Applied Ontology and Poly-Ontology in Biomedical R&D

Peter L. Gabel
Arity Corporation
200 Friberg Pkwy
Westborough, Ma 01581
508-272-9947
Peter.Gabel@Arity.com

Steven W. Adams
Pfizer Inc.
Eastern Point Road
MS 8274-1256
Groton, CT 06340
860-715-2147
Steven.W.Adams@Pfizer.com

Josh Mugele
Pfizer Inc.
2800 Plymouth Rd.
Ann Arbor, Mi 48105
734-622-7824
Josh.Mugele@Pfizer.com

Pamela T. Schaepe
Arity Corporation
200 Friberg Pkwy
Westborough, Ma 01581
508-272-5346
Pamela.Schaepe@Arity.com

ABSTRACT

Pfizer Inc. collaborates with Arity Corporation on two projects which address needs for enterprise-scale software systems to manage and use biomedical ontologies. In designing these systems, we confronted certain challenges, namely

- The presence at Pfizer of multiple inter-related and nested ontologies that must be managed in concert;
- Varying levels of abstraction in a given ontology;
- Inconsistencies and contradictions within source ontologies;
- The need for "applied ontologies" that fit into current scientific processes; and
- The need for high precision in order to accurately inference and hypothecate on collected data.

To address these issues we developed guidelines for ontology development and management in biomedical R&D. This paper describes those guidelines and the systems that were developed.

Categories and Subject Descriptors

I.2.3 [**Artificial Intelligence**]: Knowledge Representation Formalisms and Methods – *Frames and scripts, Representation Languages, Representations (procedural and rule-based), Semantic Networks*

General Terms

Algorithms, Management, Design, Experimentation, Human Factors

Keywords

Description Logic, Knowledge Representation, Lexicon, Natural Language Processing (NLP), Ontology, Applied Ontology, Poly-Ontology, Biomedical Ontology.

1. INTRODUCTION

It is well recognized that ontologies can be used in large biomedical research organizations, such as at Pfizer, to capture organizational knowledge and to enhance analysis and hypothesis generation [2,4]. Traditional ontology building (as, for example, described in Noy, et al. [8]) allows for construction of large-scale ontologies that can be based on pre-existing terminology structures (taxonomies, thesauri, etc.), on structured data schemas, and on evidence, namely facts which are often represented in RDF-style triples. Facts can be identified from data in a variety of ways, but usually require markup within the source data or identification by knowledge workers. Facts can be determined by automated data analysis tools on structured data or text mining tools using NLP on unstructured text, but the output from these tools still require curation of the collected facts by knowledge workers in order to ensure high levels of accuracy.

Traditional methods of ontology building fall short in certain areas. First, they require a priesthood of author/editors or knowledge engineers to create and maintain ontologies over time. Second, traditional methods fail to sufficiently account for the needs of large biomedical organizations. Commercial R&D organizations require ontology building systems and methodologies to be incorporated within established and standardized scientific and IT processes. These systems must be fully scalable. They must incorporate externally curated ontologies with frequent and varying update schedules. Different ontologies require different levels of abstraction over their subject matter. And different subject matter groups may require different workflow solutions. While we recognize the need to fully embrace traditional methods of ontology construction, we have found that the traditional methods fail to meet these needs of large biomedical organizations. We have extended traditional methods in the forms of "applied ontologies" and "poly-ontology" to address these problems and we advocate their study and adoption.

1.1 Applied Ontology

An applied ontology is to an ontology as an applied science is to a pure science. The applied ontology is used in service of a goal. The @tlas™ applied ontology system (described in greater detail in the "@tlas" section) extends the reach of traditional ontologies and provides an investigative science or business process interface over an ontology infrastructure and, by doing so, enables scientists (and others), in the course of normal activities, to directly interact with and contribute to relevant knowledge assets and data. An applied ontology combines the ontology concept with process modeling and user experience. The approach, comprised of both technology and methodology introduces efficiencies into scientific processes that, we contend, reduces cycle time and leverages the scale of large organizations. This applied approach also allows users (domain experts and knowledge workers) to both utilize and extend the ontology infrastructure in the course of normal scientific activity without the necessity of specialized training in ontology construction.

ECCB'05 Workshop on Biomedical Ontologies and Text Processing, September 28, 2005, Madrid, Spain.

1.2 Poly-Ontology

We advocate systems that employ a poly-ontology approach to enterprise-wide projects. This is an approach that accounts for multiple ontologies of varying constructions, abstractions, and inter-relations. The poly-ontology approach that we have implemented in the LexiLink™ system creates and maintains meta-linkages between distinct ontologies and provides a common viewing interface as well as the capability to search and reason across the different ontologies.

There are a number of reasons for maintaining multiple (but inter-related) ontologies in an organization rather than creating a single organizational ontology, which we will only summarize here. Each ontology expresses a point of view and has a particular goal in mind. For example, both the marketing function and the safety science function at a pharmaceutical organization may have reason to address the topic of Alzheimer's disease. However, each of them would require information at very different levels of abstraction from the broad description of symptoms of the disease to the interactions of cell enzymes in the brain. Further, ontologies often encode contradictory or inconsistent knowledge which cannot be completely resolved but must be managed.

Another reason for maintaining multiple, inter-related ontologies is the requirement to automatically notify the curators that are responsible for a particular ontology that specific changes have occurred within some concept definitions of an ontology that was used as a resource. For example, UMLS is updated quarterly. Curators may use portions of UMLS as a resource for their ontologies. It is a requirement that they be notified of specific changes in a new version of UMLS that apply to each and only those specific concepts within their ontologies that are related.

1.3 Overview of Projects

Pfizer Inc. and Arity Corporation collaborated on two ontology-related projects in Pfizer Global R&D: A safety science client-server platform called @tlas and an ontology management client-server system in Information Management called LexiLink. These were constructed using Arity's Information Animator™ platform. Information Animator provides a complement of components and tools for knowledge representation, information extraction from text, discovery, reasoning, and application development based on Web Services and Web GUIs. KR&R services are derived from a description logic, rule, and logic programming engines. Discovery services are provided by association and cluster data mining. Information extraction services use cascaded parsing combined with named entity recognizers and large-scale use of lexicons derived from ontologies.

2. @tlas

@tlas was created by Arity Corporation in collaboration with safety scientists from Pfizer Global Research and Development to support the investigative process to evaluate safety signals for compounds and targets through the cycles of literature and laboratory study. The underlying ontological structures serve to both provide data to the scientists as well as record the relationships and evidence about knowledge that the scientists collect in the process of their investigations. The record of strategy and point of view development has both short and long term benefit to Pfizer. The integration and exchange with data resources, search strategies and existing ontologies creates a cohesive environment for productivity.

The sections of the program map serve as a "spreadsheet for thinking", focusing attention on various critical components in the study process. Each section is supported by reference ontologies (that are assembled and edited in LexiLink, described in the next section) and data imports from various Pfizer internal sources and from additional sources such as Medline.

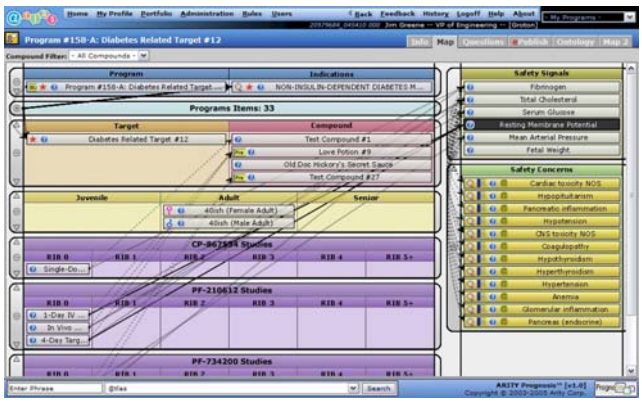

Figure 1: The @tlas Program Map

The next layer of Atlas interface focuses attention on evidence collection, exchange of ideas, and publication. In a given section, the user is guided through a series of questions that reflect the practices of scientific investigation around that topic. The questions reflect the standard workflow processes established in Pfizer Safety Sciences and in exceptions such as the use of non-standard biomarkers. For each question, suggested answers are provided to the user based on

- Term generation based on analysis of the underlying poly-ontology
- Inference using analogous reasoning
- Inference using heuristics generated from data mining of Medline and other sources
- NLP of additional Pfizer and non-Pfizer information sources.

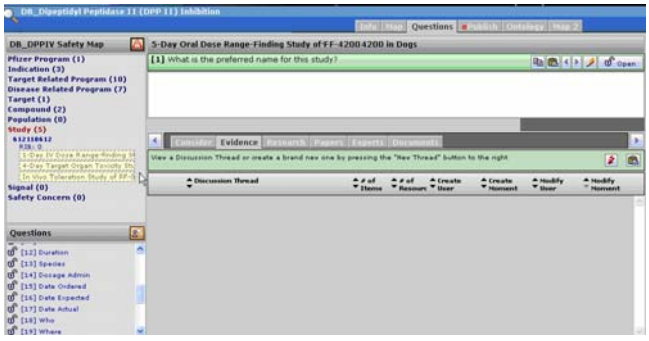

Figure 2:The @tlas evidence collection page.

Scientists answer the questions based on experimentation and research in the application. Answers are analyzed by the system and drive further inference-generated suggestions. Answers are also used to supplement the underlying ontology which in turn guides reasoning of future studies. In essence, the scientist is

generating an ontology through an application directly related to his or her work. The benefits of working within this system are twofold: increased reasoning for future studies (which leads to increased efficiency of study execution as well as increased accuracy of study results); and increased collaboration – scientists share results, eliminating duplication of effort.

Figure 3: Evidence Presented in Support of an Inference

3. LexiLink

LexiLink was implemented by Arity Corporation for the Information Management department of Pfizer Global Research and Development to manage the lexicons – taxonomies, thesauri, and controlled vocabulary – licensed or created by Pfizer (though the system currently manages lexicons as defined above, the framework supports richer ontology development and management). LexiLink allows multiple concurrent users to manage multiple ontology-like structures through a web-browser interface

3.1 Multiple Ontologies

LexiLink stores multiple ontologies. Sizes range from hundreds of thousands of terms and relationships (e.g. WordNet) to under 100 terms and relationships (e.g. proprietary Pfizer Therapeutic Areas List). Ontologies are created within LexiLink or are imported from external sources using XML files. For imported ontologies, update schedules are established according to frequency of update of the original source. Update frequencies range from daily or weekly (e.g. GO, the Gene Ontology) to quarterly or semi-annually (e.g. MeSH). LexiLink is built on a proprietary platform for knowledge-based systems called Information Animator from Arity. Information Animator optimizes the storage and retrieval of ontological structures and is advantageous compared to a traditional relational database granting the system the scalability required in a large organization.

LexiLink does not enforce a standard ontology schema. It allows for a schema definition on a per-ontology basis. As examples, LexiLink currently has ontologies defined using the traditional thesaurus schema (BT/RT/NT/Use/Use For), traditional taxonomic and partonomic schemas (IS-A and IS-A-PART-OF), and entirely unique schemas such as that of WordNet which contains grammatical relationships. Examples of the last include Antonyms, Noun Pertainyms, and Derivationally Related Verbs. Schemas are established in LexiLink manually or as part of the ontology import process. LexiLink also supports frame-oriented ontology schemas that are common to many biochemical ontologies.

3.2 Traditional Ontology Editing

LexiLink supports traditional ontology editing and provides a tree-style ontology editor similar to Protégé 2000 and to OBO-Edit (nee DAG-Edit). LexiLink also allows users to manage non-hierarchical links between domains within an ontology and between lexicons. Metadata for a given concept (i.e. descriptions, synonyms, creator, source, etc.) are also managed in the primary editing display. Standard editing functions include adding, deleting, or modifying concepts or relationships. Addition of concepts or relationships can be accomplished manually or by copying from other ontologies within LexiLink (individual concepts, synonyms, or entire sub-trees or subsets can be copied).

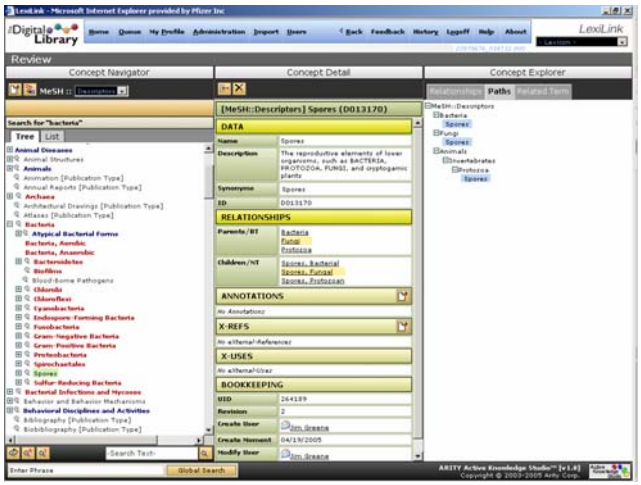

Figure 4: The primary editing interface in LexiLink.

3.3 Links Between Ontologies

LexiLink allows for links called external references or XREFs between concepts in unrelated ontologies. XREFs may be of a variety of types, including custom types. Examples of XREFs include “migrate”, which indicates a reference source for a concept, and “same as”, which indicates synonymous relationship across ontologies. XREFs may be thought of as a generalization of the representation used in the Open Biology Ontology (OBO) format that is used for the Gene Ontology (GO).

XREFs may be created manually or they may be created systematically. Two types of systematic XREF creation are current supported. First, when a concept or a synonym is copied from one ontology to another, a “migrate” XREF is automatically created indicating the source of the copied concept as well as the destination. “Migrate” XREFs serve as source references for terms in the destination ontology. If a change occurs in the source ontology (usually through the import of a new version), the destination ontology may be notified of the change through the alerting feature in LexiLink using the existing “migrate” XREF.

The second type of systematic XREF creation in LexiLink is generated by the automatic inference of similarity between concepts in different ontologies (creating a “similarity” XREF). Similar concepts are inferred based on similarities in the morphology of the concept name and synonyms or based on properties or relationships that are analogous between concepts.

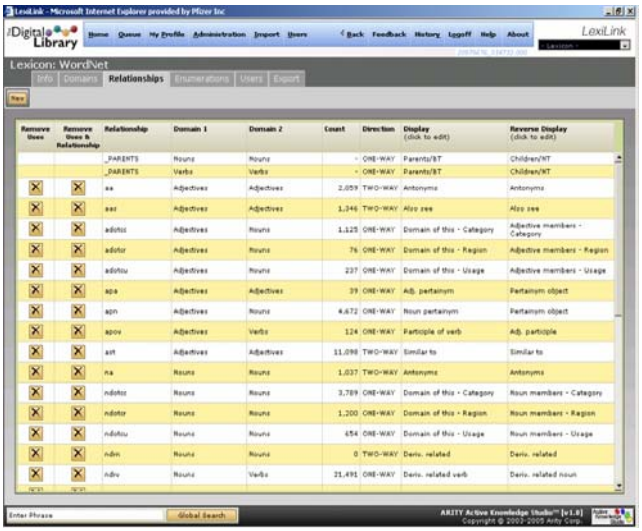

Figure 5: LexiLink Schema definitions for WordNet 2.1

The network of XREFs between ontologies creates the poly-ontology discussed previously. The poly-ontology can be thought of as an active meta-ontology where the LexiLink system itself is the meta-ontology framework.

Figure 6: The Poly-Ontology relationship editing interface in LexiLink. This pane allows users to copy concepts and to create and manage XREFs between ontologies.

3.4 Role Management

Ownership of ontologies in LexiLink is complex. Different business groups within Pfizer own different ontologies, and different licensing restrictions govern the different externally licensed ontologies in Pfizer. Ontology editing within LexiLink must also conform to business and scientific rules in place at Pfizer to guarantee minimization of mistakes and a high degree of accuracy.

To address these needs we implemented defined roles within LexiLink. The roles in LexiLink are editor, author, and administrator. Editors modify properties of a concept; authors in addition to performing editing tasks also can modify, add, or delete concepts; administrators perform the role of editors and authors and design ontologies and manage ontological properties. A user may have different roles for different lexicons. A user's role is detected and authenticated systematically in compliance with Pfizer's authentication systems.

3.5 Searching Capabilities

LexiLink contains robust search capabilities. Users can search individual Lexicons or across all Lexicons for which they have permissions. Users can search against all fields including name, synonyms, and relationships. Results of a search are displayed in a list or within the context of the ontology tree. LexiLink APIs also allow searches of concepts from other systems. These searches can use the poly-ontology nature of LexiLink to retrieve networks of concepts across ontologies.

Figure 7: The LexiLink Search Panel

4. GUIDELINES

We defined the following guidelines to accompany our implementation of LexiLink and @tlas applied and poly-ontologies. These guidelines describe the requisite steps for establishing processes and systems for using and managing ontologies in a large biomedical organization.

4.1 Set Up a Governing Body

A governing body oversees the organizational and functional goals of ontology production. It acts as sponsor and advocate for the endeavor. It also creates an authority structure to guide ontology development teams in coordinating activities and resolving conflicts that might arise in stylistic preferences or approaches to content.

4.2 Model a Range of Worldviews

Those who grapple with the genome and those who consider competitor pipelines will necessarily slice the world differently at various levels of both the abstract and the concrete. Different perspectives allow for the emergence of new ideas. In addition, some points of view might be controversial or speculative. The ability to create an ontology based on an "experimental" worldview may provide novel ways to look at data and transform

basic assumptions. Identification of the worldviews that affect an ontology system is essential.

4.3 Identify User Roles and Flow of Work

Defining work processes, workflow and collaboration model(s) is an essential step. Each user is granted authority based on their role in the business process. Beyond the basic roles of browser and editor, are roles such as legal reviewer or chemistry reviewer. Workflow is instigated when reviewers are part of the work process. The collaboration model describes how much of the ontology or applied ontology is visible to browsers while the work is under construction.

4.4 Coordinate the Efforts of Domain Experts and KM Experts

The quality of ontologies and data mining efforts is almost always much higher when an effective collaboration between domain experts and KM method and tool experts is created.

4.5 Identify Data Sources

The broadest range of relevant data sources are identified and explored. Data sources that are not immediately accessible must be exposed for greatest value. Where possible, existing ontologies are included. Existing ontologies are a great resource to mine for content, as well as examine for style to emulate or adapt. LexiLink, for example, holds several Pfizer internal and external collections for consideration: GO, WordNet, portions of UMLS, MeSH, Atlas Safety Signals, and other internal lexicons.

One must be careful in choosing sources to include in ontology systems. We have found that external systems often contain logical errors in construction that must be corrected prior to being used for analysis. For example, some versions of MeSH contain hierarchical cycles ('A' is a BT of 'B', 'B' is a BT of 'C', 'C' is a BT of 'A') as well as redundant ancestors ('A' is a BT of 'B', 'B' is a BT of 'C', 'A' is a BT of 'C').

4.6 Identify the Style of each Ontology

Ontologies are generally created in either a frame or semantic net style. The frame style emphasizes the primacy of concepts and their description. The semantic net style emphasizes the relationships between concepts with information tied to the relationship links. Either style has merit.

Overall, just as in thesaurus development there has been a broad shift in the opinion of lexicographers towards concept-centric representations over the last ten years (for example, the successes of UMLS and WordNet), there is an emergent consensus favoring frame based approaches within the KM community. Much of this emergence of consensus is because of the increasing importance of inference using ontologies as knowledge bases.

The poly-ontology approach allows relationships to be maintained between ontologies of differing styles. Thus, using the poly-ontology approach, individual managed ontologies can be maintained using its own style, whether based on historical choices or on the preferences of its curators.

Often continuity of style is constrained by the circumstances, such as when an ontology is licensed from an external source.

4.7 Identify and Precisely Describe Relationships

A central dogma of cognitive science is that a concept is an idea inferred or derived from specific instances, but *defined* in terms of more general concepts and its characterizing properties and relationships. In an ontology, information about a concept includes:

- A description or definition
- Synonyms and lexical / morphological assertions
- Properties
- Relationships to other concepts

In the complex context of life sciences, it is fundamentally important to employ semantic precision. Assumptions must be made explicit so they can be recognized and altered when appropriate. A domain should be described or represented in declarative language that is as specific as possible, so as to constrain interpretation accurately.

A potential pitfall in ontology development is shifting word sense. Because individual words can have multiple meanings and those meanings can be different in either subtle or distinct ways, the language used to describe concepts should be carefully chosen. Shifting word sense can cause serious errors in reasoning over the data. Not only does an ontology communicate the meaning of subject-matter between people, but also between people and machines as we rely on machines to process and to do the heavy lifting of volumes of data analysis.

Precision in descriptions about relationships enables the construction of intelligent software agents to act on behalf of people. As an example, accurately describing partonomic relationships (how concepts aggregate) provides more specific information about a type of "is a" relationship. This increases the ability to capture the transitivity of relationships and, consequent inferences that can be made. Winston et al [10] differentiate six types of containment, as follows:

Containment Type	**Example**
Component/object	Wheel/car
Member/collection	Tree/forest
Portion/mass	Slice/cake
Stuff/object	Wood/house
Feature/activity	Paying/shopping
Place/area	Groton/Connecticut

Chaffin, Hermann, and Winston [1] add a seventh:

Containment Type	**Example**
Phase/process	Adolescence/growing up

More precise characterization of partonomic relationships (and the creation of relationship type hierarchies) can be built using these types of containment as a foundation.

4.8 Integrate Bottom-up and Top-down Approaches in the Design

The process of designing an ontology is generally a blend of induction and deduction – general concepts to specific instances (top down), or specific instances to general concepts (bottom up). A designer interviews experts to begin to understand the framework to build and then starts to fill in the framework through text mining of documents. In the process, a designer might discover nuances or new branches of thought that require modifications to the original framework. Flexibility in the tool set and agility in its application are critical.

4.9 Apply Description Logic Technology to the Development Process

In ontology development integrity verification is critically important. Because public ontologies like MeSH and UMLS have errors in parent-child relationships, LexiLink has a built-in automated check and repair for definitional cycle integrity. @tlas uses reasoning technologies to identify similarities between concept features and suggest, for example, alternative or additional safety signals or concerns to be considered. A rule set is generated by information extraction techniques and accepted or rejected by scientists as they work, thus assisting the system in validating or learning when to apply the rule.

4.10 Deploy Applied Ontologies derived from Poly-Ontology Building Methods

Traditional ontology development techniques do not suffice for many complex applications. In particular, the nature of applied ontology applications typically require a broad set of knowledge domains to be integrated. The ontologies for each of these domains may be characterized by ontologies and knowledge sources that have origins both inside of and outside of the enterprise and may be curated or managed independently.

The poly-ontology approach provides the "conceptual glue" that unifies disparate knowledge into a useful whole. LexiLink has been successfully used to integrate many ontologies spanning biochemical and pharmaceutical domains and provides the basis for the knowledge provided to and maintained by Safety Scientists in their use of the @tlas system.

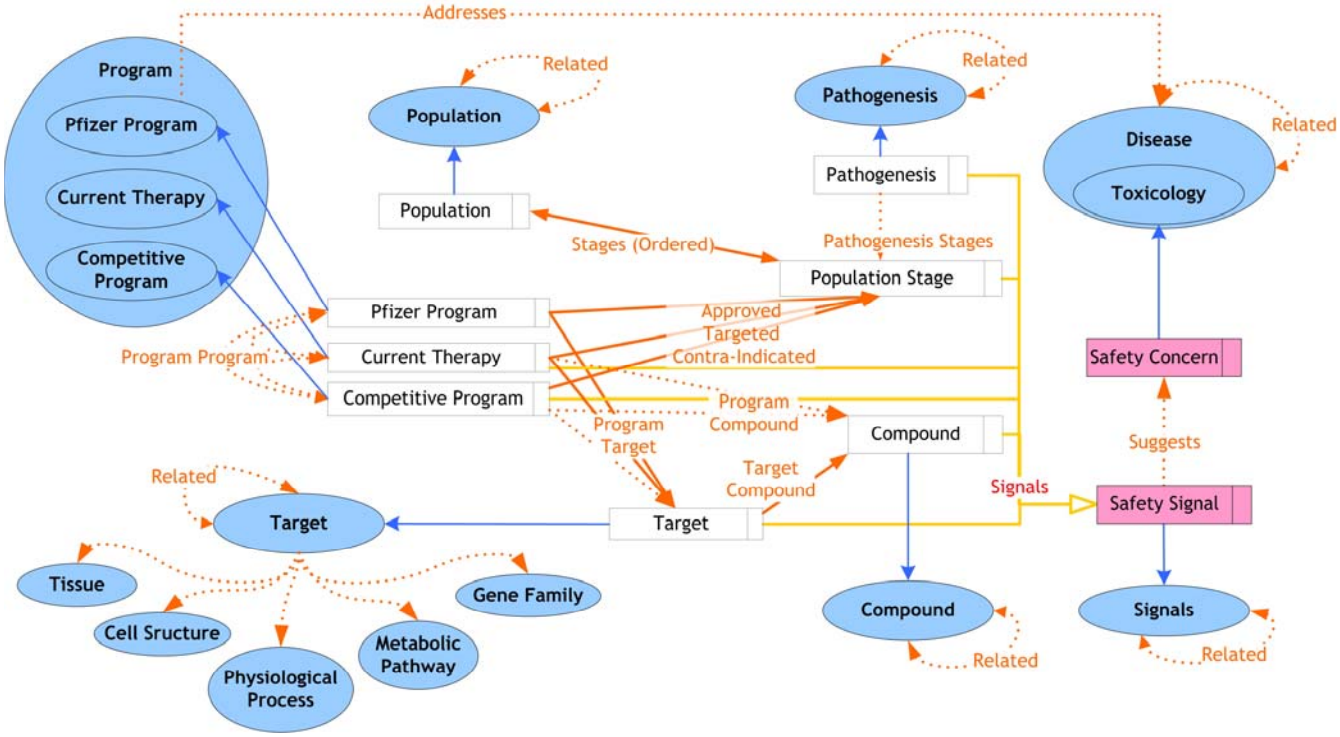

Figure 8:Simplified diagram of the @tlas Poly-Ontology integrating many separately curated ontologies

5. CONCLUSION

In conclusion, this approach to Enterprise-scaled Ontology development requires thoughtful analysis and interaction between subject-matter experts and ontology building experts. The traditional form of ontology building and the applied form interact and feed each other. This integration of styles provides a two-way street of information gathering and curation. The step-by-step process of ontology development culminating in precise relationship descriptions will enhance the ability to use inference engines. Meanwhile, the poly-ontology of nested and inter-related ontologies helps to address multiple organizational goals, isolate user roles and authority appropriately, and provide the opportunity for re-use of data and/or stylistic elements.

6. ACKNOWLEDGMENTS

Our special thanks to Jim Greene of Arity Corporation and Mike Roos of Pfizer Inc. for their contributions to this article and to the @tlas and LexiLink projects. Also thanks to the anonymous reviewers for their helpful comments.

7. REFERENCES

[1] Chaffin, R., Herrmann D, Winston, M.E. An Empirical Taxonomy of Part-Whole Relations: Effects of Part-Whole Relation Type on Relation Identification, *Language and Cognitive Processes 3*, Utrecht: VNU Science Press, 17-48.

[2] Degtyarenko, K. Chemical Vocabularies and Ontologies for Bioinformatics. In *Proceedings of the 2003 International Chemical Information Conference*, Nimes, France, 2003. http://www.infonortics.com/chemical/03proc.html

[3] Fellbaum, C., Ed. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA, 1998.

[4] Gangemi, A., Pisanelli, D., and Steve, G. An Overview of the ONIONS Project: Applying Ontologies to the Integration of Medical Terminologies, *ITBM-CNR*, (V Marx 15), 1-42.

[5] Kubinyi, H. *Changing Paradigms in Drug Discovery*. Lecture, University Heidelberg. Weisenheim am Sand, Germany, 2004. http://www.biomod.org/complexity2004/kubinyi.pdf

[6] National Library of Medicine/ Medline. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi.

[7] National Library of Medicine /UMLS http:www.nlm.nih.gov/research/umls.

[8] Noy, N. F., and McGuiness, D. L. Ontology Development 101: A Guide to Creating Your First Ontology. http://protege.standford.edu/publications/

[9] Sowa, J.F. Knowledge Representation: Logical, Philosophical, and Computational Foundations, Brooks Cole, Pacific Grove, CA, 1999.

[10] Winston, M, Chaffin, R., and Herrmann, D. *A Taxonomy of Part-Whole Relations*, Cognitive Science, 11 (1987) 417-4

[11] W3C Web-Ontology (WebOnt) Working Group http://www.w3.org/2001/sw/WebOnt