Supporting the Analytic Knowledge Manager: Formal Methods for Ontology Display and Management

Alan Chappell, Anthony Bladek, Cliff Joslyn, Eric Marshall,
Liam McGrath, Patrick Paulson, Sean Stolberg, and Amanda White

Pacific Northwest National Laboratory

***Abstract* — The Intelligence Community and other analytic-focused communities are developing and implementing large knowledge bases and semantic-based systems. These systems require new activities for managing their ontological underpinning, including a range of tasks from supporting domain description and evolution to integrating multiple source of semantic information. Beyond the role of the analyst or the traditional data base administrator, the role of the knowledge manager as the point of focus for such activities is growing in prominence. We are developing methods and tools to provide an analytical ability for the display and management of ontological systems, rooted in the formal properties of semantic relations in semantic graphs, and the semantic hierarchies in which they are valued. We describe methods for display, integration, and management of ontological resources to support the emerging Analytical Knowledge Manager with the AKEA tool.**

***Index terms* — Knowledge management, knowledge manager, ontology visualization, ontology alignment.**

I. Introduction

In this paper we address the needs of the “Analytic Knowledge Manager” (AKM), a hypothetical actor whose responsibilities are to manage not the underlying data of an analytical organization, but rather the collection of its semantic information, ontologies, and schemata. The semantic domain of an enterprise is linked both to the content of its data and the applications in which those data are used. Thus the AKM must respond to the needs of a particular analytical/scientific function in much the same way that the IT manager responds to the needs of the business function of an organization.

Since the role of the AKM is a relatively recent evolution, most organizations splinter the associated functions among multiple actors, each performing AKM functions as adjuncts to data processing pathways established before the organization incorporated semantic processing. These actors include the producers of the data; the end users of the data (e.g. intelligence analysts), those who store and provide access to the data (e.g. IT and DBMs); and intermediaries (e.g. web site managers, web programmers, information retrieval specialists, and anyone who must interpret, transform, or manipulate the data).

Large organizations typically provide partial support for AKM roles through formal groups. These groups include an IT department, a librarian, and a technical support group, each of which must understand and support multiple user communities within that organization. This ultimately leaves the end user—the intelligence analyst—with many of the tasks of the knowledge manager. These shared tasks include: how to construct requests for data, how to access the resultant data, and how to integrate them into an analysis. *And,* since the required semantics of data can be lost or modified by the many *de-facto* AKMs along the data delivery chain (including all those listed previously), it may be impossible for the analyst to retrieve information related to an intelligence problem or the metadata necessary to determine the quality of data.

We find that many workgroups within the IC already rely, formally or informally, on selected member of the workgroup to assist others with AKM functions. This person typically is technology “savvy” and skilled in the use of a wide set of data access and transformation tools. Unfortunately, this *ad hoc* role often is under-recognized and under-resourced, which can exacerbate the workload of the individual even if enhancing the effectiveness of the workgroup.

We argue that a recognition of the AKM role in terms of its responsibilities and the support it requires will allow an intelligence enterprise to more effectively find the data required for a particular analysis task, allow the analyst to understand the quality and provenance of data, and help prevent the analyst from being overwhelmed by data not pertaining to the current problem. Ultimately, a formal assessment of AKM roles may assist with understanding access control and separation of duty considerations.

In this paper we use the RASCI (Responsible, Accountable, Supportive, Consulted, Informed) framework [1] to define the AKM role, its responsibilities, and the support required for the AKM role. We then describe typical AKM tasks in the context of ontology management, including analysis and linkage. We describe our approach to supporting such AKM tasks on ontologies through the formal analysis of the mathematical properties of link types, and in particular the manipulation of semantic hierarchies. We conclude by illustrating our implementation of these methods within the AKEA tool.

II. Background

Knowledge Management (KM) is a discipline that strives to organize and preserve knowledge, making it accessible to the enterprise [8]. In the domain of intelligence analysis, the primary knowledge is fluid and tied to specific analytical problems.

A. The Potential Roles of AKMs

We envision the following to be the primary responsibilities of the AKM:

1. To enable access to information by analysts that fulfills the requirements of a particular analytic problem
2. To provide queries to data sources using the semantics and syntax expected by the data source
3. To interpret the provided information within the context of the analytic problem, and
4. To provide the supporting data required to determine the quality and provenance of the delivered information.

Table I applies RASCI charting to describe the relationship of the AKM role to the other roles within the intelligence organization. For each activity in a process, a RASCI chart identifies who is responsible for carrying the activity, who is accountable for the result, who provides support for the activity, who is consulted in carrying out the activity, and who is informed about the status of the activity. In Table I, we list only the activities for which the AKM is *responsible* (R). In carrying out or enabling the activities, the AKM may have to *consult* (C) with other roles. For example, in order to determine the type of data that corresponds to a request from an analyst, the AKM will need to consult with the analyst and subject matter experts in order to build a representation of the terminology used in the problem domain and the relationships between terms. Once an activity is completed, other roles may have to be *informed* (I)—for example, the analyst must be informed when requested data has been delivered. Finally, the chart specifies who the AKM is *accountable* (A) to for the specified activity. As an example, the analyst acknowledges that delivered data matches their requirements and that is placed in the proper context in their analysis tools.

Table I
RASCI Matrix for AKM

Activity	*AKM*	*Analyst*	*Librarian/ Manager of Data Source*
Provide analyst with appropriate data	R	C/I/A	C
Provide queries in the semantics of the data source	R	C	C/I/A
Interpret delivered data in the context of specified analytical problem	R	C/I/A	
Provide provenance and metadata for interpretation of data quality	R	I/A	C

B. The Current State: de Facto AKMs

Given the complexity of the AKM's task, the heavy dependence on knowledge that is tightly bound to particular problems and problem domains, and the need to access data with many formats and from many sources—each with their own set of semantics—it is understandable that the role of AKM either has been ignored or distributed to other parts of the organization. This leads to a lack of responsibility and accountability for the activities that should belong to the AKM.

Table II applies the RASCI chart method to this *current* state of affairs. The analyst is both responsible and held accountable for almost all activities, allowing no check on the suitability of data for an analytic task. Responsibility is often split between the analyst, who must use data supplied by tools, and the developers of tools that deliver data. Having multiple roles responsible for the same activity can lead to conflict and the activity not being completed, since the 'buck' doesn't stop at a specific doorstep.

C. Assessing the Needs of the AKM

In order to carry out the described primary responsibilities, the knowledge manager must further:

1. Understand the data requirements of the user community in terms of the *semantics* of the particular problem domains of interest to that community
2. Explore and understand potential information sources and determine the relevance of the provided data to the

Table II
RASCI Matrix for the real world

Activity	*Analyst*	*Librarian/ Manager of Data Source*	*Programmer/ Developer*
Understand problem semantics	R/A		
Explore data sources	R/A		R/C
Create appropriate queries for data sources	R/A	C	R/C
Interpret delivered data within problem domain	R/A		R
Provide provenance and metadata to ensure data quality	R/A	C	R

community of interest

3. Adapt requests for data to the format required by the selected data source, without losing the semantic meaning implied by the request or the retrieved data
4. Support delivery of information to analysts and analytic systems using representations and semantics appropriate to its intended use
5. Provide appropriate metadata – such as the original source, publication date, and processing work-flow – of any data provided in response to a request
6. Ensure that security protocols are invoked properly so that data are only available to those with the necessary credentials for obtaining the data

Given these requirements, one of the primary needs of the AKM is an ontology that describes the semantics of the target analytical domain. The ontology describes the semantic meaning of potential queries and the relationship between terms. The ontology describes basic attributes of terminology, such as composition or subsumption, and may also describe more advanced notions, such as formal definitions of terms in terms of primitive assertions. The knowledge manager will most likely need to develop or adapt much of this ontology so that it serves the needs of the user community, and will use a variety of tools to present the ontology to end users in order to validate its content and to ensure the its consistency.

In order to determine if a potential data source will be useful to their knowledge consumers, AKMs must be able to access both the semantics of a data source, preferably through an ontology, and the relationship of the data delivered by the data source to the source's domain ontology. This can be a large bottleneck, since many data sources provide neither. The resulting lack of formalized knowledge forces the knowledge manager to define both of these using whatever sources are available, including database schemas, XML schemas, and—mostly—common-sense. Identifying the correct semantics of data retrieved from the source can be particularly onerous, potentially requiring specialized tools to scrape source documents, information extraction software employing natural language processing, and, in the worse case, hand annotation.

In order to provide data that meets an analyst's needs, the knowledge manager needs the ability to understand the relationship between terms used by client analysts and the terminology used by specific data sources. Visualizing and understanding these relationships is at the core of generating appropriate queries and presenting data within the analyst's problem context.

Finally, the AKM must have access to metadata describing data provenance. For each data element, metadata describing its source, e.g. the date of publication, the original source, etc, and documenting its history of analytic or prepatory steps should be made available to end users. Such metadata enable users to understand the quality of the delivered data as well as enabling repetition of results.

We now describe tools which are relevant to support the AKM functions within the intelligence organization.

1) Tools for the Analyst as the AKM: Often the AKM role is delegated entirely to the analyst, who must determine the best key words to use to bridge from requirements to the documents of a data source, and understand the terminology used across multiple disciplines. The advantage to this approach is that the analyst has direct knowledge of the source and provenance of the data that is obtained. However, few tools are provided to support the analyst within the AKM role beyond the firm grounding of the analyst in select disciplines and the ability of the analyst to quickly adapt to changing conditions and new data sources.

2) *AKM Tools for the Data-Base Manager:* For structured data sources, there is at least some schema or description of the type of data that can be expected, and maybe even some business rules that can be used to infer relationships between data. Here standard structured data tools such as schema editors and query engines can be used, with the assistance of a knowledgeable data-base manager (DBM), to deliver appropriately annotated data to the analyst.

Intelligence organizations also work with their own knowledge and data repositories. These repositories can have some known data semantics and relationships, although those semantics often are only loosely related to individual problem semantics. The data manager can use standard database management tools to organize and provide these repositories. While an experienced DBM may have an understanding of the semantics of stored data that could be of use to the analyst and application programmers, they may not be able to address questions about semantics outside of what is needed to provide reliable performance and data security.

3) Tools for AKM Role of Application Programmers and Web Developers: AKM tasks are also supported by application programmers and web developers that provide analytical tools. Often it is left to a programmer to determine where a required piece of data resides within a source repository and where to map that data into the analyst's resident databases. It also is up to the designers and implementers of these tools to ensure that all requisite provenance and metadata is carried along with the data—failing to make this requirement known may result in the analyst obtaining interesting, but unusable, information.

There are also few tools to support the AKM role of the application programmer, who are left with the same tools as the DBM, along with less structured tools such as XML schemas and tags, to determine the semantics of data they obtain from web sites and other sources. The programmer needs to coordinate not only with the DBM to determine where to best store mined data within a structured data store, but must also use test cases and user acceptance tests to verify that the data delivered is displayed with the correct semantics in deployed tools. These approaches can be effective when such defined requirements are available, but can also be cumbersome and limiting in the dynamic environment of intelligence analysis.

III. Formal Support For the AKM Role

The AKM responsibilities revolve around the generation, maintenance, description, and alignment of ontologies for both the problem domain of the client analysts and of available data sources. Tools to support this task are only now emerging from the research community [9], and often require a large investment of time to master. Given that the AKM role is mostly filled now by application developers, DBMs, and end-users such as intelligence analysts, who already need to master a large number of processing tools, disciplines, and subject matter areas, it’s not surprising that these tools are often not understood and are underutilized.

Current tools for ontology generation and maintenance are generally ontology editors. But AKMs require additional tools to help them accomplish such tasks as:

- Representing domain and source ontologies to end-users to enable validation and understanding
- Mapping or aligning the semantics of data sources to the analyst’s problem domain
- Aiding in the generation of ontologies for new or evolving problem domains

These tools and techniques can also be applied to the metadata associated with data sources to allow data quality and provenance to be available to the intelligence analyst.

Our approach rests on being sensitive to the mathematical properties of the link types present in an ontology, and in particular to their symmetric and transitive properties. Table III shows the primary classes of link types in terms of these mathematical properties, together with their canonical mathematical structures and a simple example.

Table III
Link Type Properties

	Transitive	Symmetric	Example
Directed graph	No	No	A knows B
Simple graph	No	Yes	A friend of B
Partial order	Yes	No	A employer of B
Equivalence classes	Yes	Yes	A sibling of B

In practice, ontologies are dominated by their “hierarchical cores”, specifically their class hierarchies connected by “is-a” subsumptive and “has-part” compositional links. Mathematically, these are partial orders, each corresponding to the transitive, non-symmetric link types exemplified in Table III by the link type “employs”. Additionally, many of the most common links in RDF graphs are transitive, including “causes” “implies” and “precedes”. Any transitive link yields a mathematical structure of a partial order, and makes the machinery of order theory [2] available to exploit these hierarchical constraints. In our past work, we have described techniques based in order theory to support a variety of AKM tasks, including:

- **Clustering and Classification:** Characterizing a portion of a hierarchy (e.g. groups of ontology nodes) to identify common characteristics [10].
- **Alignment:** Casting ontology matching [3] as mappings between hierarchical structures [4].
- **Induction from Source Data:** Using concept lattices to induce ontologies from textual relations [5].
- **Visualization:** Including exploiting the vertical level structure of semantic hierarchies to achieve a satisfactory layout [6].

In general, such a hierarchical analysis, when available, promises complexity reduction, improved user interaction with the knowledge base, and improved layout and visual analytics. Fig. 1 shows a fragment of a semantic graph using the link types present in Table III. Once the hierarchical link type “employs” is identified, the fragment can be laid out according to the hierarchical layout shown in Fig. 2, the

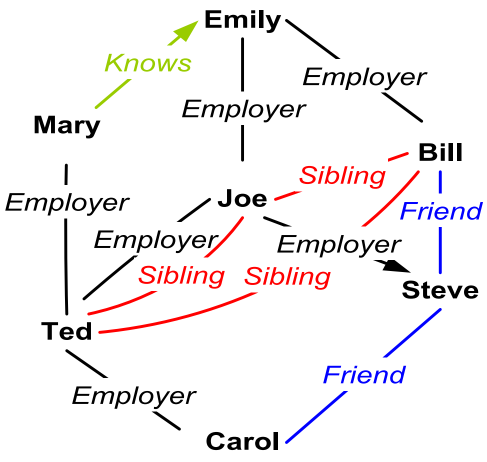


Fig. 1. A simple semantic graph.

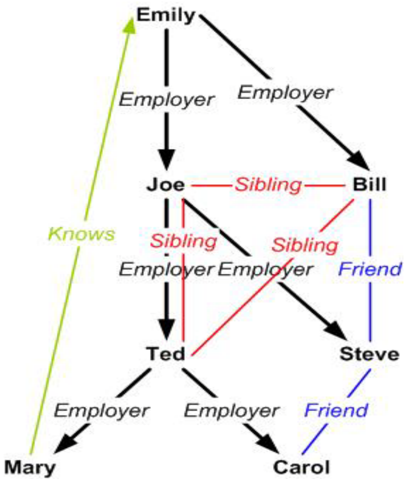


Fig 2: Semantic graph laid out by the hierarchical link type “employs.”

remaining, non-hierarchical link types moving around the central hierarchical structure. The result is a great clarification of the underlying link structure.

Additionally, mathematical properties of the semantic hierarchy, and of particular nodes within it, can be revealed to the user. Especially in large semantic hierarchies where graph drawing and visualization is difficult, it can be critical to report such quantities as:

- The number of nodes
- "Edge density": number of links per node
- "Leaf density": percentage of nodes which are terminals
- Height: maximum chain length from the top to the bottom
- Amount of multiple inheritance: percent of nodes with more than one parent

These quantities are over the whole semantic hierarchy. Additionally, it is useful to be able to provide quantitative assessments of individual nodes in the hierarchy, for example:

- Depth: Number of levels down from the top
- Height: Number of levels up from the bottom
- Number of children
- Number of total descendants
- Number of parents
- Number of total ancestors

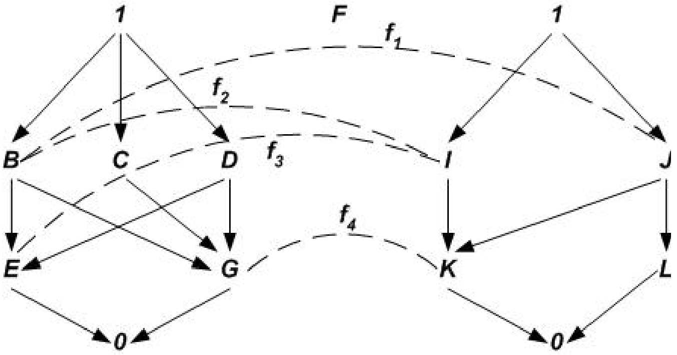
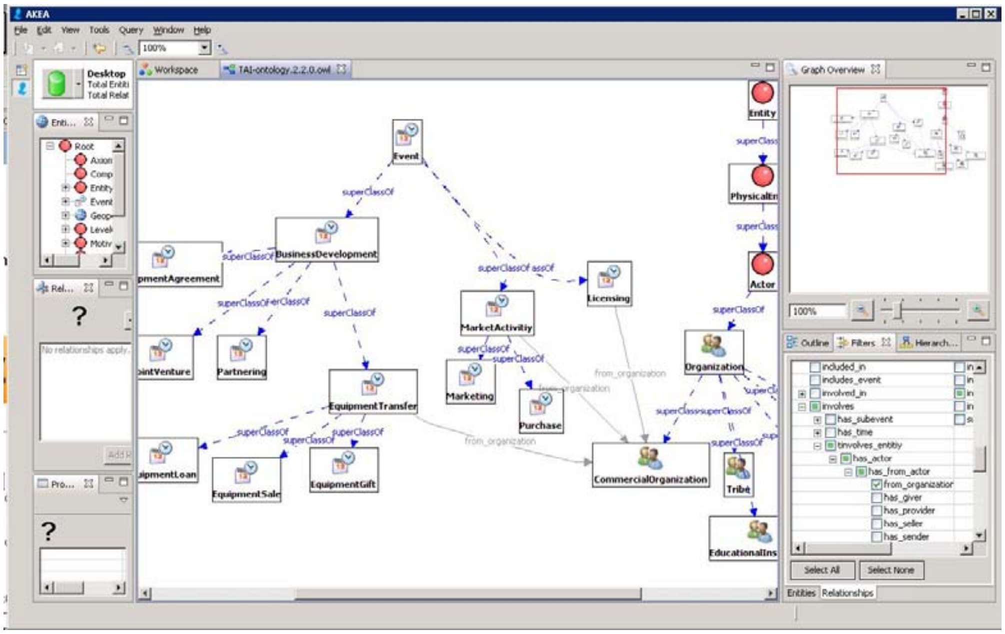


Fig. 3. A simple semantic hierarchy alignment example.

Such quantifications are very useful when performing alignment tasks. Fig 3 shows a small example of an alignment between two semantic hierarchies. Our prior work [4] has proposed methods for measuring the quality of such alignments based on such measures. And when alignment is performed interactively within a GUI-based tool suite such as PROMPT within the Protégé tool [7], augmentation with such statistics will provide the AKM with the context needed to understand the quality of the proposed mappings. For example, in Fig. 3, it is valuable to map nodes high in the structure on the left to those high in the structure on the right, requiring the kind of quantification we have proposed here.

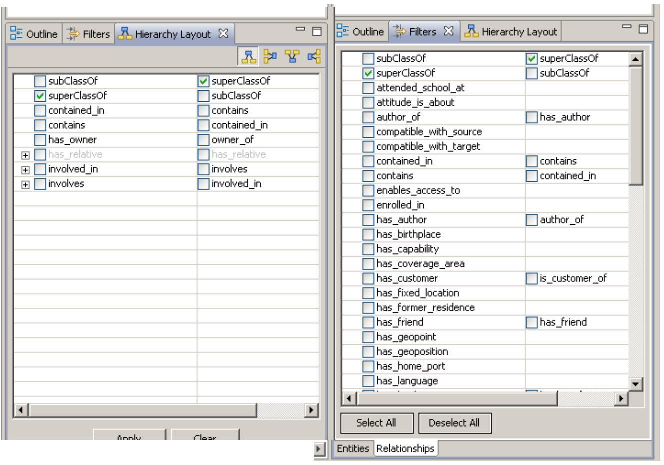
IV. Implementation Within The AKEA Tool

The methods proposed above are being implemented with the Analyst-Driven Knowledge Enhancement and Analysis (AKEA) tool at the Pacific Northwest National Laboratory. AKEA was created for clients within the IC as an environment for testing analyst interaction with semantically labeled data and for enabling automation-supported knowledge-level analysis over contents of structured and unstructured sources. While being ontology agnostic, AKEA depends on data representations which are ontologically



□ included_in □ in
□ includes_event □ in
□ involved_in ☑ in
☑ involves □ in
□ has_subevent □ s
□ has_time
☑ tinvolves_entitiy
☑ has_actor
☑ has_from_actor
☑ from_organizatior
□ has_giver
□ has_provider
□ has_seller
□ has_sender

Fig. 4. The AKEA tool showing a portion of an ontology used within the intelligence community. A portion of the "event" class hierarchy is linked to a portion of the "entity" hierarchy through the selected "from-organization" property.



⊞ □ subClassOf	☑ superClassOf
☑ superClassOf	□ subClassOf
□ contained_in	□ contains
□ contains	□ contained_in
□ has_owner	□ owner_of
⊞ □ has_relative	□ has_relative
⊞ □ involved_in	□ involves
⊞ □ involves	□ involved_in

□ subClassOf	☑ superClassOf
☑ superClassOf	□ subClassOf
□ attended_school_at	
□ attitude_is_about	
□ author_of	□ has_author
□ compatible_with_source	
□ compatible_with_target	
□ contained_in	□ contains
□ contains	□ contained_in
□ enables_access_to	
□ enrolled_in	
□ has_author	□ author_of
□ has_birthplace	
□ has_capability	
□ has_coverage_area	
□ has_customer	□ is_customer_of
□ has_fixed_location	
□ has_former_residence	
□ has_friend	□ has_friend
□ has_geopoint	
□ has_geoposition	
□ has_home_port	
□ has_language	

Fig. 5. Relation types for hierarchical layout and filtering.

backed in order to provide the variety of visualization and analytic capabilities offered.

For this effort we exploited and extended AKEA's capabilities to additionally support activities of the AKM. While many aspects of the AKM roles were already addressed, these capabilities needed to be more directly focused on the ontology itself rather than on instance data represented using the ontology.

The first step in this support was direct visualization of the ontology. Because of the complex nature of the classes and relationships typically described within an ontology, typical link-node layouts fail to communicate meaningfully. However, by integrating the visualization approached described above, layouts appropriate to understanding the conceptual and relational structures of the ontology begin to address this problem. Fig. 4 provides a snapshot of an ontology presented in the AKEA ontology viewer using the subsumption hierarchy to drive layout. Fig. 5, left side, shows the controls for selecting among transitive relationships to view other concept structure. At the right of Fig. 5 is the relationship filters used to de-clutter the display. Since the sheer number of relationships in most ontologies would obscure the concept structure, this allows the analyst to focus on only the specific relationships of interest at any given time to fully understand interactions between the concept structures and relationships.

Future work with AKEA will address additional activities of the AKM. Work is already underway to incorporate the structural characterization statistics of the ontology and of classes and relationships. However, the most important change will be the ability to address multiple ontologies. This will enable the visualization, analysis and creation of alignment mappings between ontologies for communication, documentation, and automated translation needs.

V. Conclusions

The advent of knowledge-based systems and supporting knowledge bases is augmenting and making more critical the role of the Analytic Knowledge Manager. While IC personnel already perform these activities, current organizational systems and structures lend themselves to a fractured and less than effective execution. By clearly articulating these activities, the roles and responsibilities involved, and the resultant support needs, the IC can begin to move toward better recognition of the importance and value of the AKM. Such recognition will help bring about the systemic changes necessary to take full value of ontologically-based system investments, make that value more widely available, and make these technologies more readily applicable to the dynamic problems encountered by the intelligence analyst.

Acknowledgment

This work funded by Battelle Memorial Institute under the Threat Anticipation Initiative.

References

1. S. Bonacorsi, *RACI Diagram/RASCI Matrix - A Complete Definition*, The Project Management Hut, 2008.
2. Davey, BA and Priestly, HA: (1990) *Introduction to Lattices and Order*, Cambridge UP, Cambridge UK, 2nd Edition
3. Euzenat, Jerome and Shvaiko, P: (2007) *Ontology Matching*, Springer-Verlag, Hiedelberg
4. Joslyn, Cliff; Donaldson, Alex; and Paulson, Patrick: (2008) "Evaluating the Structural Quality of Semantic Hierarchy Alignments", Int. Semantic Web Conf. (ISWC 08), http://dblp.uni-trier.de/db/conf/semweb/iswc2008p.html#JoslynDP08
5. Joslyn, Cliff; Paulson, Patrick; and Verspoor, KM: (2008) "Exploiting Term Relations for Semantic Hierarchy Construction", *Proc. Int. Conf. Semantic Computing (ICSC 08),* pp. 42-49, IEEE Computer Society, Los Alamitos CA
6. Joslyn, Cliff; Mniszewski, SM; Smith, SA; and Weber, PM: (2006) "SpindleViz: A Three Dimensional, Order Theoretical Visualization Environment for the Gene Ontology", *Joint BioLINK and 9th Bio-Ontologies Meeting (JBB 06)*, http://www.bio-ontologies.org.uk/2006/download/Joslyn2EtAlSpindleviz.pdf}
7. Noy, Natasha and Musan, Mark A: (2003) "The PROMPT Suite: Interactive Tools for Ontology Merging and Mapping", *Int. J. Human-Computer Studies*, v. 59, pp.983-1024
8. D.E. O'Leary, "Enterprise knowledge management," *Computer*, vol. 31, no. 3, 1998, pp. 54-61.
9. T Tudorache, N.F. Noy, S. Tu, M. A. Musen: (2008) "Collaborative Ontology Development in Protégé", *7th International Semantic Web Conference (ISWC 2008)*, Karlsruhe, Germany, Springer.
10. Verspoor, KM; Cohn, JD; Mniszewski, SM; and Joslyn, CA: (2006) "A Categorization Approach to Automated Ontological Function Annotation", *Protein Science*, v. 15, pp. 1544-1549