GALEN – are used for identifying structural similarity and conflicts among the anchors resulting from the lexical alignment. Structural similarity is used as positive evidence for the alignment; conflicts, on the other hand, suggest mismatches (negative evidence) and prevent the concepts from being aligned.

Structural similarity is defined by the presence of common relations among anchors across systems. In other words, for a given anchor, the relations to other anchors observed in FMA are expected to be present in GALEN (and vice versa). In practice, we search the two systems for common "beelines" between a given anchor and the neighboring anchors. Beeline refers to the shortest path between two concepts on which all edges (i.e., all relationships) are of the same type. There are four types of hierarchical beelines: *IS-A*, *INVERSE-IS-A*, *PART-OF*, and *HAS-PART*. Shared beelines of the same type denote structural similarity. For example, the concepts *Fibularis tertius* in FMA and *Peroneus Tertius* in GALEN have been identified as anchors by lexical techniques. In FMA, *Fibularis tertius* has nineteen beelines to other anchors and, in GALEN, *Peroneus Tertius* has twenty-two. Eleven of these beelines are shared (e.g., to *Muscle of leg*, *Leg*, and *Muscle cell*) and, in each beeline pair, the type of beeline is the same. Therefore, the lexical mapping between *Fibularis tertius* in FMA and *Peroneus Tertius* in GALEN is supported by structural similarity.

Conflicts are defined by semantic incompatibility between anchors, detected through the structural alignment. The first type of semantic incompatibility is represented by the presence of beelines of opposite type (e.g., *PART-OF* and *HAS-PART*) between two anchors across systems. The second type of semantic incompatibility is based on the disjointedness of top-level categories in knowledge representation systems. In practice, if anchor concepts have relationships to disjoint top-level categories in FMA and GALEN, they are declared semantically distinct and the corresponding anchor is removed. For example, *Foot* in FMA and *feet* in GALEN match after normalization, resulting in an anchor. However, the two concepts are considered semantically incompatible because FMA's *Foot* is linked to the top-level concept *Anatomical structure*, while GALEN's *feet* is linked to *Unit*, *Anatomical structure* and *Unit* being disjoint top-level categories.

Anchor classification. Each anchor *A*, i.e., each pair of anchor concepts in FMA and GALEN (A_1, A_2) is examined for structural similarity and conflicts. Six groups of anchors were identified whose definition is given in Table 1. Anchors from Groups 1 and 2 do not have any structural evidence. Anchors from these groups require further investigation (e.g., review by a domain expert) to assess their validity. The validity of the anchors from Group 3 is confirmed by structural evidence. Though with a lesser degree of confidence, anchors from Group 4 are also considered valid. Finally, anchors from Groups 5 and 6 receive negative evidence and are therefore removed. Domain

---

[5] between *Heart Chamber* and *Heart*.

[6] Inheriting upwards is not common, but is justified in this context where the properties assigned to the leaf descendants (in FMA) should in fact have been assigned to a more generic concept and inherited downwards.

knowledge is required to clarify the situation in these cases.

1	**Having no beelines.** At least one of A_1 or A_2 does not have any hierarchical beelines to the other anchors.
2	**Sharing no beelines.** Both A_1 and A_2 have beelines to some other anchors, but they do not share any beelines.
3	**Sharing beelines of the same type.** A_1 and A_2 share at least one beeline and all shared beelines are of the same type.
4	**Sharing "compatible" beelines.** A_1 and A_2 share at least one beeline and, although not all shared beelines are of the same type, all shared beelines are "compatible" (hierarchical relationships are in the same direction, making *IS-A* compatible with *PART-OF*, but not with *HAS-PART*).
5	**Sharing "conflicting" beelines.** A_1 and A_2 share at least one beeline and at least one shared beeline is "conflicting" (hierarchical relationships are in opposite directions, e.g., *PART-OF* and *HAS-PART*).
6	**Belonging to disjoint top-level categories.** A_1 and A_2 belong to disjoint top-level categories by negative structural alignment.

Table 1 –Definition of the six groups of anchors

RESULTS

Anchors identified by lexical alignment

2,353 matching anchor concepts were identified lexically, accounting for about 4% of FMA concepts and 9% of GALEN concepts.

Semantic relations acquired

The number of relations extracted and generated in Step 3 is listed in Table 2. Complementation, augmentation, and inference resulted in generating large numbers of relations, especially when compared to the limited number of relations explicitly represented. Not surprisingly, many relations come from inference rules R3 and R4, which perform similarly a transitive closure on the hierarchical relations.

Types of relations	FMA	GALEN
Explicitly represented	238,135	214,403
Complemented	104,754	107,689
Augmented (from reification)	199,598	22,126
Augmented (from other phenomena)	116,262	5,148
Inferred (rules R1 and R2)	162,700	25,544
Inferred (rules R3 and R4)	5,009,968	1,636,280
Total	5,831,417	2,011,190

Table 2 – Number of relations in FMA and GALEN

Anchor identified by structural alignment

Table 3 shows the results of structural alignment in terms of anchor classification. On the left part of the table, the alignment relied on explicit and complemented knowledge only, while on the right part, it also took advantage of augmented and inferred relations.

Groups 1 and 2. Anchors from Group 1 (no beelines) include *Root canal of tooth* because, although linked to *Tooth* (*PART-OF*) and *Body Part* (*IS-A*) in GALEN, *Root canal of tooth* has no connections to other anchors in FMA. As in this example, most of the time, the lack of links to other anchors was observed in FMA. In Group 2 (no shared beelines), although *Mucosa* is linked to fifteen anchors in FMA (e.g., *Epithelium*), and seven in GALEN (e.g., *Mucous Gland*), none of these relations are shared across systems.

Groups 3 and 4. Nearly 69% of all anchors (and 91% when using additional knowledge) receive positive evidence, most of them sharing beelines of the same type (e.g., *Fibularis tertius* in FMA and *Peroneus Tertius* in GALEN, presented earlier). An example from Group 4 (shared "compatible" beelines) is the anchor *Ascending Colon*. In both systems, this concept is linked to *Colon*, but, although going in the same direction, the relationship is *PART-OF* in GALEN and *IS-A* in FMA. For alignment purposes, sharing compatible beelines is deemed a sufficient condition.

Groups 5 and 6. About 1.5% of the anchors represent conflicts between the two representations. For example, from Group 5, the relationship between the anchors *Wall of heart* and *Apex of heart* is *PART-OF* in GALEN but *HAS-PART* in FMA. The semantic incompatibility between *Foot* (the anatomical structure) and *feet* (the unit) presented earlier illustrates conflicts from Group 6.

	Based on explicit and complemented knowledge		Adding augmented and inferred knowledge	
Group 1	482	30.43%	113	7.52%
Group 2	234		64	
Group 3	1,528	68.72%	1,675	90.95%
Group 4	89		465	
Group 5	10	0.85%	26	1.53%
Group 6	10		10	

Table 3 – Results of structural alignment

DISCUSSION

Explicit vs. implicit knowledge

Taking advantage of implicit knowledge largely increased the chances of finding structural evidence for the anchors, mostly positive evidence. As shown in Table 3, using augmented and inferred relations helped reduce the proportion of anchors in Groups 1 and 2 from 30.4% to 7.5% and essentially benefited Groups 3 and 4. In the examples presented above, *Root canal of tooth* and *Mucosa* received positive evidence through the use of implicit knowledge. Assuming positive structural evidence correlates with validity, using implicit knowledge would significantly facilitate the review of the alignment by domain experts.

Examining the respective contribution of the various techniques applied in Step 3 shows that, for the purpose of identifying structural evidence, the single most important source of additional knowledge is augmented relations. Augmentation accounted for 74% of the 523 anchors that acquired positive structural evidence. Although a large majority of the additional knowledge was acquired through inference techniques, these relations played paradoxically a lesser role in identifying structural evidence. This illustrates the importance of reification in knowledge representation and, here, the differential use of reification between FMA and GALEN.

Not surprisingly, using implicit knowledge also resulted in revealing sixteen new conflicting relations across systems.

Inconsistencies within systems

Although the objective of this experiment was not to assess the quality of either representation, a small number of errors within each system appeared – in addition to inconsistencies across systems – while acquiring semantic relations from both systems (Step 3). Examples of internal errors, found in both FMA and GALEN, include reflexive hierarchical relationships and circular hierarchical relationships. Some errors were detected using only explicit knowledge (e.g., <*Wrist Joint*, *PART-OF*, *Wrist Joint*>). Other errors were revealed during augmentation (e.g., <*Male Perineum*, *IS-A*, *Component Of Male Perineum*>, while investigating reified *PART-OF* relationships). Finally, inference also revealed hierarchical cycles. These inconsistencies were reported to and discussed with the authors of the two systems.

Future work

The next step in this project is the validation by domain experts of the anchors identified by a combination of lexical and structural techniques. We will also investigate methods for acquiring new anchors based solely on structural similarity. Finally, we will move to our broader objective, i.e., to investigate the reasoning capabilities of the two systems.

With the recent interest in merging FMA, Mouse Anatomy, and GALEN, we believe that the methods we developed for alignment can also be helpful in bridging these representations.

Acknowledgements

The research was supported in part by an appointment to the National Library of Medicine Research Participation Program administrated by the Oak Ridge Institute of Science and Education through an interagency agreement between the U.S. Department of Energy and the National Library of Medicine.

Thanks for their support and encouragement to Cornelius Rosse, José Mejino, and Kurt Richards for FMA and Alan Rector, Jeremy Rogers, and Angus Roberts for GALEN. Thanks also to Pieter Zanstra at Kermanog for providing us with an extended license for the GALEN server.

References

1. Bailin SC, Truszkowsk W. Ontology negotiation as a basis for opportunistic cooperation between intelligent information agents. In: Cooperative Information Agents V, Proceedings; 2001. p. 223-228
2. Uschold M, Gruninger M. Creating semantically integrated communities on the world wide web. Proc International Workshop on the Semantic Web 2002 http://semanticweb2002.aifb.uni-karlsruhe.de/USCHOLD-Hawaii-InvitedTalk2002.pdf.
3. McGuinness DL, Fikes R, Rice J, Wilder S. The Chimaera ontology environment. Proc of AAAI 2000:1123-1124
4. Noy NF, Musen MA. PROMPT: algorithm and tool for automated ontology merging and alignment. Proc of AAAI 2000:450-455
5. Chalupsky H. OntoMorph: a translation system for symbolic knowledge. Proc of KR 2000 http://www.isi.edu/~hans/publications/KR2000.pdf.
6. Reed SL, Lenat D. Mapping Ontologies into Cyc. Proc of AAAI 2002 http://citeseer.nj.nec.com/509238.html.
7. Mork P, Pottinger R, Bernstein PA. Challenges in precisely aligning models of human anatomy using generic schema matching. Personal Communication 2003
8. Noy NF, Musen MA, Mejino JL, Rosse C. Pushing the envelope: challenges in a frame-based representation of human anatomy: Technical Report of Stanford Medical Informatics; 2002. Report No.: SMI-2002-0925
9. Rosse C, Mejino JL, Modayur BR, Jakobovits R, Hinshaw KP, Brinkley JF. Motivation and organizational principles for anatomical knowledge representation: the digital anatomist symbolic knowledge base. J Am Med Inform Assoc 1998;5(1):17-40
10. Rector AL, Bechhofer S, Goble CA, Horrocks I, Nowlan WA, Solomon WD. The GRAIL concept modelling language for medical terminology. Artif Intell Med 1997;9(2):139-71
11. Rogers J, Rector A. GALEN's model of parts and wholes: experience and comparisons. Proc AMIA Symp 2000:714-8
12. Schulz S. Bidirectional mereological reasoning in anatomical knowledge bases. Proc AMIA Symp 2001:607-11