From data to knowledge: a tool for clustering multi-scale resources for physiology research

João D. Ferreira[1], Bernard de Bono[2,3], Francisco M. Couto[1]

[1]Departamento de Informática, Faculdade de Ciências, Universidade de Lisboa, 1749-016, Lisboa, Portugal, [2]CHIME, University College London, Gower Street, London, WC1E 6BT
[3]ABI, 70 Sysmonds Street, City Campus, Auckland 1010, New Zealand

Correspondence: joao.ferreira@lasige.di.fc.ul.pt

Summary: This paper demonstrates a new web tool to compare physiology-related resources. Resources in physiology are by nature heterogeneous, as they make use of information from different domains of knowledge pertaining to different scales. Two such domains are gross and cellular anatomy. We describe our search tool, Multiple-Ontology Semantic Similarity (available at http://lasige.di.fc.ul.pt/webtools/mossy/) that automatically compares and clusters resources based on their anatomy-related annotations, thus taking into account their multi-scale nature.

• • •

Physiology research relies on knowledge across multiple scales, such as concepts related to organ systems, tissues, cellular types, chemical reactions and electricity, and the relationships between these concepts. For instance, nutrient intake depends on the characteristics of the intestine, the cells that are responsible for holding on to the nutrients and the kinds of metabolic reactions that affect these nutrients. While organizing this knowledge in meaningful and systematic ways ensures that the community can easily search and compare information, the heterogeneity of this information has made this effort a challenging task.

To mitigate this difficulty, we created a tool to compare and cluster resources annotated with anatomy-related concepts. For example, recent work has been carried out in histology to manage knowledge about Functional Tissue Units (FTUs), which are three-dimensional blocks of cells centered around a a small advective vessel, such that each cell in this block is within diffusion distance from any other cell in the same block [1]. FTUs can be annotated with gross anatomy and cell type concepts to respectively describe their location in the human body and the types of cell that they contain. Anatomy-based navigation of a database of FTU information can be achieved through a combination of similarity algorithms as well as clustering techniques applied to semantic metadata annotating FTU-related resources.

Semantic metadata for multi-scale models of cancer mechanisms are also annotated with similar anatomy-related concepts [2]. Comparing such models based on their annotations can provide an objective means to find patterns and associate certain characteristics of the models with related characteristics of FTU-based histology data, thus matching models with relevant data on the basis of their anatomical meaning.

The Multiple-Ontology Semantic Similarity (MOSSy) tool was created with two aims in mind: (i) compare resources annotated with concepts of the physiology domain, and (ii) cluster them based on their similarity. This clustering analysis facilitates the management of knowledge by grouping together resources that describe similar real-world cases. Furthermore, we plan to include a knowledge base of annotated resources in MOSSy, thus supporting the retrieval of resources that have a topic similar to a user query.

MOSSy exploits knowledge encoded in biomedical ontologies. Ontologies are collections of facts about a domain of knowledge, stored in a machine-readable way, which support categorizing of knowledge and automatic reasoning over the domain. An ontology contains (a) a set of concepts relevant for the domain, along with their names, synonyms and descriptions, and (b) the relationships between these concepts, most notably the class-subclass relationship between a concept and its specializations (*e.g.* the relationship between concepts "Heart" and "Organ"), and the partonomy relationship between a concept and its parts (e.g. the relationship between "Heart" and "Left atrium"). As such, ontologies are commonly depicted as graphs, where each node represents a concept and each edge represents a relationship between two concepts. Figure 1 contains a snippet of the Foundational Model of Anatomy (FMA), an ontology of the human anatomy.

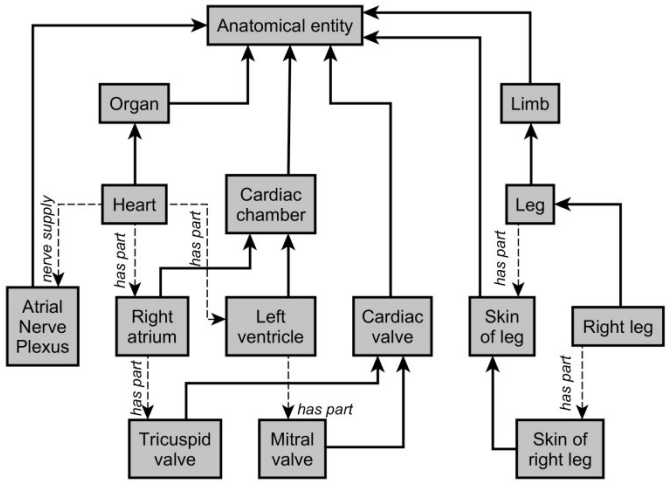

Figure 1 A snippet of the Foundational Model of Anatomy drawn as a graph. Nodes represent anatomical concepts and edges represent their relationships. Bold edges are class-subclass relationships; dashed edges have a label to describe the type of relationship.

In this context, ontologies provide two benefits. The first is the fact that they can be used as a stardard for metadata annotation, where the meaning of the concept used as annotation can be checked against its original definition in the ontology. This standardization means that integrating and sharing information whose metadata comes form ontologies is easier to do when compared to the case where the only available data is textual [3].

The second benefit is the fact that we can exploit the information they contain to compare resources, a technique known as ontology-based semantic similarity [4]. Consider again the ontology in Figure 1. With the information it contains, a resource about the heart can be regarded as more similar to a resource about cardiac chambers than to one about the skin of the leg. This idea of comparing resources based on their ontology-based metadata has been studied for several years, but one remaining open problem is multiple-domain semantic similarity.

MOSSy addresses the problem of comparing resources annotated with terms originating from multiple ontologies. This tool accepts two kinds of input: (i) annotated resources (*i.e.* semantic metadata) and (ii) textual resources, such as abstracts for automated annotation. By using an external text-mining web service (BioPortal's Annotator web service [5]), we extract from the given text a set of ontology concepts. MOSSy currently works with up to 8 different ontologies of the biomedical domain (spanning the domains of anatomy, cell lineage, protein function, biological processes, chemical compounds, human phenotypes, symptoms, human disease and biomedical investigation).

MOSSy provides a number of different similarity measures. In particular, it includes two different approaches to deal with the multiple-domain nature of the resources. It can compare two resources as if all their annotations come from the same ontology: to do this, it assumes that all the ontologies have a single root from which all the terms derive, and, as such, they can be considered as a single ontology. Alternatively, it can compare resources one ontology at a time, and then average these values into a single similarity value. For instance, given two resources with annotations from anatomy and cell lineage, this approach first compares the anatomical terms from the first model with the anatomical terms of the second model, then carries out the same procedure for cell lineage terms, and finally

averages the two values. For greater flexibility, users can choose the weight they wish to give to each ontology. Once similarity values have been calculated for the collection of resources input by the user, the tool offers a view of the resources clustered according to those similarity values.

As an illustration, we present in Figure 2 the screen-shots obtained after a user supplies with five different cancer-related models. In this case, the models consist of automatically annotating the first few paragraphs of the Wikipedia article whose name is the model name. It can be observed, for example, that the models are clustered according to the organ system that they affect (unrinary system, resporatory system, and digestive system).

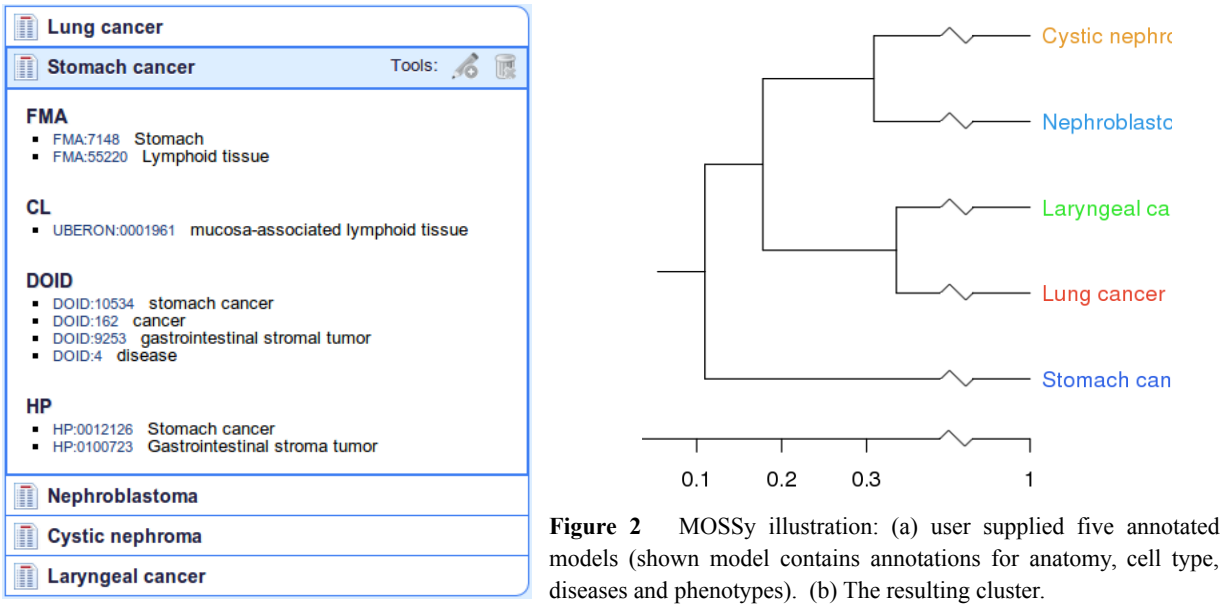

Figure 2 MOSSy illustration: (a) user supplied five annotated models (shown model contains annotations for anatomy, cell type, diseases and phenotypes). (b) The resulting cluster.

In conclusion, MOSSy supports physiology research by improving the organization of the existing knowledge through semantic similarity and clustering of the resources, in an attempt to provide (a) a mechanism through which to develop navigation facilities for knowledge bases, including things such as a section of "Related resources" or by presenting resources as part of larger groups of semantically similar resources, and (b) a means to detect patterns in real-world cases, such as annotated clinical cases, which can potentially be used to predict a prognosis or a best treatment approach.

References

1. de Bono B, Grenon P, Baldock R, Hunter P. Functional tissue units and their primary tissue motifs in multi-scale physiology. Journal of Biomedical Semantics 2013, 4:22. DOI:10.1186/2041-1480-4-22

2. http://chic-vph.eu/

3. Camon E, Magrane M, Barrell D, *et al*. The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. Nucl. Acids Res. 2004, 32 (suppl 1): D262-D266. DOI:10.1093/nar/gkh021

4. Pesquita C, Faria D, Falcão AO, *et al*. Semantic Similarity in Biomedical Ontologies. 2009. DOI: 10.1371/journal.pcbi.1000443

5. Jonquet C, Shah N, Youn C, *et al*. NCBO annotator: semantic annotation of biomedical data. International Semantic Web Conference, Poster and Demo session. 2009