# BioNavigation: Using Ontologies to Express Meaningful Navigational Queries Over Biological Resources

Zoé Lacroix, Kaushal Parekh
Arizona State University
Tempe, AZ 85287, USA
{zoe.lacroix,kaushal}@asu.edu

Maria-Esther Vidal, Marelis Cardenas, Natalia Marquez
Universidad Simón Bolívar
Caracas, Venezuela
{mvidal,mcardenas,nmarquez}@ldc.usb.ve

Louiqa Raschid
University of Maryland
College Park, MD 20742, USA
louiqa@umiacs.umd.edu

## Abstract

*Exploiting the complex maze of publicly available Biological resources to implement scientific data collection pipelines poses a multitude of challenges to biologists in accurately reflecting the scientific question at hand and in the selection of the best resources which satisfy their needs. We extended our BioNavigation system to address these challenges and aid the scientists visualize and navigate the resources, express their queries and determine the most suitable set of resources to evaluate them. For this purpose, we use an ontology that describes the higher logical level of scientific concepts and their relationships. A user can browse and visualize this ontology and then graphically select the relevant nodes and edges to build his query. We developed the ESearch algorithm that searches the physical level of resources to generate paths that express the ontological query. The algorithm also ranks the paths based on three semantic metrics; target object cardinality - to optimize the number of records in the output dataset, path cardinality - to optimize the number of links between the involved data sources, and evaluation cost - to minimize the cost that will be incurred to execute that evaluation path. These metrics allow the user to select the most optimum path that matches his requirements.*

## 1. Introduction

There exist thousands publicly available biological resources, both databases and applications, often richly interconnected with links. Scientists exploit these resources with the help of navigational queries, which start with one source and end at another with many intermediate sources, forming a path or a pipeline. Scientists face many challenges in implementing such a scientific data collection pipeline. The first challenge is to accurately reflect the scientific question at hand in terms of a query that captures adequately the scientific aim. In contrast, scientists often build their queries to adapt to the characteristics and limitations of the resources that they are familiar with. Another challenge lies in the availability of multiple resources which may have similar purposes while being highly heterogeneous with respect to the data format, number of records, level of curation, navigational capabilities or links to other resources, etc. Thus the same higher level query can be translated to various evaluation paths involving a number of different data sources, links, applications etc. Experiments in [1] have shown that depending on the data source used for obtaining information about a given scientific class and the links followed, scientists may retrieve significantly different information, both in terms of quantity and quality, for the same scientific query. Hence, it becomes important for the user to understand what path is best suited to his purpose to get the best possible set of results from the query.

## 2. Use of Ontology

An *ontology* is a definition of the properties of important concepts and their relationships in an unambiguous language, which is both machine and human readable. As described in [4] we use two levels of representation for the scientific resources: the *Physical Level* consists of the data sources, applications and navigational links or capabilities whereas the *Logical Level* is a higher level representation of the scientific concepts and relationships that map to the

physical resources. We use an ontology to describe the higher logical level of scientific concepts and their relationships. Each scientific concept or class in the ontology is mapped to the physical data sources for that scientific concept whereas each relationship is mapped to the physical links between the databases or applications whose input and output corresponds to one of the scientific classes. This ontology allows the user to browse and visualize these classes and relationships and then graphically select the relevant ones to build his query. Figure 1 shows an example of an ontology.

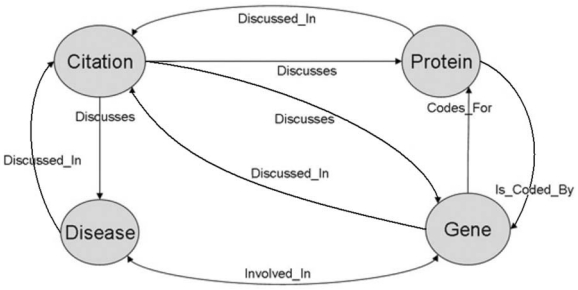


Figure 1: BioNavigation Ontology

## 3 Query Language and ESearch Algorithm

With the help of the ontology, a user can graphically build a query by selecting the relevant nodes and edges representing the concepts and relationships respectively. For example, a user interested in citations discussing a particular gene would first select, from the ontology illustrated in Figure 1, the scientific class *Gene*, then the relationship *Discussed_In*, and then the class *Citation*. This represents the ontological query for the scientific question. The Grammar specifying the query language is described in Figure 2. We use special symbols such as $\epsilon$ to specify an arbitrary number of intermediate resources to be used in the path. It allows the user to search for any-length paths. Also, the annotations allow him to specify any restrictions on the physical resources to be used or avoided in the resulting paths.

```
<RE>:=<cTerm><Y>
<cTerm>:=<EpsilonC> | <ClassName><SourceAnnotation>
<Y>:=<Epsilon><Y> | <aTerm><cTerm><Y> | empty
<aTerm>:=<EpsilonA> | <AssociationName><LinkAnnotation>
<SourceAnnotation>:=empty | "[" <SourceList>"]"
<SourceList>:=<AnnotatedSource> | <AnnotatedSource> ","
                                              <SourceList>
<AnnotatedSource>:=<OP><SourceName>
<LinkAnnotation>:=empty | "[" <LinksList>"]"
<LinkList>:=<AnnotatedLink> | <AnnotatedLink> ","
                                                <LinkList>
<AnnotatedLink>:=<OP><LinkName>
<LinkName>:=<ApplicationName> | <QueryCapName>
<OP>:="!=" | "="
```

Figure 2: BNF grammar of regular expressions

We extended the ESearch Algorithm introduced in [4] to accept the ontological queries described above. ESearch is based on an annotated DFA (Deterministic Finite State Automaton) that performs an exhaustive breadth first search on the physical level graph of resources to generate paths that express the query. It recognizes annotations in the query and includes or excludes corresponding physical implementations. In addition to generating all possible evaluation paths, the algorithm also ranks the paths with respect to three metrics. The ranking guides the user to select the path that best satisfies his needs. These metrics defined in [3] and [4] are described below:

1. *Path Cardinality* is the number of instances of paths of the result. For a path of length 1 between two sources S1 and S2, it is the number of linked pairs (e1, e2), where e1 is an entry in S1 and e2 in S2.
2. *Target Object Cardinality* is the number of distinct objects retrieved from the final data source.
3. *Evaluation Cost* is the cost of the evaluation plan, which involves both the local processing cost and remote network access delays.

## 4 Conclusion

BioNavigation acts as a helpful guidance system for scientists in their data collection efforts. BioNavigation can be combined with a system to execute the query on the path selected by the user. In the future, BioNavigation will be coupled with SemanticBio [2] that allows users to express and execute scientific workflows with an ontology and Web Services.

**Acknowledgments** The work was partially supported by NSF grant IIS-0223042.

## References

[1] Z. Lacroix and V. Edupuganti. How biological source capabilities may affect the data collection process. In *CSB*, pages 596–597. IEEE Computer Society, 2004.

[2] Z. Lacroix and H. Ménager. SemanticBio: Building conceptual scientific workflows over web services. In *DILS*, Lecture Notes in Bioinformatics. Springer, July 2005.

[3] Z. Lacroix, H. Murthy, F. Naumann, and L. Raschid. Links and paths through life science data sources. In Rahm [5], pages 203–211.

[4] Z. Lacroix, L. Raschid, and M.-E. Vidal. Efficient techniques to explore and rank paths in life science data sources. In Rahm [5], pages 187–202.

[5] E. Rahm, editor. *DILS*, volume 2994 of *Lecture Notes in Bioinformatics*. Springer, 2004.