On Querying OBO Ontologies Using a DAG Pattern Query Language*

Amarnath Gupta and Simone Santini

San Diego Supercomputer Center
University of California San Diego, La Jolla, CA 92093, USA
{ssantini, gupta}@sdsc.edu

Abstract. The Open Biomedical Ontologies (OBO) is a consortium that serves as a repository of ontologies that are structured like directed acyclic graphs. In this paper we present a language DQL for querying a database of directed acyclic graphs. The query language has a comprehension style syntax and contains a pattern specification sub-language DPL. DPL can be viewed as an extension of tree-pattern query language like XPath. The language allows extraction of nodes, paths and subgraphs from DAGs, and permits construction of result structures by composing them. We show that using such a language on OBO ontologies (such as the gene ontology), we can express more complex and scientifically valuable queries.

1 Introduction

Query languages and query evaluation techniques for the retrieval and manipulation of graph-structured data have been investigated since the late 80s [1,2], through the era of object-oriented data models [3,4,5] up to the more recent general interest in semistructured data [6,7,8] and ontologies represented in RDF [9]. Graph-structured data appear naturally in many modern applications, especially in biological information systems [10], chemical structure analysis [11], and social network analysis. In these application domains, a surprisingly large fragment of graph-structured data turn out to be directed and acyclic. Specifically in the domain of biomedical and biological ontologies, the majority of the ontological structures are designed to be directed acyclic graphs (DAGs). The Open Biomedical Ontologies (http://obo.sourceforge.net/) is an umbrella consortium that serves as a repository of many different but often inter-related ontologies, where the nodes of the graphs represent terms used in the vocabulary of a specialized biological domain, and the edges between nodes are typically labelled by the strings "isa", "part-of" or "develops-from". Furthermore, given the multiplicity and categories of ontologies emerging today, new needs are developing to query across ontologies and composing ontologies together. As the ontologies grow and become more complex, searching through them will require a more complex query mechanism that natively operates on graphs, especially DAGs.

* Supported in part by NSF ITR Grant EIA-0205061, and the NLADR grant from NSF.

U. Leser, F. Naumann, and B. Eckman (Eds.): DILS 2006, LNBI 4075, pp. 152–167, 2006.

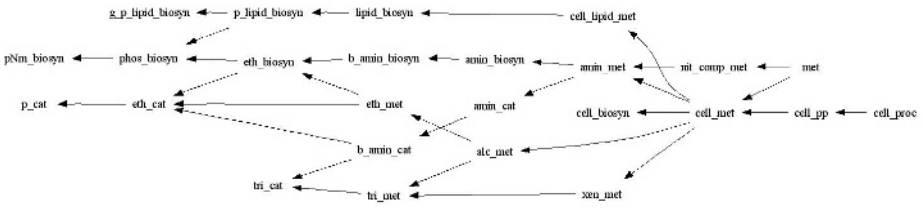

Fig. 1. A simplified fragment of the Biological Process component of Gene Ontology. The names of the nodes have been abbreviated for clarity.

Despite this need, most of the systems available to life scientists are mostly operated with visual interfaces allow only simple operations like keyword based node search, descendant enumeration, shortest path finding and neighborhood operations on graphs. This paper is an early step toward searching repositories of large ontological structures using a DAG query language, and similar in its intent as [12].

Example 1. As a motivational example, consider the well known Gene Ontology (GO) (`www.geneontology.org`) that consists of three DAG-structured components called biological processes (BP), molecular functions (MF) and subcellular components (SC). In Figure 1, a fragment of the BP DAG is shown. Here, an edge represents an *superclass* relation, such that $n_1 \rightarrow n_2$ means that the process n_2 is a specialization of the process n_1. Nodes in this graph represent tuples of a relation N which, in our simplified example, has three attributes *id*, *name* and *definition*. To make the node names simpler, just consider that a node with the substring "_met" is a metabolism process, a node with "_cat" is a catabolism process and a node with "_biosyn" is a biosynthesis process. Given this example DAG, a number of different types of queries can be asked:

1. Which biosynthesis processes under lipid biosynthesis are also classified as amine biosynthesis? (Q1)
2. How does phosphatidylethanolamine biosynthesis (phos_biosyn in Fig. 1) derive from cellular metabolism (cell_met)? (Q2)
3. Is there a case where a xenobiotic process (e.g., xen_met) is a subprocess of at least two forms of cellular metabolism? (Q3)
4. construct a reduced data graph by deleting all metabolism nodes except *met*, and connecting the non-deleted parent(s) of a deleted node n to its non-deleted children. (Q4)

Consider the first query. Since the graph represents a classification structure (i.e., an is-a graph) we interpret the expression "A classified as B" to mean "A reachable from B" in this DAG. Thus, this query can be expressed as the pattern query

$$\mathsf{reachable\_from}(X, \mathsf{lipid\_biosyn}) \wedge \mathsf{reachable\_from}(X, \mathsf{amin\_biosyn}) \wedge \mathsf{substr}(\text{'biosyn'}, X) \quad (\text{Q1}')$$