Toward an Ontological Database for Subcellular Neuroanatomy*

Amarnath Gupta, Stephen D. Larson, Christopher Condit, Sandeep Gupta, Lisa Fong, Li Chen, and Maryann E. Martone

University of California San Diego, La Jolla, CA 92093, USA gupta@sdsc.edu, slarson@ucsd.edu, condit@sdsc.edu, sandeep@sdsc.edu, lfong@ncmir.ucsd.edu, lichen@sdsc.edu, mmartone@ncmir.ucsd.edu

Abstract. We present the semantic data model for an ontological database for subcellular anatomy for Neurosciences. The data model builds upon the foundations of OWL and the Basic Formal Ontology, but extends them to include novel constructs that address several unresolved challenges encountered by biologists in using ontological models in their databases. The model addresses the interplay between models of space and objects located in the space, objects that are defined by constrained spatial arrangements of other objects, the interactions among multiple transitive relationships over the same set of concepts and so on. We propose the notion of parametric relationships to denote different multiple ways of parcellating the same space. We also introduce the notion of phantom instances to address the mismatches between the ontological properties of a conceptual object and the actual recorded instance of that object in cases where the observed object is partially visible.

1 Introduction

An Ontology-based database (OBDB) is a new class of information systems that consists of a domain ontology and a database content that references the ontology, such that the system can be queried both as a database and through the ontology. In recent work, [1,2] have developed an object-oriented framework for managing and querying OBDBs. We consider this class of systems to be especially important for life science applications and develop elements of a semantic data model to address a number of specific modeling issues we have encountered in developing ontology-based databases for neuroscience. A primary difference between the work in [1,2] and ours is that we consider the ontology to be much richer, specified in terms of OWL-DL or any other SHOIN-compliant description logic, supplemented by constructs and rules of entailment from the Basic Formal Ontology (http://www.ifomis.uni-saarland.de/bfo) that describes, among other things, a formalization of space and objects in space.

This paper investigates how the semantic data model of an ODBMS is influenced when the application domain contains information about interacting

^{*} This work is supported by NIH BIRN-CC Award No. 8P41 RR08605-08S1, NIH Human Brain Project Award No. 5RO1DC03192.

extended objects in space. To focus the issue around a life science problem, we consider the domain of neuroanatomy as our application context.

Application Domain. To situate the application context of our problem domain, consider that one of the fundamental organizing principles of biological information is anatomical, i.e., based on the physical structure of living beings, and how these entites relate to each other. Anatomy provides one of the *covering* disciplines that spans every single entity in any biological organism. Further, a number of biological functions can be associated with anatomical entities, the functional relationship between two entities is often effected through their anatomical relationship. Anatomy presents a number of interesting issues. As a simple example, any anatomical structure (at least for multicellular organisms) can be recursively broken down into increasingly finer substructures – starting from substructures that are at the same scale (i.e., measured by the same spatial units) like forearm and phallanges, and extending to substructures that span scales such as tissues and cellular ultrastructures. This meronymic continuum is complemented by other relationships among anatomical enities such as physical and chemical connectivity, functional groups, and cytoarchitectural associations. It is the task of a conceptual model to capture the intricacies of their relationships to various biological functions. Because of the central role of anatomy in structuring biological information, it may also be used to serve as a common backbone for problems in biological information integration. For example, an organ-level biological database and a cell-level biological database can be semantically integrated through the multi-granular meronymy of an anatomical ontology.

The central role of anatomy in modeling and understanding biological systems has inspired a number of formal modeling efforts. For example, the cellular component fragment of the well-known Gene Ontology (http://www.geneontology.org/) captures is-a and part-of relationships among intracellular structures of generic cells. The Foundational Model of Anatomy (FMA), on the other hand, captures gross anatomical object including spatial relationships like has_boundary. These two ontologies overlap in that the cell-level terms of the FMA is a subset of that in Gene Ontology. And yet, their combined collection of terms and relationships leave some holes in the eyes of the neuroanatomist. The present work is an extension and re-formalization of SAO [3], the ontology for subcellular anatomy arises from our recognition that (a) these ontologies do not reflect the complexity needed to express cellular and subcellular neuroanatomical information, and (b) there is no available ontology that formally reflects *mesoscale* information, where those structures that sit between more gross anatomical scales and the level of individual protein and other macromolecules. In our example area of the nervous system, it comprises the dimensional range of nanometers to microns, encompassing cellular networks, subcellular microdomains, and their macromolecular constituents. These spatially extended structures lie at the heart of information processing in the nervous system, providing the adaptive spatial framework in which molecular and biochemical processes occur. In order to uncover the complex, detailed structural and dynamic inter-relations critical to the functional properties of nervous system, entities at this scale must be observed and described, and inferred using a combination of experimental observations and formal treatment. The purpose of the semantic modeling framework proposed in this paper is to create an OBDB that serves this goal.

Contributions. In this paper we propose the notion of parametric relationships to denote different multiple ways of parcellating the same space. We also introduce the notion of phantom instances to address the mismatches between the ontological properties of a conceptual object and the actual recorded instance of that object in cases where the observed object is partially visible.

2 Data Model

2.1 Preliminaries

An OBDB consists of the triple (S, \mathcal{O}, μ) , where S is a schema, \mathcal{O} is an ontology and μ defines a set of mappings between \mathcal{O} and S. For simplicity, we start with the assumption that S is a relational schema containing relations $R_1, R_2 \dots$ We also make the simplifying assumption at this point that we have a single ontology \mathcal{O} , and that the ontology roughly corresponds to the SHIN description logic. The focus of this paper is to define a variety of mapping relationships μ that can be concretely used for anatomical information.

Any relation $R_i(\mathbf{A}_o, \mathbf{A}_d)$ may have two kinds of attributes – \mathbf{A}_o is the subset of attributes that are *ontology-mapped*, while \mathbf{A}_d is the subset of attributes having data domains. If $R_i.A_j$ is an ontology-mapped attribute of relation R_i , the domain of $R_i.A_j$ is defined to be an ontology expression that specifies which part of the ontology will provide values for the attribute. In our example domain, consider the relation neuron-image containing images of traced neurons.

```
neuron-image(image-id, image-type, neuron-name, x-size, y-size)
```

where **neuron-name** is mapped to the subcellular anatomy ontology (SAO) and the mapping μ_{dom} is described as:

```
\mathsf{dom}(\mathsf{neuron-image.neuron-name}) = \mathsf{SAO:subclass*('neuron')}
```

where 'neuron' is a term of the SAO (Subcellular Anatomy Ontology) [3] ontology and subclass* refers to the transitive closure of the subclass relationship of SAO. We will present more examples of ontology-expressions as we proceed. Suppose t (e.g., 'basket cell') is a term belonging to SAO:subclass*('neuron') and r_i is a tuple in neuron-image such that r_i .neuron-name = t. We say tuple r_i is an evidence-of t. The set of all evidences of a term t in a relation R is called the range of t in R. Why should one invent yet another relationship like evidenceof? Semantically, it will be incorrect to say that tuple r_{20} of neuron-image is an instance of 'basket cell', for example, because r_{20} is actually an instance of class image which happens to depict a basket cell. Thus this image is like a container of an instance of an ontological object. We capture this semantic relationship through the distiguished relationship evidence-of. Now consider the relation that stores information about all structures that have been segmented from images:

segments(structure-name, cell-name, image-id, length, volume)

where the domain of structure-name is SAO:component-of*('neuron') and the domain of cell-name is SAO:subclass*('neuron'). Note that has-component (hc) is a kind of partonomic relationship such that the component has a specific function within the cell. Clearly, every tuple of the relation segments is an evidence of some 'neuron' and some 'component'. But in our model this depicts a stronger semantic association. Since the primary key of the segments relation consists of two attributes having domains in two parts of the SAO ontology, this implies that every tuple in segments is an *instance-of the direct or inferred path connecting the corresponding neuron and the structure of the ontology*. Let us say the SAO ontology has the following entries:

Purkinje-neuron \square neuron dendrite \square neural-component neuron $\equiv \forall$ has-component.neural-component neuron $\equiv \exists$ has-component.dendrite

Here Purkinje-neuron $\xrightarrow{\square}$ neuron \xrightarrow{hc} dendrite is a path p_1 . Consequently, Purkinjeneuron \xrightarrow{hc} dendrite is a path p'_1 in SAO. If the relation segments has tuples:

 r_1 : segments(dendrite, Purkinje-cell, 12, 4.67, 2.11) r_2 : segments(dendrite, Purkinje-cell, 39, 3.93, 1.52)

we can say r_1, r_2 are instances of path p'_1 (or p_1 for that matter). However, it is not meaningful to create instances of all paths in an ontology. For example, while it is possible to have an instance of $\exists R.C_1$ it is not possible to have an instance of $C_2 \sqcup \exists R.C_1$. In general, instances of paths containing \sqcap or \sqcup are disallowed. Instantiations of more complex concepts are possible with more specialized constructs. We will return to this issue when we present our ontology model for aggregates in Section 2.2.

2.2 A Semantic Model for Extended Objects

Anatomical objects are spatial objects – they are located in a spatial context, occupy space, and maintain functional relationships with other objects often by virtue of their (relative) positions in space. It is therefore imperative that both the ontology component and the schema component of an OBDB has an adequate semantic model of extended objects and their spatial context. The fact that spaces and objects are conceptually distinct but are related intricately through relationships of *spatial inclusion, meronymy* and *topology* has been a topic of philosophical research [4,5,6]. In our work, we encapsulate all this development through our adoption of the Basic Formal Ontology (BFO) [7], described next.

Basics. BFO serves as an upper-ontology for biomedical science. BFO uses a concept called **continuant** for everything that is not dependent on time. Two subcategories of **continuant** relevant to us are **spatial-region**, that represents space,

and independent-continuant, that represents material objects or named portions of objects. According to the BFO catergorization, spatial-region represents space that is not relativized, i.e., not specified with respect to any object (e.g., extracellular space surrounding the post-synaptic density). In contrast, independentcontinuants have some characteristic shape (that may be specified with respect to other continuants) in which they are completely enclosed, and are further subcategorized into object, object-boundary, site, object-aggregate and fiat-objectpart. Briefly, sites are entities that can be occupied by other continuant entities (e.g., a post-synaptic-density is also occupied by the dendrite of which it is a part); objects are independent, spatially extended, maximally self-connected and self-contained entities, and possesses an internal unity; object-boundary (e.g., the outer membrane of a mitochondrion) is an independent continuant entity that is a lower-dimensional part of some other continuant entity; an object-aggregate is a named independent continuant entity that is a mereological sum of separate objects (e.g., a gap junction consists of the two membranes of two different neurons apposed to each other); finally, a fiat-object-part is a named part of an object that does not have an identifiable boundary (e.g., the distal dendrites of a neuron). In addition to these entities, BFO also admits a number of relationships – the two relevant ones are part-of that holds between two continuants or between two spatial-regions and located-in, that holds between an independent continuant and a spatial-region.

Recently [8] has investigated the problem of representing space and extended objects in the light of biomedical ontologies. They introduced the primitive function region-of, which given an extended object instance e, returns the spatial-region s that e fully occupies, such that

$$\forall x, y : \text{located-in}(x, y) =_{def} \text{part-of}(\text{region-of}(x), \text{ region-of}(y))$$

For extended objects they also distinguish between two forms of located-in, namely, location due to parthood and that due to spatial containment. If *o* is the instance of a spatial extended object, they define:

 $\forall x, y : \text{contained-in}(x, y) =_{def} \text{object}(x) \land \text{object}(y) \land \text{located-in}(x, y) \land \neg \text{part-of}(x, y)$

Our semantic data model adopts this ontological framework and extends it to provide a more realistic ontology to instance mapping constructs that can model extended data objects observed in images and videos and stored in the database.

Multiple Paritioning. As in BFO, in our model, a spatial-region, extended object, or site can be partitioned using part-of relationship. However, our anatomymotivated viewpoint persuades us to model spatial partitioning in a finer detail. The following snippet shows an example of our partition specification:

```
\begin{array}{l} \mathsf{s1} \sqsubseteq \mathsf{SAO:site} \\ \mathsf{p1} \equiv \exists \ \mathsf{complete-partition-of.s1} \\ \mathsf{p2} \equiv \exists \ \mathsf{partial-partition-of.s1} \\ \mathsf{s11} \equiv \forall \ \mathsf{part-of(p1).s1} \\ \mathsf{s12} \equiv \forall \ \mathsf{part-of(p1).s1} \\ \mathsf{s21} \equiv \forall \ \mathsf{part-of(p2).s1} \\ \mathsf{disjoint(s11,s12)} \end{array}
```

The first line establishes s1 as a SAO:site (which is the same as a BFO:site). The next two lines show that our role hierarchy for the partition-of relationship has two values complete-partition-of and partial-partition-of. The first sub-role means that all the spatial parts that constitute the particular partition-of relationship are completely enumerated in the ontology, while the latter sub-role states that this is not the case. The next three lines illustrate a *parametric role* or a *parametric relationship*, our extension to the standard OWL-DL. It simply states that s11 and s12 are parts of site s1 according to the partitioning scheme p1, whereas s21 is part of site s1 according to the partitioning scheme p2. The last line implies that our partition-of relationship does not assume spatial disjoint-edness of the parts – otherwise this constraint must be explicitly asserted as in this example. Also note that although p1 is a complete partitions11 s12 are not the only parts of s1 because we don't claim the OBDB is completely known.

Multiple Inheritance of BFO Categories. Our model permits multiple inheritance from the basic BFO categories. For example, the class synapse inherits from both an aggregate-object and from a fiat object part because a synapse is a *junction*, i.e., it is a portion of extracellular space (hence it has no demarcated boundaries) where axon terminals and dendritic processes are situated (hence it is an aggregate) closely enough such that chemical neurotransmitters can pass from the axon terminals to the neurotransmitter receptor portions (e.g., post-synaptic density) of those dendrites. We show the ontological definition of synapse in Figure 1. Other neuroanatomical entities like the *gap junction* or the *node of Ranvier* share the same characteristics – they are aggregates of fiat object parts such that certain predicates over selected properties of the participating objects hold.

Implicit Subclasses. Our model also supports the creation of *implicit subclasses* but considering all possible values of a class based on its data properties. Consider a class called post-synaptic-density that has a data property called *morphological-type* whose value is a 2-tuple (m_1, m_2) where m_1 can be 'symmetric' or 'asymmetric' and m_2 can be 'macular' or 'perforated'. When implicit subclasses are used, the system implicitly creates a cross-product of 4 subclasses of post-synaptic-density. If one of the system-generated classes is not valid, the ontologist has to specify a constraint like invalid-class('asymmetric_perforated_post-synaptic-density'). One might argue that this creates a huge proliferation of possible classes. We view this as a necessary evil because thanks to the bottom-up nature of biological discoveries, very often a scientist would discover a new characteristic of an biological entity that had not been materialized as a separate class before because there was no prior evidence that such a class would be biologically significant.

2.3 Bridging the Instantiation Gap

We return to the issue of mapping between the ontology part and the database part of an OBDB in the context of our semantic model.

An implicit assumption in almost all ontology literature is that the primary technical problems lies in the expressive power and computational complexity



Fig. 1. A rough ontological definition for the concept synapse

related to the specification of the ontology and the reasoning over instances. We introduce a problem called the *instantiation gap* problem between the ontology and its instance repository, which in an OBDB, has a schema constructed independently of the ontology. To our knowledge, this problem has not been considered by prior research. Simply, an *instantiation gap* occurs when either of the following cases hold:

- 1. an instance store has a data object d marked as instance of a concept c from the ontology, but one or more of the following occurs:
 - (a) d does not have all the inherited properties of c
 - (b) d does not have all the local properties of c
 - (c) d's extended object properties violate the expected object properties of \boldsymbol{c}
- 2. an instance store has instantiations of all properties that an instance of concept c from the ontology is supposed to have, but there is no data object d in the instance store that can be assigned as an instance of c.

It is easy to see that all three conditions under case (1) may occur when an image depicts one or more extended object that are partially visible. We show that case (2) also occurs due to partially visible structures. For example, consider an image (image-id = 23) showing a chemical synapse from a granule cell to a Purkinje cell as part of its content. As a constituent of the synapse, the presynaptic axon terminal exists in the image but is partly visible, the postsynaptic density of the synapsing dendrite is also partly visible, other dendritic processes of the same postsynaptic neuron are also partial. Let us imagine the image also shows a number of synaptic vesicles, some in the synaptic cleft, some in the presynaptic axon, and the rest in the extracellular space as well as parts of other neural structures such as the endoplasmic reticulum. Now consider Table 1, a slightly extended version of the segments relation we saw earlier, with the contents as recorded in from the scientist's annotation of image 23.

Obj-ID	Structure-name	Cell-name	Image-id	Visibility	length	volume	class
1001	axon-terminal:02	granule-cell:21	23	partial			
1002	synaptic-vesicle:45	granule-cell:21	23	partial			fused
1003	psd:15	Purkinje-cell:19	23	partial			
1004	dendritic-process:25	Purkinje-cell:19	23	partial			
1005	dendritic-process:26	Purkinje-cell:19	23	partial			
1006	dendritic-spine:58	Purkinje-cell:19	23	partial			thin
1007	cell-membrane:34	granule-cell:21	23	partial			
1008	synaptic-vesicle:46		23	$\operatorname{complete}$		0.10	normal
1009	synaptic-vesicle:47		23	complete		0.11	
1010	synaptic-vesicle:48		23	partial			
1011	synaptic-vesicle:48		23	partial			
1012	synaptic-cleft:03		23	complete			
1013	ER:09	granule-cell:21	23	partial			
1014	extracellular-space:28		23	partial			none

Table 1. A portion of the segments table showing the experimenter's data about the synapse

Notice that since the database has been developed to support experimental documentation and ontology has been developed to model the scientist's perception of biological reality, there are a number of disconnects between the ontology's depiction and the database's depiction of the objects. First, for oid = 1003, the instance of the **post-synaptic-density** could not be assigned any classification (e.g., 'symmetric_macular') due to the partial observation, although the ontological data property of the class posits this as a mandatory property¹ A different mismatch occurs for oid = 1001. We know from the ontology that an **axon-terminal** is a part of an **axon** and can't really exist without the axon. However, the database does not indicate the presence of the axon. In absence of the axon however, oid = 1007 can't be considered to be an instance of **axonal-membrane**, which is defined as the part of the cell-membrane covering the axon, and is a mandatory part of **axon**. Along the same lines, the fiat aggregate object called **synapse** as defined in the ontology (Figure 1) does not exactly exist, although pieces that consitute the synapse do.

To bridge this instantiation gap, our system computes an additional conceptto-instance mapping called *unobserved-instance-map* that first copies partially observed extended objects, and then recursively fills-in *phantom instances* of

¹ Notice that in contrast, the classification for the dendritic spine could be filled in even if the spine is not completely visible.

objects for which there is sufficient evidence (more than x% of mandatory properties). This would create an axon instance and an neuron instance given the axon terminal. The expected properties of these phantom instances are then matched against the observed properties in the database to derive further correspondences between the concepts and properties in the ontology to the data records. This will instantiate the synapse when most of its expected components are located. When the process terminates, the unobserved-instance-map is committed to the database, that can be verified by the domain expert.

3 Querying the OBDB with OntoQuest 2.0

In [9], we presented OntoQuest 1.0, the first version of our ontology exploration software. OntoQuest 1.0 is built upon the IODT framework from IBM (http://www.alphaworks.ibm.com/tech/semanticstk) and allows navigational queries over the ontology content and SPARQL+aggregate queries over instances in distributed stores. In general however, one should be able to pose arbitrary ad hoc queries to an OBDB that permit ontology navigation and instance fetching together, and utilize all the concept-to-instance mappings that are available.

OntoQuest 2.0 uses a storage model for the concept graph that is similar to that of IODT. In this model, all distinguished relationships permitted by OWL (e.g., subclass, allValuesFrom, disjoint ...) are stored in separate tables, while all user-defined relation names are stored in a quad-store. However, the instances (both observed and phantom) are stored in a graph-like manner. The properties of the nodes and the edges are stored in relational column stores, while the connectivity of the nodes and edges are stored in triple stores with additional path indexes. The mappings between the concepts and the relations are maintained using standard primary key foreign key relationships between the concept tables and the node/edge property tables.

OntoQueL, the query language of OntoQuest 2.0 is influenced by SPARQL, OntoQL [1], OWL-QL [10]; since a full treatment of the language is beyond the scope of this paper, we illustrate it with examples.

The query select succ*(concept:'neuron', part-of, 3) from MyOnto as graph returns a 3-level deep part-of hierarchy from the concept-graph starting with the term 'neuron'. If we want to get the part-of hierarchy of all *observed instances* of the concept 'neuron', we will write the query as select succ*(instance-of('neuron') \$X, part-of, 3) from MyOnto as graph where observed(\$X). Due to the builtin predicate observed, the query does not return phantom instances, but does return both complete and partially visible instances. If partially visible instances are not desirable, we add the condition not(partial(\$X)).

OntoQueL also permits *instance navigation* queries. The query select neighborhood(instance:'Purkinje-cell' X) as G from MyOnto where X.length > 4 stop before (exists Y and phantom-instance(Y) and Y in G) starts with each instance of Purkinje-cell satisfying the length predicate and follows graph edges until it touches a phantom instance of any kind. The condition stop before ensures that the phantom instance does not get included

in the result graph; if want to stop the graph navigation after the phantom node is included, we will use the construct **stop when** instead of **stop before**.

4 Conclusion

In this paper, we presented our model of capturing the semantics of an ontologybased database system, and the extensions we made to the standard OWL and BFO semantics. We introduced the problem of instantiation gap and presented our current solution. However, we believe more elegant solutions to the problem should be possible. We briefly discussed how the OntoQueL, the query language of the OntoQuest 2.0 system treats concept-grapha and instance graph queries, and allows the formulation of more complex queries that current ontology-based query languages do not support.

References

- Jean, S., At-Ameur, Y., Pierra, G.: Querying Ontology Based Database Using OntoQL (an Ontology Query Language). In: Ontologies, DataBases, and Applications of Semantics (ODBASE'2006), pp. 704–721 (2006)
- Jean, S., Dehainsala, H., Xuan, D.N., Pierra, G., Bellatreche, L., At-Ameur, Y.: OntoDB: It is Time to Embed your Domain Ontology in your Database. In: Database Systems for Advanced Applications (DASFAA 2007) (2007) (demo paper)
- Fong, L.L., Larson, S.D., Gupta, A., Condit, C., Bug, W., Chen, L., West, R., Lamont, S., Terada, M., Martone, M.E.: An ontology-driven knowledge environment for subcellular neuroanatomy. In: 3rd Int. Workshop on OWL: Experiences and Directions, Innsbruck, Austria (2007)
- Casati, R., Varzi, A.: The structure of spatial localization. Philosophical Studies 82, 205–239 (1996)
- Casati, R., Smith, B., Varzi, A.C.: Ontological tools for geographic representation. In: Guarino, N. (ed.) Formal Ontology in Information Systems, pp. 77–85. IOS Press, Amsterdam (1998)
- Smith, B.: Mereotopology: a theory of parts and boundaries. Data Knowl. Eng. 20, 287–303 (1996)
- Spear, A.D.: Ontology for the Twenty First Century: An Introduction with Recommendations. The Institute for Formal Ontology and Medical Information Science (2006), http://www.ifomis.uni-saarland.de/bfo/manual
- 8. Schulz, S., Marko, K., Hahn, U.: Spatial location and its relevance for terminological inferences in bio-ontologies. BMC Bioinformatics 8, 134 (2007)
- Chen, L., Martone, M.E., Gupta, A., Fong, L., Wong-Barnum, M.: Ontoquest: Exploring ontological data made easy. In: VLDB. Proc. 31st Int. Conf. on Very Large Databases, pp. 1183–1186 (2006)
- 10. Fikes, R., Hayes, P.J., Horrocks, I.: OWL-QL a language for deductive query answering on the semantic web. Journal on Web Semantics 2, 19–29 (2004)