

Proposal for Collaborative Building Environment of Animal Trait Ontology

October 5, 2005

1 Description of research support position requested

Funding for 1 graduate research assistant is requested. Anticipated total salary and fringe benefits for each year amount to \$21,676.

2 Plan for utilization of position

The graduate research assistant will report to Vasant Honavar and James Reecy who are collaborating on the development of informatics tools to support annotation, storage, retrieval, and querying, and analysis of disparate data sources in support of animal genomics research.

3 Research Statement

3.1 Introduction

The need for biological ontologies has risen in recent years in large part due to the development of large biological databases. Successful ontologies in biology, such as Gene Ontology(GO)¹, Rice Ontology², Plant Ontology³ and Phenotype and Trait Ontology (PATO)⁴, have emerged in the past few years. Ontologies that identify and define the entities and relationships in specific domains of interest offer a powerful approach for annotating biological data in a form

that allows users and software tools to retrieve, interrelate, and extract biological knowledge from such data.

With the rapid proliferation of biological ontologies, the complexity of their size, structure and thus the building process is also increasing.

A typical biological ontology used today contains ten of thousands of terms. For example, GO contains 2×10^5 terms and Gramineae Taxonomy Ontology contains 7×10^5 terms. Large-scale ontology requires better mechanism to create and process on

- Collaborativeness: The ontology building is usually a collaborative process. In order for a ontology to be broadly useful to a certain community, it needs to capture the knowledge based on the collective expertise of multiple experts and research groups. Typically, a large ontology is built and curated by a community. Each member of the community only contributes a part of the ontology and the final ontology is compiled from the contributed pieces.
- Scalability: Stronger tools are needed to handle large-scale ontology for storage, editing, browsing, visualizing, reasoning and reusing. Those tools should enable processing of the ontology with limited time and space(e.g. memory) resources.

Such characteristics call for mechanism and tools to support both scaleable and collaborative ontology building.

¹<http://www.geneontology.org/>

²<http://www.ro.dna.affrc.go.jp/docs/index.html>

³<http://www.plantontology.org/>

⁴<http://obo.sourceforge.net/cgi-bin/detail.cgi?poav>

3.2 Animal Trait Ontology

With the rapid development of animal genome sciences, there is also an urgent need for precise definition of animal trait terms (phenotypes) to capture the biologically relevant distinctions at the desired level of detail in an unambiguous fashion.

Previous works by Hu *et al* [4] on PigQTLdb has introduced simple ontologies in the form of controlled vocabularies to describe pig phenotypes/traits, and to link them to QTL information. With the dramatic increase in genomic information available to researchers it is imperative that projects like this be completed. Otherwise, querying databases in the future will be extremely problematic due to the unstructured nature of the nomenclature. For example, back fat, rib fat, fat thickness could all mean the same things. Thus, we need to develop a way to link these terms in a queriable fashion. That leads to the development of software tools to facilitate the creation, editing, curation, and management of animal trait ontologies.

Animal Trait Ontology (ATO) describes animal trait on several aspects, such as meat quality, health factors, exterior features, production and reproduction parameters. A well established ATO will help animal science as well as the livestock production community to use standardized terminology to effectively record, describe, and compare traits.

3.3 The Collaborative Editor for ATO

3.3.1 Limitation of Existing Tools

Ontology building has received considerable attention during the past few years, and there are many available ontology editors, such as Protege⁵, OilEd⁶, OntoEdit⁷ and DAG-Edit⁸. For the state-of-art of ontology editor, please see survey by OntoWeb (2002)[5], Michael Denny (2002) and [2] (2004)[3], and Irene Polikoff (2003) [6].

However, due to the characteristics of ATO (and other large-scale biological ontologies), such existing

tools have some limitations such as

- **Efficiency:** ATO ontology is a hierarchy-based ontology with DAG skeleton, which demands a succinct ontology representation. However, most of the general-purpose ontology editors are based on expressive knowledge representation paradigms, such as Description Logic (or its variances of DAML, OIL and OWL), Frame Logic (such as OKBC) or First Order Logic (FOL). Although such expressive languages are capable to build hierarchy, tools based on such languages are less efficient than a devoted hierarchy tool. This is also the reason why Gene Ontology has its own tool DAG-Edit.
- **Collaboration:** Many tools, such as Protege, OilEd and DAG-Edit, lack the support for collaborative ontology building. Such tools support single user at certain moment and hardly support the development for team work.
- **Scalability:** Many file-based systems face scalability problem when the ontology is large. Although some systems support database storage, they need a whole in-memory model for user interaction and reasoning, such as Protege-OWL.

No tool, to the best of our knowledge, has met all the requirements for ATO application. We have developed an animal trait ontology editor (ATO editor) that overcomes most of these problems.

The building of ATO is boosted by allowing team collaboration of modular ontology.

3.3.2 Modular Ontology

Modularity in ontologies is beneficial to ontology engineering in several ways. It simplifies the maintenance and evolution of ontology and the mappings between ontologies; it enables flexible partial reuse of large ontology; it offers a more efficient organization for ontology; it also improves the processing and query of ontology.

In this research, we introduce **package**[1] as the basic module of ontology. In such a modular ontology representation, ontology is divided into smaller

⁵ <http://protege.stanford.edu/>

⁶ <http://oiled.man.ac.uk/>

⁷ <http://www.ontoknowledge.org/tools/ontoedit.shtml>

⁸ <http://www.godatabase.org/dev/java/dagedit/docs/index.html>

components called packages, and each package contains a set of highly related terms and their relations; packages can be nested in another package, and form a package hierarchy; the visibility of a term is controlled by scope limitation modifier such as **public**, **private** and **protected**. The whole ontology is composed of a set of packages.

Packages help to handle large-scale ontology on several aspect:

- Package serves as organization unit. User can check out one or more packages for editing, while other packages are still open to other users. Such dynamic and partial loading of ontology also demands much less memory than full in-memory loading.
- Package can be partially reused. For example, if pig component and horse component of ATO shares information on “production”, the shared part can be organized in a package and be reused by the pig component and the horse component.
- During the editing, ontology may evolve. To reduce the risk of unwanted coupling and inconsistency diffusion, package can be used with “visible interface” (a set of terms that can be used by other packages) and “invisible” inner contents. The inner contents can be changed (for example, insert new node between exiting node, or merge two inner nodes) while the visible interface is still unchanged.
- In reasoning, package has localized semantics, thus an error in a package can be more easily controlled and traced.

3.3.3 Team Collaboration

We provide a modular ontology editor to enable the building and deployment of ATO (see Figure 1). The editor enables the ontology developer to create a community-shared ontology server with database storage which supports concurrent browsing and editing of the ontology. The editor supports team collaboration with such features:

- The ontology is stored on a relational database server; a user can connect to the server and check out part of the ontology, do editing on the checked-out part, and check it in when finished editing.
- The editor supports multiple users to concurrently edit the ontology. Locking mechanism is provided to avoid conflict and abnormalities. Modules of the ontology can be developed by different parts and be assembled in latter stages.
- There is user profile management. Authors for the ontology have different level of privileges (such as ontology admin and package admin) over modules in that ontology. Author of a package can authorize other user the access of certain terms therefore controls the extendibility of that package.
- The editor provides a handy graphical user interface (GUI) to edit and browse a DAG.
- The editor can import and export ontology in standard format, such as OBO and OWL.

The tool is also ideal for other ontologies in OBO (Open Biomedical Ontologies) format, for example, Gene Ontology.

3.4 Proposed Work

Based on the existing work, we will improve the ATO ontology and its building tools on following problems.

1. Expand the animal trait ontology for multiple species. We have developed database structures to manage trait ontology, and entered over 300 pig traits into the pig trait ontology database. We have also created ontology database structure for cattle, chickens, sheep and other livestock species. Actual development of each trait ontology requires that the scientists from relevant fields work together. We plan to continue our research in ontology management and maintenance to make this process a smooth one.

2. Promote the developing and use community of ATO. Researchers from the animal genomics community are among the first to realize the utility of ontology in effective communications involving trait

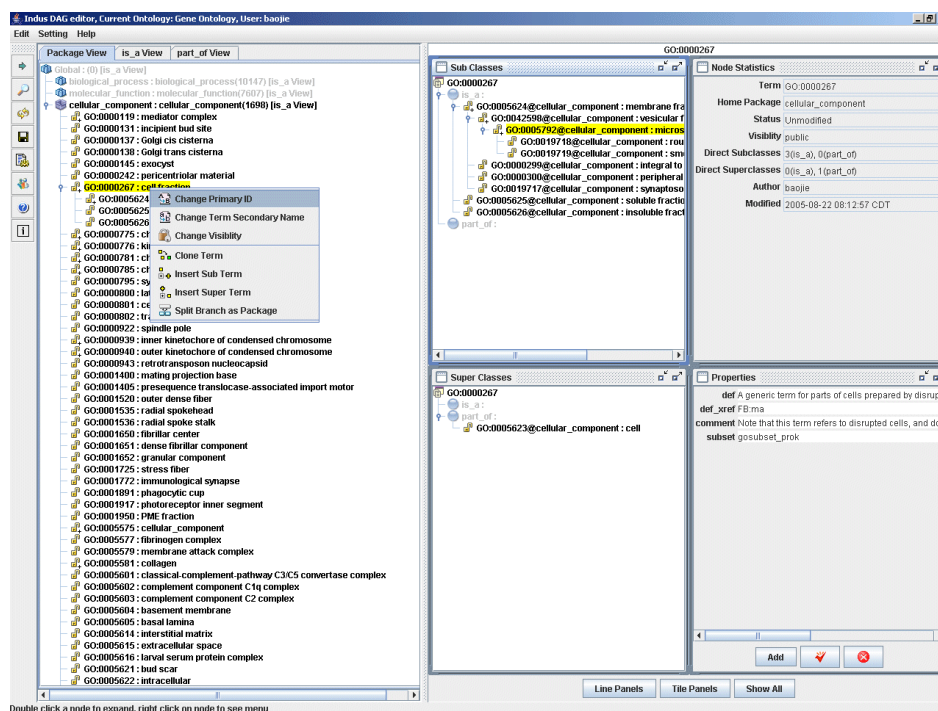


Figure 1: ATO Editor, loaded with Gene Ontology as example

comparisons. The use of ontology as a standard by all scientist in the animal science field will facilitate their research and routine as well. We will submit ATO as one of the core OBO ontologies.

3. ATO is semantically related to some other ontologies, such as PATO. We will create tool to build mappings between ATO and other ontologies. The mapping building tool will also support scaleable storage and collaborative building.

4. The editor will ensure the consistency of submitted ontology, the integrity of entire ontology and the correctness of ontology mappings. Therefore, a reasoner should be connected to the editor. The reasoner will be optimized for DAG ontologies and supports database based storage.

5. We will apply ATO to annotate practical data sources.

ATO and associated software tools for collaborative creation, editing, curation, and management of ontologies will provide the infrastructure necessary for engaging the animal genomics and livestock com-

munities in the process of creating comprehensive genomic resources for annotating, integrating, and analyzing phenotype and genomic data, and for performing genome-wide cross species analysis.

4 Current and Future Projects to be Supported by this Position and Funding Sources

- Discovering Protein Sequence-Structure-Function Relationships. National Institutes of Health. Vasant Honavar (PI), Robert Jernigan and Drena Dobbs (Co-PIs), (2003-2007). \$1,022,000.
- Integration of Functional Genomics and Quantitative Genetics to Improve Feed Efficiency in Pigs. United States Department of Agriculture (2005-2008), Vasant Honavar (Co-PI) with Jack Dekkers (PI), Chris Tuggle (Co-PI), Dan Net-

Table 1: CIAG Research Support Grant Fall 2005 Competition Budget (2 years)

		Year 1	Year 2	Project Total
1- Graduate Research Assistant	\$1620*12 months	\$19,440	\$19,440	\$38,880
Benefits	11.5%	\$2,236	\$2,236	\$4,472
Total		\$21,676	\$ 21,676	\$43,352

tleton (Co-PI), Lloyd Anderson (Co-PI), Rondhane Rekaya (Co-PI), Richard Barb (Co-PI), and Gary Hausman (Co-PI), \$876,000.

- Developmental Proteomics of Retinal Progenitor Cells, National Institutes of Health (2003-2007). Vasant Honavar (Co-PI), (with Heather West-Greenlee and Jan Buss), \$438,000.
- National Research Sponsored Program - 8 Livestock Database Coordination. United States Department of Agriculture. Cooperative Research Extension Education Service (2003-2008) James Reecy (PI), Susan Lamont (Co-PI), Max Rothschild (Co-PI), and Chris Tuggle (Co-PI) \$540,000.

5 Biographical Sketch

CVs of Vasant Honavar and James Reecy are in the attachment

6 Budget

Please see Table 1

Funding Requested from CIAG: \$10,838

Matching Funds from CCILD: \$5,419(497-17-25)

Matching Funds from Reecy: \$5,419(415-40-05)

Justification: Funds are requested to support Jie Bao, a Ph.D. student working under Dr. Honavar's supervision. Jie will be responsible for design, implementation, documentation, and testing of the proposed ontology tools, under the guidance of Vasant Honavar and James Reecy. Jie Bao has established good working collaborations with Zhiliang Hu, who works for Jim Reecy on the USDA

Bioinformatics project. This collaboration has led to the prototype software which will be further developed as part of the proposed research.

References

- [1] Jie Bao and Vasant Honavar. Ontology language extensions to support localized semantics, modular reasoning, collaborative ontology design and reuse. In *3rd International Semantic Web Conference (ISWC2004), Poster Track, 7-11 November 2004, Hiroshima, Japan*, 2004.
- [2] Michael Denny. Ontology building: A survey of editing tools. XML.com, <http://www.xml.com/pub/a/2002/11/06/ontologies.html>, 2002.
- [3] Michael Denny. Ontology tools survey, revisited. XML.com, <http://www.xml.com/pub/a/2004/07/14/onto.html>, 2004.
- [4] Zhi-Liang Hu, Svetlana Dracheva, Wonhee Jang, Donna Maglott, John Bastiaansen, Max F. Rothschild, and James M. Reecy. A qtl resource and comparison tool for pigs: Pigqtl db. *Mammalian Genome*, 15:792–800, 2005.
- [5] Asuncion Gomez Perez and et.al. Deliverable 1.3: A survey on ontology tools. OntoWeb, <http://citeseer.ist.psu.edu/525623.html>, 2002.
- [6] Irene Polikoff. Ontology tool support: Ontology development lifecycle and tools. TopQuadrant, <http://www.topquadrant.com/documents/TQ1202.Ontology%20Tool%20Survey.pdf>, 2003.