Semantic models in VIP

Germain Forestier and Bernard Gibaud

November 15, 2010

Contents

Abstract

This document presents the early work on the semantic modeling for the ViP project. We present the different avenues of research concerning the use and integration of existing biomedical knowledge (ontologies, terminologies, etc.). We quickly review these different sources of knowledge and their use in already existing projects for medical information management. Then, we present how these knowledge sources can be specifically used in th ViP project and we propose a first model for the semantic representation of the entities in ViP. Finally, we detail the following steps of the modeling in the future of the project.

1 Introduction

Medical image simulation has become an essential tool to improve the understanding of biological processes, pathology diagnosis and treatment. Wide-scale availability of simulated medical images would greatly help designing new image acquisition techniques, developing realistic physiological models and validating image analysis procedures. The Virtual Imaging Platform (VIP) proposes to develop an environment enabling multi-modality, multi-organ and dynamic (4D) medical image simulation.

In this document we present the early work on the semantic modeling for the ViP project. Two different aspects are covered. Firstly, we review the major attempts of biomedical information modeling and their use in already existing projects. Secondly, we present our first study to identify the relevance of these resources for ViP, and the modeling of the semantic for the project.

In recent years, a strong interest has been given to the development and the formalization of biomedical knowledge. Even if formal representation is not yet the standard, the use of the tools provided by the semantic web (OWL/RDF) is increasing. In section 2, we review the major biomedical knowledge resources. In section 3, we present related work of semantic integration for the management of medical data. Finally, in section 4, we present the work carried out for the ViP project.

2 Existing biomedical knowledge resources

2.1 Heterogeneous representations of biomedical knowledge

The need of standard and well defined biomedical knowledge is not new but recent years has witnessed a strong intensification through different projects and initiatives. A relatively new trend is the development of biomedical ontologies [1] to store and formally represent medical knowledge. Many different projects aim at developing a controled vocabulary, specification and definition of medical terms. The BioPortal [1] project proposes on a unique website a great amount of resources. The objective of this portal is to ease the access and sharing of ontologies that are actively used in biomedical communities. The website counts 202 resources (in November 2010). There is however, a strong heterogeneity among these different sources, especially on the aim of the projects and the format in which the knowledge is available. Indeed, some of them are thesaurus, lexicon, list of terms or formal ontologies.

Each of these different formats of biomedical knowledge sources offer an interesting point of view on biomedical knowledge. However, it also brings confusion to the general user which can easly be lost in the proliferation of the information. Consequently, it can be difficult to select the right source (or sources) for a specific use. For example, biomedical resources can be used to control the vocabulary used in an application. In that case, even informal representation (e.g. a lexicon) could be use. However, many recent applications want to integrate biomedical knowledge in the development of semantic application where formal definition of the knowledge is needed. This is the case of the ViP project which wants to rely on the Web Ontology Language (OWL) as the standard format for the representation of ontology.

The Foundational Model of Anatomy Ontology (FMA) [12] is an evolving computer-based knowledge source for biomedical informatics. It is developed and maintained by he Structural Informatics Group at the University of Washington. The FMA proposes a comprehensive taxonomy of human anatomy. It also provides several axioms and definitions of conceptual attributes: part of, location, etc. The last version of FMA (v3.1) contains 81053 concepts and 176 properties. The main format of the FMA is Protege-FRAME which is a language allowing the development of frame-based domain ontologies. For some applications, there is a need for a formal representation of the knowledge following description logic (DL) principles. Consequently, there have been several attempts of converting the Protege-FRAME representation of the FMA to a formal representation in OWL. This conversion is far from being straightforward, as the Protege-FRAME representation and the OWL representation do not share the same philosophical concept (e.g. closed world assumption vs. open world assumption). The work of [4] proposed to transform the FMA exclusively in OWL-DL to allow efficient inference while [3] also introduced disjoint and closure axioms. The OWL version proposed on the FMA website was created by [11] and is composed of two parts. One part in OWL-DL and another one in OWL-Full which imports the OWL-DL version and adds extra knowledge which can only be represented using the expressiveness provided by OWL-Full.

The RadLex [5] project aims at building a lexicon for uniform indexing and retrieval of radiology information resources. The project is sponsored by the RSNA, which has enlisted the collaboration of other radiology organizations, including the American College of Radiology (ACR), to develop a comprehensive radiology lexicon. The last version of RadLex (v3.2.1) is proposed in Protege-FRAME on the BioPortal website. A translation of the version 3 is available in OWL-Full. However, there are no information about how this version has been converted from the Protege-FRAME version. The RadLex lexicon contains a terminology of radiology terms ranging from anatomy to radiology image acquisition and pathology. The structure of the project is however difficult to grasp and some choices of modelization are not explained in details. The project lacks of clear ontological foundation, but starts to rely on foundational ontologies.

Several other projects exist and it will be difficult to give a comprehensive listing. From the most popular we can cite the NCI (National Cancer Institute) thesaurus, the UMLS (Unified Medical Language System) and also the SNOMED CT (Systematized Nomenclature of Medicine Clinical Terms). The difficulty of moving from an unstructured representation to a formal representation in OWL is presented for the SNOMED-CT in [2].

[1] `http://bioportal.bioontology.org/`

2.2 Extracting information from existing biomedical knowledge

Some of the biomedical ontologies like the FMA are called *reference ontologies* and are not meant to be used directly in semantic information systems. Indeed, their aim is to comprehensively cover a domain and all the information present is not always interesting for a specific use. In that case, *application ontologies* have to be created or extracted from relevant subsets these reference ontologies. Most of the time, this extraction is done manually by an expert who selects the interesting resources to extract. However, some recent works like the vSPARQL [15] language propose an alternative in creating these application ontologies automatically through query over the reference ontologies. vSPARQL proposes to extend the SPARQL query language, which the standard to query RDF/OWL repositories. This extension consists in allowing to create *views* by adding new features to the SPARQL language (e.g. recursive query).

3 Semantic management of medical information

There has been a recent interest in creating knowledge-based information management system which deeply integrate semantic knowledge. The goal of these approaches is to systematically integrate semantic in the development of information management platform. This integration aims at creating new information systems which carry the semantic of the data along with the data itself through annotations and semantic modeling. These systems, which embed knowledge about the data (i.e. meta-data) can then propose intelligent ways of accessing and managing the data like reasoning.

In the biomedical field, a strong effort has been devoted to create platforms for medical image annotation and retrieval. The goal of these platforms is to annotate medical images using well-identified and controlled terminologies and ontologies. This meta-information can then be used to access the data using semantic query allowing the use of interesting features like reasoning to improve the results of the queries.

The recent, but very active, MEDICO project covers different aspects of access to medical information and proposes to use semantic technologies to developed semantic searches of medical databases. The project is supported by the THESEUS Program which is funded by the German Federal Ministry of Economic and Technology. In the frame of this project several propositions have been made to develop an interactive platform of medical image annotation and retrieval. Different existing biomedical knowledge resources are used, like the FMA for representing the anatomy, the RadLex lexicon to control the terminology of the radiology terms and the ICD-10 to represent diseases. In [9], the authors propose a system called RadSem using FMA, RadLex and ICD-10 for image annotation. An annotation model is proposed which aims at centralizing information provided by these three resources. A graphic user interface is proposed to help the radiologist to annotate the images and a query system is proposed. This query system uses the mechanism of *query expansion* to make the most of the semantic annotations. This query expansion system is used to retrieve resources which have been annotated by concepts related to the queried concept. For example a request on `Heart` related resources could retrieve resources annotated `Valve` as these two concepts are somehow linked in the FMA. In [14], the authors also included an image parsing module which consists in an automatic segmentation and annotation of medical images. They also recently proposed an OWL version of the ICD-10 classification of diseases [8] in order to help the annotation of the resources.

In a similar way, [13] proposes a model called AIM (Annotation & Image Markup) for semantic annotation of medical images. This work proposes an information model for image annotation and and image annotation tool. A related work presents a system named RadiO [6] which uses FMA and RadLex for medical image annotation. This project makes a clear distinction between the domain of the body and the medical image domain. An ad-hoc mapping between the concepts of FMA and Radlex is proposed to ensure semantic coherence of the annotation. A similar mapping is proposed by [7] which aims at comprehensively map the anatomical concepts of RadLex with the concepts of the FMA.

The use of FMA and RadLex for image annotation is proposed in the domain of neuroimaging in [17]. In this work, an application ontology is built by extracting information from the FMA to propose an ontology called NeuroFMA. This *view* of the FMA is then used to enhance RadLex through a mapping of the concepts of NeuroFMA and RadLex. These resources are then used to annotate neuroimaging data and to proceed

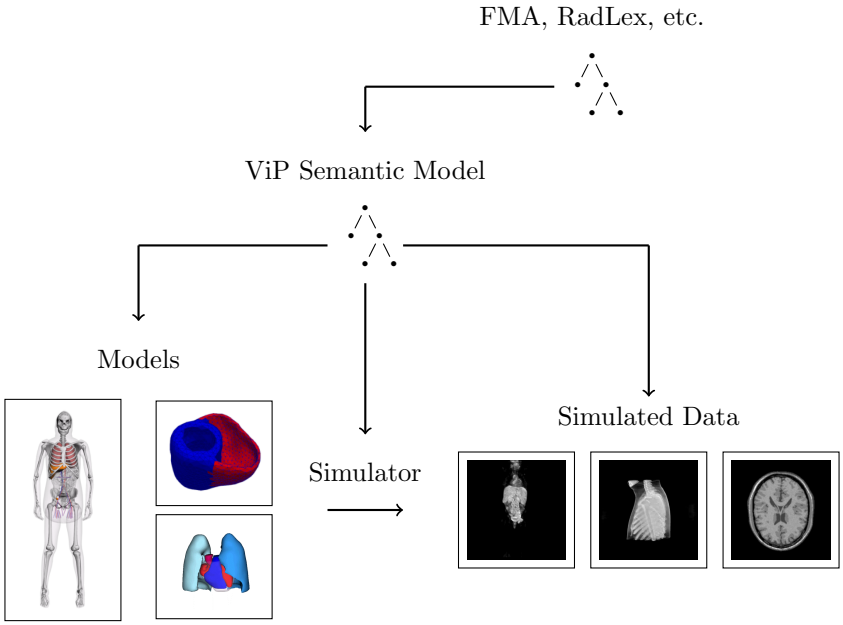

Figure 1: Illustration of the semantic integration in ViP.

to intelligent queries using the knowledge stored in FMA.

Semantic web technologies have also been used in [16] for the annotation of regions of interest in neuroimaging. This paper presents the design of a common semantic model providing a unified view on all shared data and tools. A multi-layered and multi-components formal ontology relying on DOLCE foundational ontology and several core ontologies of domains is proposed.

4 Semantic integration in ViP

4.1 Biomedical knowledge needs for ViP

Several needs of semantic integration have been identified throughout the ViP platform. Indeed, we would like to annotate all the resources which are involved in the platform ranging from the Models, to the Simulators and the Simulated Data. The figure 1 illustrates these different resources. According to the first analysis of the existing biomedical knowledge representation, we have to clearly identify the needs of the ViP project.

Concerning the reference to the anatomy, a need for well defined terminology is present. Indeed, the Models along with the Simulated Data will be annotated with the anatomical structure they represent in order to allow an easy access to them. The FMA appears to be a good candidate to represent this information. We have however to clearly identify how we are going to refer to the FMA and which subpart is related and useful for ViP.

Concerning the description of the Simulator and the Simulated data, the use of some terms extracted of the RadLex lexicon is envisaged. However, we also have to clearly identify which part of RadLex is relevant for the ViP project. However, the use of this resource should help to quickly progress in the modeling.

4.2 Concepts matching for semantic annotation in the Models

The use of biomedical terminology will be relevant in the definition of the Models and especially in the definition of the Intermediate Anatomical Model Format (IAMF) format. Indeed, most of the Models have

Model	#Terms	#Matchs FMA	#Matchs RadLex
Brainweb	10	7 (70 %)	4 (40 %)
Adam	22	16 (72 %)	14 (63 %)
Zubal CT	58	39 (67 %)	38 (65 %)
Zubal MR	63	40 (63 %)	39 (61 %)

Table 1: String matching between the terms used in the Models definition and the FMA/RadLex terminologies.

Model term	FMA term	FMAID
Cerebrospinal fluid	1. `Cerebrospinal_fluid`	20935
	2. `Cerebrospinal_vasculature`	73746
	3. `Synovial_fluid`	12277
	4. `Seminal_fluid`	62967
	5. `Tissue_fluid`	9673
	6. `Pericardial_fluid`	9887
	7. `Serous_fluid`	20932
	8. `Prostatic_fluid`	66884
	9. `Pleural_fluid`	12273
	10. `Peritoneal_fluid`	16515
Gray Matter	1. `Gray_matter_layer_of_neuraxis`	83142
	2. `Gray_matter_structure_of_tectum`	77868
	3. `Internal_gray_matter_component`	223151
	4. `Gray_matter_of_neuraxis`	67242
	5. `Gray_matter_of_telencephalon`	83911
	6. `Gray_matter_of_hypothalamus`	83915
	7. `Subcortical_gray_matter_structure`	61831
	8. `Right_intermediate_gray_matter`	74011
	9. `Gray_matter_of_pons`	83921
	10. `Gray_matter_of_midbrain_tegmentum`	83918

Table 2: Example of matching between two terms of the Models and the terms of FMA

to maintain a mapping between the simulated anatomical entity and their corresponding physical properties. These physical properties will be the information used to actually perform the simulation of the anatomical entity they are referring to. This mapping can therefore greatly benefit from the use of semantic annotation in order to clearly reference anatomical terms. For example, these references could be used to semantically query the models. To study the suitability of FMA and RadLex terminology for this purpose we carried out a string matching analysis between the terms used in the definition of several Models whose integration is foreseen in the ViP project and the terms available in FMA and RadLex. The string matching approach uses the open source framework Lucence [2] which uses the Levenstein distance to proceed to a fuzzy comparison of the string to evaluate their similarity. A similar approach has been used in [10] to match the terms of FMA and Radlex. For this first analysis, only perfect matches were considered. The table 1 presents the results of this matching. From these first results, we can conclude that almost 70% of the terms used in the Models definition can be automatically mapped in FMA/RadLex terminologies. A more detailed analysis would be necessary for the remaining terms. The table 2 gives an example of the matching for two terms. In this example, Cerebrospinal fluid has a perfect match while Gray Matter will need an expert intervention as various concepts are candidates. More advanced term matching strategy could be envisaged to improve the automatic discovery of the mapping [18]. However, a human interaction seems to be mandatory to ensure the semantic validity of the results.

[2] `http://lucene.apache.org`

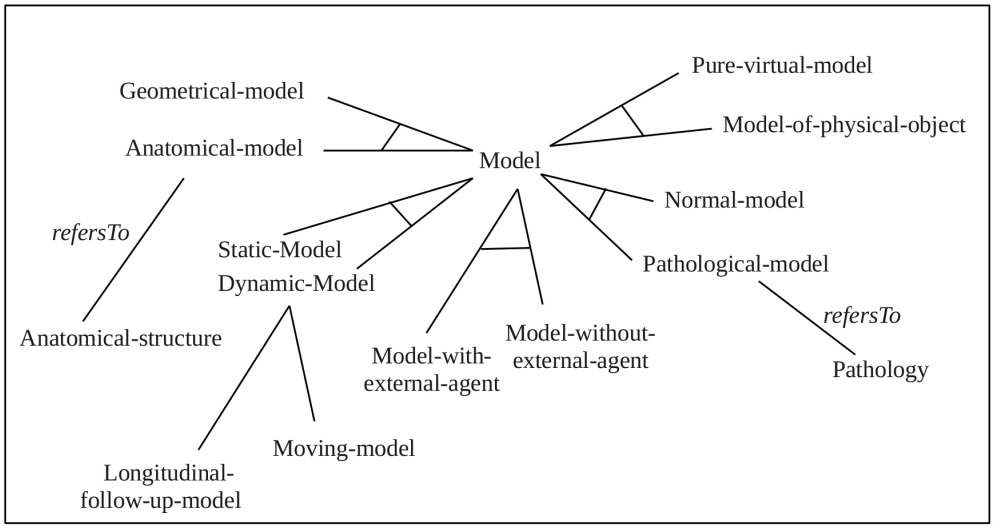

Figure 2: Model.

4.3 First semantic model

We also started to define the relevant entities and relations for the semantic modeling in ViP. The following diagrams represent the first attempt at modeling the different resources of the platform. We focused on the entities which are important according to the scenarios which presented the future use of the platform.

5 Conclusion

In this document we have presented the first step of the semantic modeling for the ViP project. We have presented several sources of existing knowledge that could be reused in the project. The first semantic model has also been proposed. The next step will be the release of a first version of the semantic model in a formal representation.

References

[1] O. Bodenreider. Biomedical ontologies in action: role in knowledge management, data integration and decision support. *Yearbook of medical informatics*, pages 67–79, 2008.

[2] O. Bodenreider, B. Smith, A. Kumar, and A. Burgun. Investigating subsumption in SNOMED CT: An exploration into large description logic-based biomedical terminologies. *Artificial Intelligence in Medicine*, 39(3):183–195, 2007.

[3] O. Dameron, D. L. Rubin, and M. A. Musen. Challenges in converting frame-based ontology into owl: the foundational model of anatomy case-study. In *AMIA Symposium*, pages 181–185, 2005.

[4] C. Golbreich, S. Zhang, and O. Bodenreider. The foundational model of anatomy in OWL: Experience and perspectives. *Web Semantics: Science, Services and Agents on the World Wide Web*, 4(3):181–195, 2006.

[5] S. Kundu, M. Itkin, D. A. Gervais, V. N. Krishnamurthy, M. J. Wallace, J. F. Cardella, D. L. Rubin, and C. P. Langlotz. The ir Radlex project: An interventional radiology lexicon–a collaborative project of

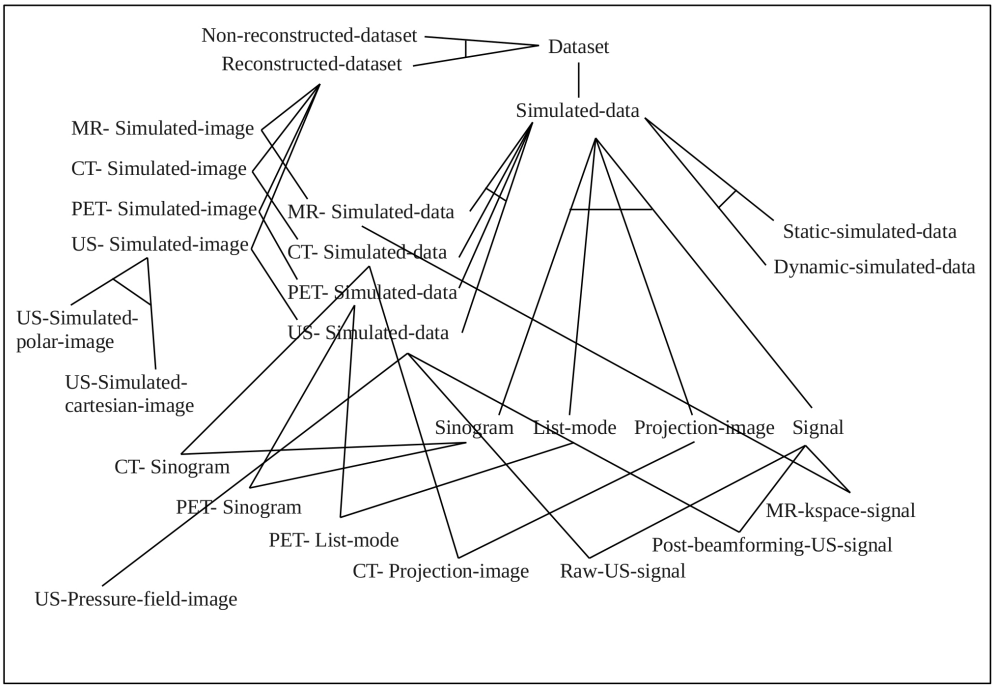

Figure 3: Dataset.

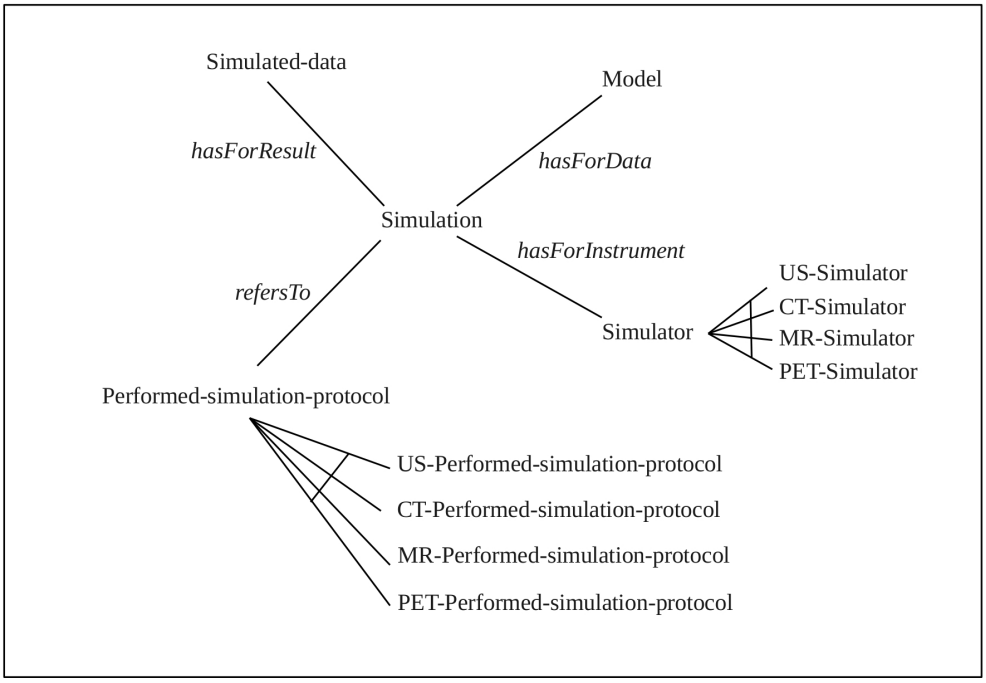

Figure 4: Simulation.

the radiological society of north america and the society of interventional radiology. *Journal of Vascular and Interventional Radiology*, 20(4):433–435, 2009.

[6] D. Marwede, M. Fielding, and T. Kahn. Radio: a prototype application ontology for radiology reporting tasks. In *AMIA Symposium*, pages 21–35, 2007.

[7] J. L. Mejino, D. L. Rubin, and J. F. Brinkley. FMA-Radlex: An application ontology of radiological anatomy derived from the foundational model of anatomy reference ontology. In *AMIA Symposium*, pages 465–469, 2008.

[8] M. Möller, P. Ernst, M. Sintek, R. Biedert, A. Dengel, and D. Sonntag. Representing the international classification of diseases version 10 in OWL. In *Proc. of the International Conference on Knowledge Engineering and Ontology Development (KEOD)*, 2010.

[9] M. Möller, S. Regel, and M. Sintek. RadSem: Semantic annotation and retrieval for medical images. In *European Semantic Web Conference, ESWC 2009*, pages 21–35, 2009.

[10] M. Möller, N. Vyas, M. Sintek, S. Regel, and S. Mukherjee. Visual query construction for cross-modal semantic retrieval of medical information. In *Malaysian Joint Conference on Artificial Intelligence (MJCAI)*, Kuala Lumpur, Malaysia, 2009.

[11] N. F. Noy and D. L. Rubin. Translating the foundational model of anatomy into OWL. *Journal of Web Semantics*, 6:133–136, 2008.

[12] C. Rosse and J. L. V. Mejino. The foundational model of anatomy ontology. In A. Burger, D. Davidson, and R. Baldock, editors, *Anatomy Ontologies for Bioinformatics*, volume 6 of *Computational Biology*, pages 59–117. Springer London, 2008.

[13] D. L. Rubin, P. Mongkolwat, V. Kleper, K. Supekar, and D. S. Channin. Annotation and image markup: Accessing and interoperating with the semantic content in medical imaging. *IEEE Intelligent Systems*, 24:57–65, 2009.

[14] S. Seifert, M. Kelm, M. Möller, S. Mukherjee, A. Cavallaro, M. Huber, and D. Comaniciu. Semantic annotation of medical images. In *Proc. of SPIE Medical Imaging*, 2010.

[15] M. Shaw, L. T. Detwiler, N. Noy, J. Brinkley, and D. Suciu. vSPARQL: A view definition language for the semantic web. *Journal of Biomedical Informatics*, In Press, 2010.

[16] L. Temal, M. Dojat, G. Kassel, and B. Gibaud. Towards an ontology for sharing medical images and regions of interest in neuroimaging. *Journal of Biomedical Informatics*, 41:766–778, 2008.

[17] J. A. Turner, J. L. V. Mejino, J. F. Brinkley, L. T. Detwiler, H. J. Lee, M. E. Martone, and D. L. Rubin. Application of neuroanatomical ontologies for neuroimaging data annotation. *Frontiers in Neuroinformatics*, 4(0):12, 2010.

[18] P. Wennerberg, M. Möller, and S. Zillner. A linguistic approach to aligning representations of human anatomy and radiology. In *International Conference on Biomedical Ontologies (ICBO)*, 2009.