ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 6, Issue 4, pp: (25-32), Month: October - December 2018, Available at: www.researchpublish.com

A Case Study of the Gene Database Ontology

Kahtani, Abdullah M

Saudi Aramco, Information Technology

Abstract: **As the data regarding biological genetics has grown exponentially, the ability for scientists to access this data has diminished. This made it more essential to have a consistent representation in a specific domain for Gene Ontology (GO) where it will be easier to retrieve information about genes of interest. Promising advances in data semantics and artificial logic have led to the development of ontologies that have the ability to tie together vast amounts of information from disparate locations and formats into a single, useable interface. One such project, the Gene Ontology, is discussed below along with an evaluation if the GO's effectiveness and weaknesses. In this research, we first review the motivation behind the development of the GO and some of its major design principles. Then we describe the possibility of extending this ontology for use in other fields. After that, we will do an evaluation based on some evaluation metrics. Finally, we highlight some of the practical uses and applications of GO.**

Keywords**: Database, Gene, Ontology, GO, Semantics, Biology.**

1. INTRODUCTION

The amount of data that has been accumulated for various areas of science has become staggering over the last fifty to sixty years. Nowhere is this more apparent than in the field of biology. The problem is exacerbated by the fact that such data is held in a variety of database with different formats and naming conventions. One way to deal with the immensity of data in diverse and complex formats is to develop an ontology based on data semantics that provides a framework for referencing the data. This paper focuses on the Genetic Ontology (GO), an attempt to develop an ontology of biological information.

The concept of ontology actually began in the philosophies, where the idea connoted a logical way of categorizing and analyzing knowledge. As early philosophers attempted to devise a system of understanding everything, they ran into two dichotomies: continuous vs. occurrent entities and dependent vs. independent entities (Kumar and Smith 2003).

Continuous entities are entities that maintain their identity at all times, such as a particular organism, while occurrent entities are temporal in nature, such as the process of photosynthesis. Independent entities can be clearly defined without referring to another entity, while dependent entities cannot. Thus a living cell is an independent entity while the mass of the cell is dependent because mass is always a property of something else.

The two dichotomies can be combined to create four classes of entities: occurrent dependent, occurrent independent, continuous dependent and continuous independent. Since occurrent entities are generally dependent on some other entities, the occurrent independent class if often left out, leaving a tripartite oncology that includes continuous dependent, continuous independent, and occurrent entities as shown in Figure 1 below. As we shall see, the proper demarcation of these classes becomes critical if an ontology is to be effective.

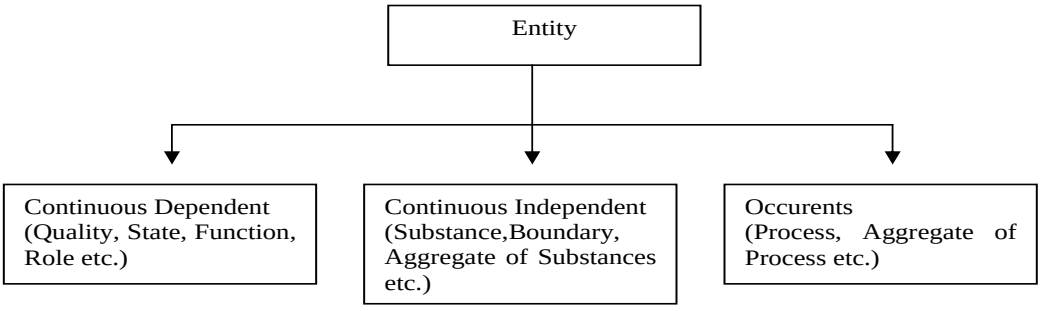

Figure 1: Triparate Oncology (Kumar and Smith 2003)

Formal philosophical ontologies were created to allow the rules of logic to be applied when reasoning about various phenomena. This concept applies quite nicely to the problem of organizing vast amounts of data in a way that can be analyzed via computer logic. Even though the data may exist in disparate locations and formats, it is possible to create a formal set of rules that will categorize the data into predefined classes. Such an ontology defines the entities as well as the relationships between them. An information ontology commonly consists of three components:

- Classes (entities)
- Attributes of classes
- Relationships between classes

The classes are the entities. Attributes are things that we know about the class whereas Relationships define how classes interact. By following a set of formal rules, navigation of the ontology can be automated based on tested algorithms. Thus one purpose for a data ontology is to allow the automated retrieval and analysis of data contained in the ontology (Hill, et al. 2008).

As the ontology is constructed and populated, the names of the classes, attributes, and relationships are chosen so that they define the acceptable terms that can be used when describing a particular domain. This set of terms becomes a controlled vocabulary that defines the acceptable language that will be used when defining the domain. Existing and well understood algorithms from the realm of graph analysis and data semantics can be applied to controlled vocabularies. In such a scenario, each class becomes a node in the graph and the relationships become the edges. One such application is the Open Biological and Biological Ontology library (OBO), which is a repository of controlled vocabularies dealing with the biological and medical domains (Smith, Ceusters, et al. 2005).

The problem of defining an ontology for an existing body of knowledge is that the data is already being held and defined in a variety of disparate and incompatible formats. The goal of the ontology is to standardize the vocabulary for the domain so that information from otherwise incompatible databases can be queried and analyzed via the ontologies schema. Thus, one practical purpose of an information ontology is to map each term in the controlled vocabulary to other terms, which may be in use and have the same meaning. In this sense, the ontology can be thought of as a meta-database that can be used to cross-index other databases by mapping the controlled vocabulary into the diverse terms that are being used in various databases. Thus, the controlled vocabulary becomes a data dictionary for the meta-database.

2. MOTIVATION

As noted above, the main motivation behind the GO is to create a way to manage the existing complexity of information related to the fields of biology and genetics. One example of the wealth of information is the PubMed literature database which contains over 15 million citations. Obviously, this single database is beyond the ability of anyone to navigate without computational help (Hill, et al. 2008).

The main motivation behind developing GO is to improve semantic coherence of various resources by software tools, which are specifically used for searching genes and proteins data by Geneticists and Biologists. It's a major initiative in bioinformatics to unify the representation of genes and gene attributes among all organisms. Moreover, the project main goals include but not limited to: (1) developing and maintaining its vocabularies of various genes and gene product attributes. (2) Annotating genes and genes products, which might include assimilating and disseminating the annotated data (3) Continuing to provide many valuable tools for accessing, testing, and manipulating the data provided by the GO project. (Wikipedia, 2011).

The principle use of the GO is to support curators of model organism databases and genome annotation centers. Such users need access to the vast amount of research regarding gene products and biological systems (Hill, et al. 2008). A structured system such as the GO allows computers to access the information across disparate databases in a uniform way.

It isn't enough for computers to be able to navigate the ontology. Another key goal of the GO is to code information in a way that is useful to humans. In other words, it is not good enough for the GO to represent an abstract, mathematical view of the biological universe. Thus, the GO uses a structured system of annotations, which allows users to add additional information to entities that otherwise would not be captured in a purely ontological system (Hill, et al. 2008).

ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 6, Issue 4, pp: (25-32), Month: October - December 2018, Available at: www.researchpublish.com

3. DEVELOPMENT

Several projects have been developed that support the work of the GO and upon which the designers of the GO looked for inspiration. For example, a system called PANTHER specializes in classifying proteins (Kumar, Smith and Borgelt 2004). Similarly, The Foundational Model of Anatomy (FMA) focuses on classifying the parts of the human body (Smith and Rosse 2004). The Gene Ontology project has been developed alongside these systems as an attempt to classify all biological and genetic information. In fact, one long-term goal of some biologists is to align these various biological systems so they are more compatible.

The development of any ontology must begin with a controlled vocabulary of terms that define the key concepts. The field of canonical anatomy focuses on defining such a vocabulary for biological concepts, while the field of instantiated anatomy focuses on data relating to actual instances of biology. One role of an ontology is to relate the empirical data to its corresponding canonical terms. This task was first realized in the FMA (Smith and Rosse 2004). This process ties the practical to the theoretical.

Another example of a controlled vocabulary is the Unified Medical Language System (UMLS) that serves as a semantic data dictionary for the field of biomedicine. The UMLS Semantic Network defines over six-thousand relationships between these terms and divides them into two branches: entities and events (Smith and Rosse 2004). The classification of terms into entities and events closely corresponds to the philosophical concepts of continuants and occurrents discussed above.

The Gene Ontology (GO) was considered one of the most crucial early steps towards solving one of the big problems in the biology field, which is representing all biological terms (genes) unambiguously in a computationally traceable way. The GO project defines a controlled vocabulary for the description of genes and gene products. It encompasses many databases and attempts to create a standard interface for retrieving and annotating information. Terms are organized into parent-child hierarchies such that the parent is always more general that its child (Kumar and Smith 2003).

Terms in the GO are divided into three domains as shown in Figure 2 below:

1. Molecular Function (Activity) Ontology – activities at a molecular level.
2. Cellular Component Ontology – inside the cell including the cell.
3. Biological Processes Ontology – changes that lead to a particular result at the cell and whole organism level.

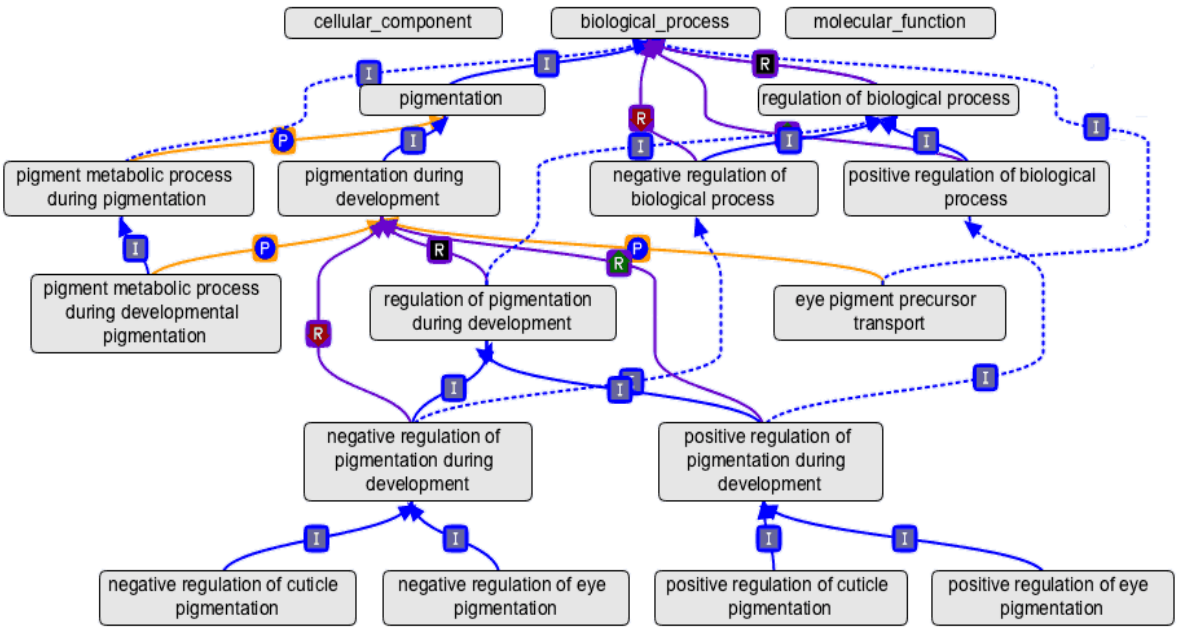

Figure 2: Gene Oncology Structure (The Gene Ontology 2011)

These three domains are actually three distinct ontologies/domains because there are no relationships defined between them. For example, terms in the Cellular Component domain are not cross-referenced with terms in the Molecular Function domain even if they are related in real life. This design has both positive and negative consequences, which will be detailed below.

Ontologies use very specific relationships to describe how entities are related to other entities. Although there are a wide variety of formal relationships to choose from, the designers of the GO decided to use only two:

1. Is a – entity B is an example of entity A
2. Part of – entity B is part of entity B

The decision to use only two relationships greatly simplifies coding in the GO because the user only has to choose from two possible relationships. As discussed below, this has also been the source of many problems in the GO.

The designers of the GO also wanted to provide a way for users to add information that went beyond defining entities and their relationships. Therefore, the designers included a system for adding annotations to each entity. Annotations allow users to add data from research and other sources to enhance the GO with current, real-world data.

The process of adding annotations is known as curation. Annotations are established electronically and later validated by a process of manual curation, which requires the annotator to have expertise both in the biology of the genes the structure of GO. (Smith, Köhler and Kumar 2004) Each annotation must be supported by an evidence code, which is a three-letter designation that indicates that type of experimental evidence that supports the annotation (Hill, et al. 2008). Thus, the GO becomes not only a repository of terms and relationships, but of all knowledge relating to genetic entities and processes.

4. EVALUATION

4.1 Metrics used in Evaluation

The framework for any ontology must follow very specific rules of logic to be effective. It is adherence to such rules that allow the ontology to be searched using automated algorithms. The Gene Ontology was not developed primarily with computers in mind, but instead, was primarily designed for human understanding. For this reason, there have been quite a few articles written that point out flaws in the design of the GO specifically because it does not strictly follow the formal rules. Several of the criteria used to evaluate an ontology are described below:

- Formal integrity – The ability to automatically retrieve data because of strictly defined rules and relationships.
- Positivity: complements of classes are not themselves classes.
- Objectivity: the classes are non-fictitious, and not arbitrary.
- Levels: the terms in a classificatory hierarchy should be divided into predetermined levels.
- Single inheritance: no class in a classificatory hierarchy should have more than one parent on the immediate higher level.
- Exhaustiveness: the classes on any given level should exhaust the domain of the classificatory hierarchy.
- Intelligibility: the terms used in a definition should be simpler than the term to be defined.
- Substitutivity: a defined term must be substitutable by its definition in such a way that the result is both grammatically correct and has the same truth-value as the original sentence.

It is understood that at some point the definitions will contain terms that are so elementary that they cannot, of themselves, be defined. These terms, known as level 0 terms, are considered primitives that cannot be further divided (Smith and Kumar 2004).

4.2 The Richness of GO

By many measures, the Genetic Ontology has been very successful. By the end of 2003 the three domains contained nearly 17,000 entries. (Kumar, Smith and Borgelt 2004). By now that number would have grown greatly.

As noted above, the GO is not primarily designed to support automated reasoning, but rather, it is designed for biologists. In this regard, it has had great success. The GO has enabled and facilitated the analysis of very large datasets. A great deal of this utility is due to the annotation system, which has been praised for enhancing the ontology with practical scientific

findings (Hill, et al. 2008). The GO has also been successful in defining a controlled vocabulary that biologists now use for describing and coding genetic information. The GO also provides a framework for defining new terms and concepts, which then become part of the structured vocabulary (Kumar, Smith and Borgelt 2004).

Another design aspect that is generally applauded is the decision to use only two relationships. This greatly simplifies the choices that must be made when coding relationships (although this is also a source of several problems discussed below) because the biologist does not have to be formally trained in the use of multiple logical relationships. This design choice is also cited for the short timeframe within which the GO was populated with its initial data. (Kumar, Smith and Borgelt 2004).

Finally, one of the biggest strengths of the GO design is that it is continually updated by biologist-curators who are experts in understanding the field. The validity of the entities and relationships in the ontology is continually checked against the real-world. GO curators refer to this as annotation-driven ontology development. In addition, the GO community works with scientific experts for specific biological systems to evaluate and update the entire framework (Hill, et al. 2008).

4.3 Issues

Technically, the Gene Ontology is not an ontology as the term is used by information scientists and by philosophers (Smith, Köhler and Kumar 2004). As discussed above, the GO uses annotations to extend the information above and beyond what would normally be found in an ontology. Most of the criticisms of the GO are directly related to its failure to strictly hold to the formal rules of data semantics.

The first criticism of GO stems from its focus on human usability. Because of this focus, the GO follows neither the rigidity required by computer applications nor the logic required by philosophical standards (Smith, Williams and Schulze-Kremer 2003). As a result, there is quite a bit of concern about the ability to use automated algorithms to access the data, especially as the amount of data grows. Some specific ontological rules that are breached in the GO include:

1. The failure to differentiate between universals and instances.
2. The failure to differentiate between continuants and occurrents.
3. The failure to differentiate between dependent and independent entities.
4. The failure to adhere of classes to be exhaustive and disjoint.
5. The failure to differentiate definitions between parents and children.
6. Inaccuracies in treatment of the is-part relationship.
7. The failure to guarantee consistent levels of granularity.
8. The failure to respect standard principles in the formulation of definitions.
9. Loose principles for synonymy.
10. No concern for consistent treatment of the compositional structure of terms.
11. The failure to deal adequately with time.

(Kumar and Smith 2004)

The next major shortcoming originates from one of GO's greatest advantages: the simplicity of using only two relationships. Although this simplicity makes it easier for biologists to code data, it also limits the expression of different types of relationships. The standard GO only allows *is-a* and *part-of* relationships. Unfortunately, life does not fit into only two dimensions. As a result, biologists stretch their interpretation of these relationships to accommodate to describe relationships that do not exist.

To compensate for the lack of relationships, the GO allows curators to create and use special terms that are appended to the existing relationships. These are known as special operators and include terms such as *with*, *within*, *without*, *in*, *site-of*, *acting-on*, or *resulting-in*. Unfortunately, these terms are never formally defined, and it is strictly at the whim of the curator to pick and choose the correct terms to use. This greatly reduces the ability to construct automated search algorithms that can find meaningful data (Smith and Kumar 2004).

Even the two standard relationships in GO are used in incorrect ways. For example, the A *part-of* B relationship can mean three different things: A is sometimes a *part-of* B, A can be *part-of* B, or A is included in B. Each of these relationships has different connotations. Similar problems occur with the *is-a* relationship, which is used both to mean A *is-a* B or A *is-a [part of]* B (Smith, Ceusters, et al. 2005).

Another problematic term is the use of sensu, which is used as a kind of wildcard between terms that have one meaning in some situations and another meaning in other situations. The term is also used to introduce new terms that are not clearly defined (Cell walls in bacteria and in fungi have a completely different composition.) This introduces a form of ambiguity that could never be navigated by an automated algorithm (Smith and Kumar 2004).

GO also uses syntactic operators, such as ',', '/', and ':', in ways which seem to contradict the underlying idea of a controlled vocabulary and have multiple meanings. For example, ',' might mean 'while' or 'of the type which is'. The '/' might mean 'and', 'or', or 'and/or' (Smith and Kumar 2004).

Another weakness of the GO is the failure to relate the three ontologies. This means that obvious relationships between biological and molecular processes are completely ignored by the GO. This is partly compensated for by common annotations. For example, given two terms from different ontologies, if they are related they will be given similar annotations, which can be used to form an implicit relationship. This results in duplicate annotations being created and maintained between the various ontologies (Kumar, Smith and Borgelt 2004).

The end-result of the failures discussed above is that the GO requires human intervention to retrieve, code, and interpret the data. In a way, it is the expertise of these biologists that effectively hides the inadequacies of the GO, because they are able to navigate and interpret the GO. Unfortunately, as the GO becomes larger and larger, this type of human intervention will become impractical (Smith, Köhler and Kumar 2004).

5. EXTENDING GO

The original GO has been extended to include more relationships in an attempt to provide more flexibility and accuracy. The use of these additional terms should help alleviate the need to force the existing two relationships to fit everything. The two new relationships in the extended GO include:

- The *has-part* relation
- The *regulates* relation

(The Gene Ontology 2011)

Another project is focused on creating meaningful relationships between the three separate domains of the GO. Instead of modifying the GO to incorporate relationships between the three ontologies, this project uses statistical analysis of the annotations to find such relationships. Thus, this project can be thought of as a database of meta-relationships. (Kumar and Smith 2004).

A project known as the OBO Foundry has been developed as a tool for normalizing the GO. The project consists of logical definitions that are cross-referenced with classes in the GO. While the GO entries are typically human readable and unfriendly to automated searches, the terms in the OBO Foundry are designed with to meet the rigorous standards of automated reasoning. Furthermore, since the terms in the OBO Foundry also cross-reference other ontologies, the OBO Foundry provides a tool for linking the GO with other ontologies. Furthermore, the newest GO standards require logical definitions to be assigned to new classes. As a result, new entries into the GO will automatically be included in the OBO Foundry (Mungall, et al. 2011).

6. APPLICATIONS BASED ON GO

Several projects have been started to correlate the GO with other biological databases. One such project will incorporate GO into UMLS (Kumar and Smith 2003). Another project is underway to create associations between the GO and The Institute for Genome Research (TIGR) databases (Kumar, Smith and Borgelt 2004). Both projects will greatly enhance the ability for biologists to reference an even greater universe of biological and genetic information.

One project that is based directly on the GO is the Reference Genome Project. The project will start by creating annotations for twelve reference genomes. These initial references will then be used as a model for annotation all fully sequenced genomes. The goal is to provide an automated process for researches to retrieve information on genomes (The Gene Ontology 2011).

Another project is using the GO to provide detailed information for cardiovascular research. This project, known as the GO Annotation for the Cardiovascular System Initiative, attempts to cross-reference all genes in the GO related to cardiovascular processes and diseases. This will greatly facilitate research by tying together new information in gene sequencing with existing knowledge of heart disease and related issues (The Gene Ontology 2011).

A final project is the GO Annotation for the Renal System. Like the GO Annotation for the Cardiovascular System, this project attempts to create a reference for all genes and processes related to the renal system. Again, this will provide researches with a tool for accessing the vast amount of research and knowledge relating to the renal system and tie this to the most recent advances in gene sequencing (The Gene Ontology 2011).

7. FUTURE WORK

Most of the work being considered for the future of the GO deals with improving the limitations discussed above. There is an increasing urgency to correct such problems before the ontology grows very large and unorganized as to make the fixes impossible or impractical. The current problem is that the GO is already so complex that such improvements will require automated tools to make the fixes. The current inconsistencies make it difficult to make such tools (Smith and Kumar 2004).

The Open Biological Ontologies (OBO) is an umbrella organization that is currently embarking on a program of reform for the GO. This will include a review of the formal ontological principles that must be enforced if the GO is to be accessible to automated search algorithms. (Smith and Kumar 2004). The following key goals have been set to improve the GO over time:

- Use of tools to avoid coding errors.
- Ensure that computer systems will be able to assume more of the burden of ontology curation.
- Ensure that such systems are better able to use GO as a basis for automated reasoning.
- Facilitate GO's interoperability with other biological databases and ontologies.

(Smith, Köhler and Kumar 2004)

A great deal can be done to ensure and enforce more consistent coding in the GO by using editing tools to enforce consistency and rules. One of the more popular tools is DAG-Edit. Unfortunately, this tool does not enforce the use of standard relationships because the curators can delete or modify the relationships at will. DAG-Edit, and other such tools, should be shipped with a fixed set of well-defined relation types that cannot be removed or modified by the user (Kumar and Smith 2004).

The same tool that was used above to create relationships between the three ontologies provides one step toward creating an algorithm that can locate internal inconsistencies. The authors have already tested this premise and were able to identify terms and definitions, which were defined in a circular or unintelligible way. Another test was able to identify a subset of 6001 problematic GO terms (Smith, Ceusters, et al. 2005). Such tools will go a long way toward verifying the integrity of the GO, while identifying key areas that are problematic (Köhler, et al. 2006).

8. CONCLUSION

The Genetic Ontology has already proven itself to be an excellent tool for biologists and geneticists. It has already begun to tame the immense amount of data that has been accumulated in this field, and also serves as a central repository for current and future experimental data. Practical applications have made the vast information in the GO available to researchers and doctors in the areas of cardiovascular and renal health. Other projects are in place that link the GO and related ontologies to cross-reference a universe of knowledge.

While the GO has met its primary goal, serving as a tool for experts in the field, certain design decisions tend to limit the ability to accurately mine data using automated algorithms. To continue to be useful into the future, the GO must be brought more in line with the formal principles of data semantics. Several projects are currently under development to automate the process by enforcing logical relationships that can be navigated by computerized algorithms. These projects will serve not only to increase the quality of new entries into the GO, but will also assist in retrofitting earlier GO entries so that they are transparent to computerized searches.

REFERENCES

[1] Hill, David P., Barry Smith, Monica S. McAndrews-Hill, and Judith A. Blake. "Gene Ontology annotations: what they mean and where they come from." BMC Bioinformatics, 2008.

[2] Köhler, Jacob, Katherine Munn, Alexander Rüegg, Andre Skusa, and Barry Smith. "Quality control for terms and definitions in ontologies and taxonomies." BMC Bioinformatics, 2006: 212.

[3] Kumar, Anand, and Barry Smith. "Enhancing GO for the Sake of Clinical Bioinformatics." Proceedings of Bio-Ontologies Workshop, 2004.

[4] —. KI 2003: Advances in Artificial Intelligence (Lecture Notes in Artificial Intelligence 2821),. Edited by A. Günter, R. Kruse and B. Neumann. Berlin: Springer, 2003.

[5] Kumar, Anand, Barry Smith, and Christian Borgelt. "Dependence Relationships between Gene Ontology Terms based on TIGR Gene Product Annotations." CompuTerm International Workshop On Computational Terminology, 2004: 31-38.

[6] Mungall, Christopher J., et al. "Cross-product extensions of the Gene Ontology." Journal of Biomedical Informatics, 2011: 80–86.

[7] Smith, Barry, and Anand Kumar. "On Controlled Vocabularies in Bioinformatics: A Case Study in the Gene Ontology." Drug Discovery Today: BIOSILICO, 2004: 246-252.

[8] Smith, Barry, and Cornelius Rosse. "The Role of Foundational Relations in the Alignment of Biomedical Ontologies." MEDINFO, 2004: 442-448.

[9] Smith, Barry, et al. "Relations in biomedical ontologies." Genome Biology, 2005.

[10] Smith, Barry, Jacob Köhler, and Anand Kumar. "On the Application of Formal Principles to Life Science Data: A Case Study in the Gene Ontology." Lecture Notes in Computer Science, 2004: 79-92.

[11] Smith, Barry, Jennifer Williams, and Steffen Schulze-Kremer. "The Ontology of the Gene Ontology." Proceedings of AMIA Symposium, 2003.

[12] The Gene Ontology. "Extended GO Ontology Relations." the Gene Ontology. 2011. http://www.geneontology.org/ (accessed April 5, 2011).

[13] —. GO Annotation for the Cardiovascular System. 2011. http://geneontology.org/GO.cardio.shtml (accessed April 15, 2011).

[14] —. GO Annotation for the Renal System. 2011. http://geneontology.org/GO.renal.shtml (accessed April 15, 2011).

[15] —. The Reference Genome Annotation Project. 2011. http://geneontology.org/GO.refgenome.shtml (accessed April 10, 2011).

[16] Wikipedia contributors, "Gene Ontology," Wikipedia, The Free Encyclopedia, ttp://en.wikipedia.org/wiki/Gene_Ontology (accessed March 9, 2011)