Aligning Multiple Anatomical Ontologies through a Reference

Songmao Zhang[1], Ph.D., Olivier Bodenreider[2], M.D., Ph.D.

[1]Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, P. R. China
[2]U.S. National Library of Medicine, NIH, Bethesda, Maryland, USA
[1]smzhang@math.ac.cn, [2]olivier@nlm.nih.gov

Objective: To investigate the feasibility of deriving an indirect alignment between two ontologies from the two direct alignments of these ontologies to a reference ontology. The three anatomical ontologies under investigation are the Adult Mouse Anatomical Dictionary (MA), the NCI Thesaurus (NCI) and the Foundational Model of Anatomy (FMA). **Methods:** The direct alignment employs a combination of lexical and structural similarity. The indirect alignment simply derives mappings from direct alignments to the reference ontology. Each of the three ontologies is used, in turn, as the reference and evaluated against the other two ontologies. **Results:** Number of direct mappings identified: MA-NCI: 715, MA-FMA: 1,353 and NCI-FMA: 2,173. Number of indirect mappings identified through the reference: FMA: 703, NCI: 771 and MA: 741. Mappings specific to direct and indirect alignments are presented and discussed. **Conclusions:** This study confirms the feasibility of aligning two ontologies through a reference ontology. We also show that both the number of concepts and the number of concept names in the reference ontology are important parameters determining the suitability of an ontology to serve as a reference for deriving indirect mappings.

1 Introduction

Mappings among ontologies constitute an enabling resource for applications such as knowledge sharing and application system communication. In particular, such mappings represent a crucial component of the Semantic Web in which the semantic annotation of resources will inevitably draw on multiple ontologies [1]. In previous work, we developed methods for aligning ontologies of anatomy, including the Foundational Model of Anatomy and the Adult Mouse Anatomical Dictionary, as well as the representation of anatomical entities in broader ontologies covering anatomy (GALEN, NCI Thesaurus and SNOMED CT) [2, 3].

While most ontology alignment techniques result in direct, pairwise mappings between ontologies, we have also demonstrated the feasibility of using one ontology as the reference in order to derive indirect mappings between two ontologies themselves mapped to this reference. More specifically, we developed an indirect mapping between the NCI Thesaurus and the Adult Mouse Anatomical Dictionary using the Foundational Model of Anatomy as the reference ontology [4]. The indirect mapping through the FMA was evaluated not only against the direct mapping between NCI and MA, but also against a gold standard alignment established manually between these two ontologies. The main finding of this previous study is that 91% of the matches identified by the direct alignment were present in the indirect alignment. Additionally, a small number of matches not present in the direct alignment were identified indirectly. Such additional matches come from additional synonyms and relations provided by the FMA that are not present in MA or NCI. In contrast, some matches are specific to the direct alignment, i.e., could not be discovered through the indirect alignment. Differing coverage and differing representation were identified as the causes for failure to find these mappings indirectly.

While encouraging, these results also raised the following question. Would we achieve a similar performance if NCI or MA – not FMA – were used as the reference ontology for deriving indirect mappings between the other two ontologies? The objective of this study is to examine this issue, and more generally, to assess the suitability of ontologies to serve as reference in an indirect alignment setting. To this end, we create three variants of our original experiment, using each of the three ontologies, in turn, as the reference to derive indirect matches between the other two.

Ontology matching is an active field of research. It is beyond the scope of this paper to give a detailed account of the various approaches proposed for aligning ontologies. For a survey of such techniques, the interested reader is referred to [5-9]. The most common approach to aligning ontologies is to create direct point-to-point mappings between concepts across ontologies, using a combination of lexical and structural methods (e.g., [10]). However, the role of reference ontologies in ontology alignment is also discussed in the literature. [11] suggests that a better solution for

creating, integrating and maintaining multiple local ontologies is to adopt a global reference ontology and a group of mapping rules between them. IF-MAP [12] is an ontology mapping system whose goal is to generate an isomorphism between local ontologies (populated with instances by different communities) and a reference ontologies (unpopulated). In contrast to this approach, we propose to map the "local ontologies" not only to the reference, but also to themselves, *through* the reference. More formally, we use the direct mappings of two ontologies O_1 and O_2 to a reference domain ontology O_R to derive an indirect mapping between O_1 and O_2. [13] proposes a similar approach, but for database integration purposes. Their system builds matchings between local database schemas and a reference ontology, and then composes these matchings to form mappings between schemas. Analogously, TAMBIS (Transparent Access to Multiple Bioinformatics Information Sources) uses ontologies to form a global schema over multiple heterogeneous resources [14]. Here the ontology forms a mechanism for building queries using a common ontological form which is mapped to each of the underlying resources. More recently, both [15] and [16] addressed the related issue of missing background knowledge in ontology matching. The former proposes a fully automatic solution by using semantic matching iteratively, while the latter first aligns the two ontologies with the background ontology, and then uses the structure of background knowledge to derive semantic relationships between the two ontologies.

3 Materials

The **Adult Mouse Anatomical Dictionary** (MA) is a structured controlled vocabulary describing the anatomical structure of the adult mouse [17]. It comprises 2,404 concepts. Each concept has one name (e.g., *Head/neck* and *Adrenal artery*). Additionally, 240 concepts have a total of 259 synonyms (e.g., *Limb* has synonym *Extremity*). The ontology is represented as a directed acyclic graph whose edges represent the relationships *IS-A* and *PART-OF*. Every concept is connected to others through *IS-A* or *PART-OF* relationships. The version used in this study was downloaded on December 22, 2004 (under the name Mus adult gross anatomy in the Open Biomedical Ontologies[1]).

The **NCI Thesaurus** (NCI) provides standard vocabularies for cancer research [18] and its anatomy class describes naturally occurring human biological structures, fluids and substances. The ontology is available in the Ontology Web Language (OWL). There are 4,410 anatomical concepts (accounting for about 12% of all NCI concepts). Every concept has a preferred name (e.g., *Abdominal esophagus*). 1,207 concepts have a total of 2,371 synonyms (e.g., *Orbit* has synonym *Eye socket*). Except for the root (*Anatomic Structure, System, or Substance*), every anatomical concept has at least one *IS-A* relationship to another concept. In addition, anatomical concepts are also connected by a *PART-OF* relationship (named *ANATOMIC STRUCTURE IS PHYSICAL PART OF*). The version used in this study is version 04.09a (September 10, 2004).

The **Foundational Model of Anatomy** (FMA) is an evolving ontology with an objective to conceptualize the physical objects and spaces that constitute the human body [19]. The underlying data model for FMA is a frame-based structure implemented with Protégé. 71,202 concepts cover the entire range of macroscopic, microscopic and subcellular canonical anatomy. In addition to preferred terms, 52,713 synonyms are provided (e.g., concept *Uterine tube* has synonym *Oviduct*). Every concept (except for the root) stands in a unique *IS-A* relation to other concepts. Additionally, concepts are connected by seven kinds of *PART-OF* relationships (e.g., *constitutional part of*, *regional part of*) and their inverses. For the purpose of this study, we considered as only one *PART-OF* relationship (with *HAS-PART* as its inverse) the various kinds of partitive relationships present in FMA. The version used in this study was downloaded on December 2, 2004.

4 Methods

We compare the direct alignment between two ontologies O_1 and O_2 to the indirect alignment automatically generated from mapping both O_1 and O_2 to O_R, the reference ontology. In practice, we perform: 1) three direct alignments O_1-O_2, O_1-O_R and O_2-O_R; 2) the indirect alignment between O_1 and O_2 through their direct alignments with O_R; and 3) a comparison of the direct alignment O_1-O_2 to the indirect alignment obtained through O_R. In [4], the FMA was selected as O_R, and MA

[1] http://obo.sourceforge.net/

and NCI as O_1 and O_2, respectively. In the present study, we examine the following two variants: NCI (O_R) with MA (O_1) and FMA (O_2), and MA (O_R) with NCI (O_1) and FMA (O_2).

4.1 Direct Alignment

Our approach to aligning two ontologies directly first compares terms across ontologies lexically in order to identify one-to-one concept matches. The second step is the identification of structural matches. The interested reader is referred to [3] for additional precisions about our method.

The **lexical alignment** compares two ontologies at the term level, by exact match and after normalization. Both preferred terms and synonyms in the two ontologies are used in the alignment. For example, the concepts *Heart valve* in MA and *Cardiac valve* (synonym: *Heart valve*) in FMA are identified as a match. Moreover, synonymy is used to identify additional matches. For example, *Cardiac valve* in NCI and *Heart valve* in MA, although lexically different, are considered a match because they name the same anatomical concept in the Unified Medical Language System® (UMLS®) [20].

The s**tructural alignment** first acquires the inter-concept hierarchical relationships, *IS-A* and *PART-OF*, and their inverses, *INVERSE-ISA* and *HAS-PART*, respectively. Missing relations are generated through complementation, augmentation and inference techniques [3]. Once all relations are represented consistently, the structural alignment is applied to the matches resulting from the lexical alignment in order to identify similar hierarchical paths to other matches across ontologies. For example, the match concepts *Heart valve* in MA and *Cardiac valve* in FMA exhibit similar hierarchical paths to other matches in these two ontologies, including paths to *Heart* (*PART-OF*) and to *Aortic valve* and *Mitral valve* (*INVERSE-ISA*). Such structural similarity is used as **positive evidence** for the alignment. Instead of similar paths, one match may exhibit paths to other matches in opposite directions in the two ontologies. Such paths suggest a structural conflict across ontologies. For example, in MA *Pericardial cavity* stands in a *HAS-PART* relation to *Pericardium*, while in the FMA *Pericardial cavity* is defined as a part of *Pericardial sac*, which is part of *Pericardium*. These conflicts are used as **negative evidence** for the alignment, indicating the semantic incompatibility between concepts across ontologies in spite of their lexical resemblance.

4.2 Indirect Alignment through a Reference

When a concept c_R from O_R is aligned with both a concept c_1 from O_1 ($\{O_1: c_1, O_R: c_R\}$) and a concept c_2 from O_2 ($\{O_2: c_2, O_R: c_R\}$), the concepts c_1 and c_2 are automatically aligned ($\{O_1: c_1, O_2: c_2\}$). The direct alignment MA-FMA identifies the match {MA: *Heart valve*, FMA: *Cardiac valve* (synonym: *Heart valve*)}, supported by positive evidence. The direct alignment NCI-FMA identifies {NCI: *Cardiac valve*, FMA: *Cardiac valve*}, also supported by positive evidence. Therefore, {MA: *Heart valve*, NCI: *Cardiac valve*} is derived automatically, through the FMA concept *Cardiac valve*, and supported by positive evidence in both direct alignments.

5 Results

Results for **three direct alignments** are summarized in section A of **Table 1**. The alignment NCI-FMA yielded the largest number of matches (2,173) and MA-NCI the smallest (715). A very small number of conflicts was identified in the two direct alignments to FMA; none in the direct MA-NCI alignment. In the three direct alignments, a vast majority of the matches (> 90%) was supported by positive structural evidence. No evidence (positive or negative) was found for 5-9% of the matches in three direct alignments. For example, although *Elbow joint* has relations to other matches in both MA (e.g., *PART-OF Forelimb*) and NCI (e.g., *PART-OF Skeletal system*), none of these paths are shared.

Results for the **three indirect alignments** are summarized in section B of **Table 1**. 703 matches between MA and NCI, 771 between MA and FMA, and 741 between NCI and FMA were automatically obtained from using FMA, NCI and MA as a reference, respectively. The majority of the three indirect alignments (88-92%) received positive evidence in both corresponding direct alignments they were derived from. 7-12% of them received no evidence and 0.4-1% received negative evidence in at least one of the direct alignments.

Taking the three ontologies pairwise, we compared the matches obtained in their direct alignment to the matches resulting from their indirect alignment through the reference. The results of

these **three comparisons** are summarized in section C of **Table 1**. For **MA-NCI**, 654 matches are shared by both alignments, leaving 61 matches specific to the direct alignment (accounting for 8.5% of the direct matches) and 49 specific to the indirect alignment through the FMA. For **MA-FMA**, 708 matches are shared by both alignments, leaving 645 matches specific to the direct alignment (accounting for 47.7 % of the direct matches) and 63 specific to the indirect alignment through the NCI. For **NCI-FMA**, 710 matches are shared by both alignments, leaving 1,463 matches specific to the direct alignment (accounting for 67.3% of the direct matches) and 31 specific to the indirect alignment through the MA.

88-89% of the shared matches in the three groups received positive structural evidence in all three direct alignments, e.g., {MA: *Heart valve*, FMA: *Cardiac valve*} in MA-FMA. Moreover, about 10-11% of the shared matches in the three groups received no evidence in at least one of the three direct alignments. For example, although linked to other matches in MA (e.g., *HAS-PART Lung*) and FMA (e.g., *HAS-PART Ear*), *Body* has no hierarchical paths to any other matches in NCI. This is why the matches of *Body* received no evidence in the two direct alignments MA-NCI and NCI-FMA, while receiving positive evidence in direct alignment MA-FMA. Lastly, nearly 1% of the shared matches in the three groups received negative evidence in one of the three direct alignments. For example, although a concept *Nephron* exists in the three ontologies, the corresponding match received negative evidence in the direct MA-FMA alignment (i.e., links to *Renal tubule* (synonym: *Uriniferous tubule*) through *HAS-PART* in MA but links to *Uriniferous tubule* through *PART-OF* in FMA), while receiving positive evidence in both direct alignments MA-NCI and NCI-FMA. Domain knowledge is required to evaluate such matches.

Table 1. Number of matches in the direct and indirect alignments

		MA – NCI	**MA - FMA**	**NCI - FMA**
A	Direct alignment	715 matches (91.3% positive evi.)	1,353 matches (94.8% positive evi.)	2,173 matches (90.1% positive evi.)
B	Indirect alignment	FMA as reference	NCI as reference	MA as reference
		703 matches (92% positive evi.)	771 matches (88.1% positive evi.)	741 matches (87.6% positive evi.)
C	Shared by direct & indirect alignment	654 matches	708 matches	710 matches
	Specific to direct alignment	61 matches	645 matches	1,463 matches
	Specific to indirect alignment	49 matches	63 matches	31 matches
D	Shared / direct alignment	91.5%	52.3%%	32.7%

6 Discussion

Alignment through a reference ontology is feasible and efficient. This study confirms the feasibility and efficiency of aligning two ontologies through a reference ontology. The proportion of matches from the direct alignment also identified in the indirect alignment is particularly good (91.5%) in the alignment MA-NCI with FMA as the reference. Assuming a good reference ontology is available, alignment through a reference is cost-effective: aligning *n* ontologies requires *n(n-1)/2* pairwise mappings, but only *n-1* mappings to a reference ontology. For five ontologies – which is a small number by Semantic Web standards – the difference already represents a 60% economy (4 vs. 10).

Suitability as a reference Ontology: Size matters. As shown in section D of **Table 1**, using the FMA as a reference resulted in the identification of a vast majority (91.5%) of the direct matches between MA and NCI. The large size of the FMA and its comprehensive set of synonyms contributed to this high percentage of mappings [4]. In contrast, when using NCI or MA as the reference in indirect alignment, only a fraction of the direct matches could be identified. Only one half (52.3%) of the corresponding direct matches were identified through the NCI and one-third (32.7%) through the MA as a reference. These findings confirm our intuition that ontologies offering a small number of concepts and a limited number of names for each concept are less suitable as a reference for deriving an indirect alignment between two ontologies. In the case of MA, for example, there are only 2,404 concepts and 2,663 names in comparison to over 70,000 concepts and 120,000 names in the FMA.

Every ontology, large or small, contributes specific indirect matches. Regardless of its size, as shown in section C of **Table 1**, every ontology contributes specific indirect matches, i.e., matches that are not identified in the direct alignment. For example, using MA as a reference generated 31 specific matches, of which 19 received positive evidence in both direct alignments. For

example, *Glomerular capillary* in NCI was not mapped directly to *Glomerulus* in FMA because the two terms are not synonyms in either ontology. However, the match {NCI: *Glomerular capillary*, FMA: *Glomerulus*} was identified indirectly when using the MA as a reference because *Glomerulus* and *Glomerular capillaries* are synonyms in MA. The match also received positive evidence in both direct alignments MA-NCI and MA-FMA. This indicates that the MA synonyms, although in relatively small number, play a significant role in the identification of mappings across two larger ontologies.

In summary, the most important finding of this study is that deriving an indirect alignment through a reference ontology is not only feasible, but also reasonably efficient. Moreover, this study confirms the intuition that both the number of concepts and the number of concept names in the reference ontology are important parameters determining the suitability of an ontology to serve as a reference for deriving indirect mappings. These findings are compatible with Burgun's "desiderata for domain reference ontologies in biomedicine", including good lexical coverage, good coverage in terms of relations and compatibility with standards [21].

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, and Key Laboratory of Multimedia and Intelligent Software at Beijing University of Technology. Thanks for their support to Cornelius Rosse, José Mejino and Todd Detwiler for the Foundational Model of Anatomy. Terry Hayamizu from the Jackson Laboratory contributed the evaluation of the direct mapping between NCI and MA.

References

1. Uschold, M., Gruninger, M.: Creating semantically integrated communities on the world wide web. Proc. International Workshop on the Semantic Web (2002)
2. Bodenreider, O., Hayamizu, T.F., Ringwald, M., de Coronado, S., Zhang, S.: Of mice and men: Aligning mouse and human anatomies. Proc AMIA Symp (2005) 61-65
3. Zhang, S., Bodenreider, O.: Aligning representations of anatomy using lexical and structural methods. AMIA Annu Symp Proc (2003) 753-757
4. Zhang, S., Bodenreider, O.: Alignment of multiple ontologies of anatomy: Deriving indirect mappings from direct mappings to a reference. Proc AMIA Symp (2005) 864-868
5. Doan, A., Halevy, A.Y.: Semantic integration research in the database community: A brief survey. AI Magazine **26** (2005) 83-94
6. Kalfoglou, Y., Schorlemmer, M.: Ontology mapping: the state of the art. Knowledge Engineering Review **18** (2003) 1-31
7. Noy, N.F.: Semantic integration: a survey of ontology-based approaches. SIGMOD Rec. **33** (2004) 65-70
8. Rahm, E., Bernstein, P.A.: A survey of approaches to automatic schema matching. VLDB Journal **10** (2001) 334-350
9. Shvaiko, P., Euzenat, J.: A survey of schema-based matching approaches. Journal on Data Semantics **4** (2005) 146-171
10. Noy, N.F.: Tools for mapping and merging ontologies. In: Staab, S., Studer, R. (eds.): Handbook on Ontologies. Springer-Verlag (2004) 365-384
11. Uschold, M.: Creating, integrating and maintaining local and global ontologies. Proceedings of the Workshop on "Applications of ontologies and problem-solving methods" at the 14th European Conference on Artificial Intelligence (ECAI 2000) (2000) 20-25
12. Kalfoglou, Y., Schorlemmer, M.: IF-Map: an ontology mapping method based on information flow theory. In: Spaccapietra, S. (ed.): Journal on Data Semantics (LNCS 2800), Vol. 2800 (2003) 98-127
13. Dragut, E., Lawrence, R.: Composing Mappings Between Schemas Using a Reference Ontology. In: Meersman, R., Tari, Z. (eds.): On the Move to Meaningful Internet Systems 2004: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2004, Agia Napa, Cyprus, October 25-29, 2004. Proceedings, Part I (LNCS 3290), Vol. 3290. Springer, Berlin / Heidelberg (2004) 783-800
14. Goble, C.A., Stevens, R., Ng, G., Bechhofer, S., Paton, N.W., Baker, P.G., Peim, M., Brass, A.: Transparent Access to Multiple Bioinformatics Information Sources. IBM Systems Journal Special issue on deep computing for the life sciences **40** (2001) 532-552
15. Giunchiglia, F., Shvaiko, P., Yatskevich, M.: Discovering missing background knowledge in ontology matching. In: Brewka, G., Coradeschi, S., Perini, A., Traverso, P. (eds.): Frontiers in Artificial Intelligence and Applications - Proceedings of the 17th European Conference on Artificial Intelligence (ECAI), Vol. 141. IOS Press (2006) 382-386
16. Aleksovski, Z., Klein, M., ten Kate, W., van Harmelen, F.: Matching unstructured vocabularies using a background ontology. In: Svatek, S.S.a.V. (ed.): Proceedings of the 15th International Conference on Knowledge Engineering and Knowledge Management (EKAW'06) (2006) (in press)
17. Hayamizu, T., Mangan, M., Corradi, J., Kadin, J., Ringwald, M.: The Adult Mouse Anatomical Dictionary: a tool for annotating and integrating data. Genome Biology **6** (2005) R29
18. De Coronado, S., Haber, M.W., Sioutos, N., Tuttle, M.S., Wright, L.W.: NCI Thesaurus: Using Science-based Terminology to Integrate Cancer Research Results. Medinfo **2004** (2004) 33-37
19. Rosse, C., Mejino, J.L., Jr.: A reference ontology for biomedical informatics: the Foundational Model of Anatomy. J Biomed Inform **36** (2003) 478-500
20. Bodenreider, O.: The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res **32 Database issue** (2004) D267-270
21. Burgun, A.: Desiderata for domain reference ontologies in biomedicine. J Biomed Inform **39** (2006) 307-313