Computer-Supported Collaborative Knowledge Modeling in Ecology

Deana D. Pennington
University of New Mexico
MSC03 2020
Albuquerque, NM 87131
+1-505-277-2595
dpennington@LTERnet.edu

Joshua Madin
Univ. of California SB
735 State Street
Santa Barbara, CA 93101
+1-805-893-7108
madin@nceas.ucsb.edu

Ferdinando Villa
University of Vermont
617 Main St
Burlington, VT 05405
+1-802-656-2968
ferdinando.villa@uvm.edu

Ioannis N. Athanasiadis
Istituto Dalle Molle di Studi
sull'Intellienza Artificiale
Manno, Lugano, Switzerland
+41-586-666-671
ioannis@idsia.ch

ABSTRACT
We describe collaborative efforts between a knowledge representation team, a community of scientists, and scientific information managers in developing knowledge models for ecological and environmental sciences. Formal, structured approaches to knowledge representation used by the team (e.g., ontologies) can be informed by unstructured approaches to knowledge representation and semantic tagging already in use by the community. Observations about the process of collaboration between the team and the community are used to generate an interaction model for supporting software tools.

Categories and Subject Descriptors
H.5.3 [Information Interfaces and Presentation]: Group and Organization Interfaces – *Computer-supported cooperative work, Web-based interaction.*

General Terms
Design, Human Factors

Keywords
Collaboration, observation, ontologies, concept maps, ecological knowledge

1. INTRODUCTION
Understanding and solving global environmental problems requires a new kind of science: science that is interdisciplinary, collaborative, and responsive to the needs of decision-makers [9, 17, 29]. Cross-disciplinary networks of scientists worldwide are marshalling their understanding in efforts to provide scientific results that target complex problems. Formal networks of scientists—such as the Long Term Ecological Research (LTER) networks originally developed in the US (http://www.lternet.edu/) and now located worldwide (http://www.ilternet.edu/)—employ information managers whose primary task is to provide online access to relevant information. With available resources rapidly increasing, the difficulty of discovering and making use of those resources (e.g., knowledge synthesis) is increasing as well, especially in conjunction with rapid expansion of the Web as a whole. A number of efforts are underway to enable better sharing of data, information, and knowledge within the natural sciences, as discussed in [1, 16, 22, 26]. These efforts include ontology-driven applications that make use of formal semantic reasoning to enable integration of heterogeneous resources.

Ontology-based approaches require eliciting shared knowledge from large communities of domain scientists and decision makers. The authors are part of several large-scale initiatives that will use shared ontologies: the National Science Foundation-funded projects Science Environment for Ecological Knowledge (SEEK; http://seek.ecoinformatics.org) and Assessment and Research Infrastructure for Ecosystem Services (ARIES; http://ecoinformatics.uvm.edu/projects/the-aries-framework.html) that focus on automated integration of environmental and economic data with models and analytical pipelines; and the EU-funded SEAMLESS project (http://www.seamless-ip.org), aimed at generating integrated assessment tools to understand how future alternative agricultural and environmental polices affect sustainable development in Europe. In all these projects, the need to crystallize community knowledge into formal ontologies has emerged paramount. However, each of these projects has been confronted by the challenges identified by Grudin [10] specific to groupware development, particular the following two problems:

- Disparity in work and benefit. Scientists who have the knowledge that must be incorporated into ontologies lack understanding of the benefits that semantic modelling will ultimately provide them and are unwilling to engage in activities that do not provide clear, short-term benefits. Information managers who might be able to provide some of the knowledge and may even understand the long-term benefits for the scientists have more immediate problems and focus their time on developing short-term solutions. Hence, ontology development requires "additional work from individuals who do not perceive a direct benefit" [10].
- Critical mass and Prisoner's dilemma. Ontology-driven applications are expected to be most useful when multiple users share their resources. The work involved in ontology development and annotation of resources is not justified by a single user. Hence, these projects require a "critical mass of users to be useful" [10] and early adopters must commit to substantial effort with no guarantee that others will follow.

Grudin makes a number of relevant suggestions for addressing these problems [10]:

- Reducing the work required of non-beneficiaries and indirect beneficiaries.
- Design processes that create benefits for all group members.
- Build in incentives for use.

Developing an innovative approach to community-based ontology development that incorporates these suggestions presents an ill-defined, unstructured problem requiring creative thinking. Development of solutions to such problems can be conceived as

two-phased [27]: 1) an idea generation phase that requires a combination of divergent thinking and domain expertise, and 2) an implementation phase. In this paper, we focus on the idea generation phase, envisioning systems that could effectively link short-term user needs supported by informal semantics with longer-term formal ontology development. The ideas are based on our experiences working with these science communities, understanding of their tasks, and ongoing efforts at community-based ontology development. The goal of this paper is to propose innovative designs for systems that enable collaborative ontology development derived from our particular case, and also to stimulate vibrant debate and creative thinking about generic issues that confront interdisciplinary ontology development efforts.

We begin with a brief description of the participants. That is followed by a brief description of our ontology needs and an upper-level ontology that we have created. These sections provide context for understanding the kinds of knowledge that we need to elicit from the community and the resources that we have available to apply to the problem. Next, we present a set of use cases for supporting semantic-based work tasks that are commonly undertaken in our communities. We describe how these tasks provide an opportunity to capture knowledge relevant to formal ontology development while providing immediate benefits to the users. We provide a high-level conceptualization of a system that we are currently designing to implement these ideas. Then, we describe methods that we have already undertaken to extract knowledge from users in direct and indirect ways, without the support of enabling systems. These provide real examples of tasks that inform ontology development. We discuss how these could be incorporated into our hypothetical system in ways that limit the work required from the user. Lastly, we abstract our specific problems and proposed solutions into a simple model for enabling collaborative ontology development.

2. PARTICIPANTS

Initially, each project had its own Knowledge Representation (KR) research and personnel. Several years ago we began to collaborate with a view towards constructing ontologies that would interoperate between projects, providing an opportunity to leverage each others' work but also creating a larger, multi-disciplinary group that was more capable of critical evaluation of different proposed ontologies.

The KR team has cross-disciplinary expertise in computer science and domain science. It consists of two computer scientists with expertise in ontologies, reasoning, and semantic mediation, and four domain scientists with differing disciplinary expertise, relatively high levels of computing experience, and varying backgrounds in knowledge representation. The team has met regularly to devise strategies for ontology development. Discussion at these meetings ranges from formal symbolic logic to philosophy of science to targeted discussion about implicit knowledge embedded in datasets. Time and effort was required to bridge disciplinary boundaries and understand inherent assumptions that impact the teams' ability to collaborate on what is clearly an interdisciplinary task. Numerous real examples of environmental data and analyses obtained from scientists and information managers have guided and informed these discussions. One of the domain scientists is tasked with knowledge engineering, and is responsible for developing and maintaining the ontologies in Protégé (http://protege.stanford.edu/). Another is tasked with acting as liaison to the scientific community.

The KR team collaborates with the scientific and information management communities to elicit domain-specific knowledge. Few of the community collaborators have the time or interest to cultivate an understanding of formal ontologies. Nor do they fully understand the benefits of ontology-driven systems, since few examples of these systems exist. Hence, their personal commitment to ontology development is limited. Yet they recognize that semantic approaches may provide future benefits to them and are willing to help to the extent that it does not impede their more immediate objectives.

3. ONTOLOGY NEEDS

In each of our projects, KR is tightly integrated into technical research and development. We are working toward semi-automated and automated resource discovery and integration, including finding and merging heterogeneous datasets and construction of workflows that pipe data through heterogeneous computing environments [4, 5, 6]. We are also constructing knowledge-driven rule-based systems. These applications require high-quality ontologies and formal reasoning provided by description logics for consistency checking and validation. Much of the functionality provided by ontological reasoning will be hidden from the user, yet will automate many low-level tasks that the user would otherwise have to undertake manually.

Our ontology development has been two tiered: 1) development of an upper-level structuring framework for observation and measurements (core ontology), and 2) development of domain-specific extensions to the core ontology. Our early work was more focused on the first though the need for domain extensions was known and information was continually gathered from the community whenever possible. Recently, the core ontology has been finalized and is currently being documented [15].

Scientists make observations about the world that are recorded as measurements. The core ontology is the *Extensible Observation Ontology* (OBOE), which is a formal and generic conceptual framework for describing the semantics of observation and measurement. The objective of OBOE was to separate knowledge that is essential for describing observation and measurement from knowledge that is asserted by a scientist and therefore a function of opinion, interpretation, or even space and time. OBOE requires that an observation is about an *entity* (concept or thing), and a measurement is of a *characteristic* of the entity. Measurement relates a value to a *measurement standard* as well as an estimate about the confidence level of the value (e.g., measurement *precision*). OBOE prescribes a structured approach for organizing domain-specific ontologies through the use of "extension points," i.e., specific classes, properties, and constraints that are elaborated by different areas or views/models of science. Therefore, OBOE can serve as an upper level framework for defining new domain ontologies as well as interoperating and relating existing domain ontologies.

While OBOE enforces a formal framework for describing the semantics of observational data, extension of this framework with domain ontologies requires the knowledge and experience of domain scientists. The KR team is continually involved in outreach to acquire community-based vocabularies and informally-structured knowledge. These outreach activities

provide a flow of informally-structured semantic description among collaborators (Figure 1).

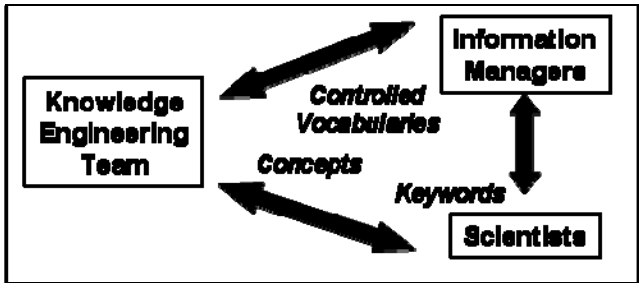

Figure 1. Collaborative relationships between the knowledge engineering team, scientists, and information managers, and the types of semantic information that are shared.

Ultimately the knowledge representation team must make some independent decisions about how best to model the domain within a formal ontology. This, therefore, necessitates at least one and perhaps many iterations of review by scientists. When the team has an ontology ready for review, we would like to recruit people from that domain to view it, comment, and propose changes. While trees may be used effectively to review the hierarchical structure, relationships are more difficult to communicate effectively. They do not understand symbolic logic commonly employed in editors. Usability testing of graphic visualizations conducted by the SEEK project indicates that they are confusing to domain users (Downey, personal communication). Additionally, community-wide ontological commitment [8] requires collective decision making, difficult to achieve without synchronous communication. Currently, there is no obvious mechanism by which to obtain the needed input from reviewers.

4. USER SEMANTIC TASKS AND COMPUTER-SUPPORTED USE CASES

There is a need for more collaboration between our KR team, scientists, and information managers. The complexity of ontologies and the difficulty of the knowledge modeling task presents a daunting obstacle to those who are not familiar with knowledge representation. We need tools that link knowledge elicitation with tasks in which the community is already engaged, and development of methods and tools that enable rapid mapping from those to formal ontologies.

There are many reasons to capture and represent knowledge in science, separate and apart from the resource discovery and integration goals of the Semantic Web. Smith [23] suggested that oftentimes philosophers turn to science as a reliable way to learn about the things and processes of a given domain. Much effort in science is focused on acquiring knowledge through scientific discourse. This begins during formal education but is ongoing throughout the life of a scientist, who must be able to share his own perspective and understand those of competing explanations. Those semantic perspectives are implicit in the artifacts of science: tools, models, datasets, and publications. Creation of these artifacts involves tasks that are inherently semantic and could both contribute to ontology development and be assisted by a knowledge base. Here we provide four use cases of some example tasks, knowledge-based computer support for those tasks, and a vision for interaction mechanisms between and among different stakeholders.

4.1 Controlled vocabulary use case

Karen Mann is an information manager for one of the LTER field sites. She and several of her colleagues at other field sites have decided to construct a standard set of terms and definitions to be used as metadata keywords, to enable better data discovery by scientists across the LTER network. She is aware of the observation ontologies that are being developed, but doesn't really understand them. She is reluctant to attempt to make use of an approach that she doesn't understand. She does understand that ontologies enable even better data discovery and integration than her approach. Therefore, she wants to work within the context of keywords and controlled vocabularies since that is what she understands, but she would also like to link her list of keywords to the ontology to take advantage of whatever additional functionality is made available.

Karen enters a website that provides an intuitive interface to a knowledge base that holds many ontologies, both private and shared. From this website she can create and manage her own private knowledge base. She imports a list of terms that she has previously generated. She can also import informal definitions (not constrained logical definitions), or she can enter the definitions on the website. Her colleagues import their lists into their own private knowledge base as well. They all indicate to the system that they want to share (or not) their private knowledge bases. Karen selects her colleagues' shared knowledge bases from a list, generates a collaborative knowledge base, and sends a message through the website asking them to collaborate with her. From a collaboration screen, they are able to merge their vocabulary lists into a single unfiltered list. The system maintains a link between their individual lists and the collective list, so that any changes made during collaboration can optionally be copied back to their individual knowledge bases. Their screens are linked. When one person selects or edits a term everyone else's screen automatically shows the change. They can make use of VoIP or a chat window to discuss their vocabularies. In this case, because there are a number of participants they prefer to use chat [14]. Their chat session is recorded and at the end of their discussion they can request for the chat session to be copied to a blog attached to the collaborative knowledge base, providing a permanent record.

They collaboratively review duplicate terms and definitions to determine semantic relationships. They identify synonyms and can drag and drop synonyms on the screen so that they are adjacent to one another. Where there are semantic conflicts they resolve them and edit the collective vocabulary.

Once they have a complete collective list of terms, they can choose an option to annotate the terms in their list with an ontology. A list of ontologies is provided to them, which includes a list of "Our Favorite Ontologies" that the system generates from each individual's list of "My Favorite Ontologies." They decide on the ontologies they want to use (all of which are extensions to the OBOE observation ontology), and begin to the annotation process. For each term, the system automatically shows them syntactically exact matches from their selected ontologies along with definitions. They can easily explore parent, sibling, and child concepts as well as other related concepts to ensure that they understand the context of any given concept in the ontology and to reconsider their term selection. They are able to search the knowledge base using a google-style interface to see what other concepts might be relevant. They can ask the system to analyze their searches and suggest concepts based on the choices by other

users who have made similar searches. If they are uncertain about whether a concept is appropriate, they can request several levels of help: tips and tricks, online documentation of annotation procedures, examples, live chat with a knowledge engineer, or e-mail support.

If they do not find a concept that fits, they can suggest terms to be added to the ontology. They recommend a concept and the system provides them with a wizard to capture their recommendations about where the concept belongs in the ontology. The system allows them to go ahead and use the term with a tentative annotation. Asynchronously, a knowledge engineer will consider where to place the term in the ontology. The system will provide him with information about the term from their knowledge base and from their search history; he may also request additional information from them. If he decides to add the concept as suggested, the system makes any needed adjustments to their knowledge base. If the concept is not added, the knowledge engineer can identify it as a synonym or make some other link from that term to the ontology such that the user can continue to use that term but the system can resolve it to the correct annotation. They will get automatic notification of the final decision made by the knowledge engineer. Task support for the knowledge engineer is further discussed in Section 4.4.

When Karen and her colleagues apply keywords to resources such as datasets or publications, they each apply terms from their individual controlled vocabulary. They can then select an option for automatic annotation that runs a script that constructs the correct ontological annotation. The metadata therefore includes keywords from the local vocabulary and annotation to one or more ontologies allowing the resources to be used with ontology-driven discovery and integration tools.

4.2 Data description use case

John Green is an ecologist with LTER who collects field data on plants. He has numerous spreadsheets with similar but slightly varying schemas that he has collected over a number of years. John is interested in contributing his data to a portal so that he can participate in a new collaborative project that will analyze plant species from around the globe. In order to do so, he must provide metadata that includes ontological annotation.

The LTER information managers have previously developed a web application that walks users through the process of creating metadata for datasets. Their knowledge base is accessed by this application, providing access to the site's controlled vocabulary linked to ontologies. His information manager has provided some training on how to make use of the application. John has never actually used the system, but has a vague recollection of how to do it and enters the website with confidence knowing that both the description and annotation tasks are supported with intuitive user interfaces online help for novices.

John creates metadata for the first dataset. He loads the dataset into the web application, which analyzes the dataset and is able to automatically generate a fair amount of metadata. The system prompts him for the remainder of the metadata. Then he must begin the semantic annotation process. He starts with the controlled vocabulary for his site. The system prompts him to select keywords for the dataset as a whole, then for each attribute in the dataset. Because the keywords are linked to an upper-level ontology, the system prompts him to annotate the relationships between attributes required by that ontology and guides him through that task. If John has an attribute that he does not think is adequately expressed by any of the terms in the controlled vocabulary, he has all of the same ontology exploration functionality available to the information managers. He can suggest terms to be added to the controlled vocabulary and/or to the ontology using the same procedure as the information manager. In this case, his recommendation is forwarded to the information manager who can assess the term, add it to the controlled vocabulary and link it to the ontology, or forward it to the knowledge engineer if it requires modification of the ontology.

Once the first dataset has been described and annotated, John has several datasets that used the same schema. He loads the second dataset and indicates to the system that it is a duplicate of the first in terms of physical, logical, and semantic description. The system analyzes both datasets using a metadata ontology and verifies that that seems to be the case. The system duplicates the metadata and annotations then prompts John for any edits that might need to be made. The system "knows" which parts of the metadata or annotations could possibly change because of the existence of the metadata ontology and leads him through those. If the datasets are not duplicates, the system will inform John where there are discrepancies and support him through the process of comparing datasets, resolving issues and generating correct metadata and annotations.

The remaining datasets are similar to the first dataset but vary in different ways. John loads a new dataset into the tool and indicates to the system that it is similar to the first dataset. The system compares table structures, data types, and column content and recognizes where there are differences. Again, the system knows where metadata and annotations could possibly change, and prompts John to enter the correct information.

John wants to generate a template dataset that is already described and annotated (to the extent possible) for future use. He can pick any of the datasets already described and annotated, and request a template. The system generates a blank table with associated metadata and annotations, then prompts for other information that is likely to be constant, such as project descriptions and personnel. John can elect to fill these in automatically from the original dataset or he can enter new information manually. Once the template is finished, he can save it and easily generate new datasets from it. Every time he does so, the system prompts him for information that is collection-specific.

Now that John has his datasets described and annotated, he contributes them to the portal, which is also tied to the knowledge base. He and a number of other scientists then begin to collaboratively decide which data should be integrated. They enter a web application that allows them to load up multiple datasets and collectively discuss them. As with the information managers, they can link their screens such that changes by one person automatically appear on everyone else's screen. They also have chat, blog, and VoIP options. As they discuss the datasets they are able to map between them semi-automatically using the knowledge base and attribute annotations. They can modify any of the mappings that the knowledge base suggests plus add new mappings. They can generate integrated datasets based on their mappings that inherit relevant metadata and annotations from the source datasets, prompting them to complete whatever new metadata or annotations are needed. As they collaboratively decide on the mappings between datasets, the knowledge base

tracks their decisions. For instance, the scientists decide that dataset 1 attribute 12 maps to dataset 2 attribute 6. These two attributes were annotated differently and there currently is no relationship between those concepts in the ontology. Through their collaborative mapping, however, they have indicated that there is indeed a relationship between these concepts. As they work through semi-automatic mapping of many attributes from many datasets the system is able to analyze their choices and suggest changes to the ontology to the knowledge engineer.

4.3 Concept mapping use case

Through the data portal, John has begun a dialogue with several scientists from different disciplines about potentially working together on a research project. Because they are familiar with different theories, research paradigms, and study methods, they need to spend a significant amount of time developing a conceptual framework that is well thought out and integrates their different perspectives. They are located in different universities and they can't take enough time away from their teaching to adequately develop a collaborative approach. They decide to make use of a new web application that provides collaborative concept mapping and is linked to the knowledge base.

They enter the website and rather than choose specific ontologies, they select the portal and request to use the same ontologies as the portal. Independently, they each draw concept maps and process flow diagrams that represent their research interests. Each term that they use, if present in the selected ontologies, is automatically completed as they type it in. Again, if they want to use a term that isn't in the ontology they can suggest terms. The linkages between terms in the diagram provide information about relationships between concepts that the system tracks, analyzes, and can use to suggest changes to the knowledge engineer.

Once they have each constructed their own diagrams they can collaboratively view and discuss each others work using linked screens, chat, blogs, and VoIP. They can draw diagrams together representing their collective views. As they discuss the diagrams they begin to resolve semantic issues. They determine that there is a close relationship between certain concepts in their different disciplines but they use different terminology for those concepts. As they find these differences they draw links on their diagrams. The system tracks these linkages and can use them to suggest links across domain-specific extensions of the ontology.

They can request the system to "show datasets," and next to each term on their maps it will provide titles of datasets in the portal that are annotated with that term or related terms. They can explore these datasets in the same collaborative way as described above, and construct integrated datasets. The portal is linked to a repository of publications that have been annotated. Therefore "show publications" can be used to display publications that have been annotated with the terms related to those they have used.

After drawing many diagrams, exploring datasets, and reading relevant publications they are ready to design their research project. They make use of a "workflow design" module that provides some structure for diagramming a conceptual scientific workflow using concepts from the knowledge base. Each node in the workflow represents a computational analysis or procedure [16]. Links between the nodes represent flow of output data from one component to input data for the next. They use terms from model and process ontologies, with the system using automatic word completion. They can indicate specific datasets from the portal that are to be input to the workflow. When they are satisfied with their workflow, they can export it as a beginning workflow for a scientific workflow system and the annotations are transferred with the workflow.

4.4 Ontology review use case

Bob Card is a knowledge engineer working with the LTER community. He works on a tightly-coupled team that includes both computer and domain scientists. Combining the teams' collective knowledge with information from text mining he has generated the knowledge base used in the above cases. He is rapidly receiving input from all of the suggestions made by his colleagues, as well as analysis of user actions from the system. He needs some sort of semantic management system to help him track all of these recommendations, make sense of them, and generate automated response to users who are affected by a given decision that he makes.

He is able to generate term lists from any combination of the above sources, flexibly sort and group terms, and try out tentative hierarchical structures before making any changes to his formal ontology. As he works with the tentative hierarchies he can invite participants to collaborate with him using linked screens. Or, he can request that colleagues review and modify a copy of any tentative hierarchy. The system will compare the modified copy with his tentative structure and show him where changes have been proposed. At any point he can modify the tentative ontology. When Bob is ready, he can request the system to align his tentative ontology with the existing ontology and show changes. When he is satisfied with the tentative ontology he can "commit" it and the system will automatically replace the affected portion of the existing ontology with the necessary changes. The earlier version is stored in case he needs to return to it. The system analyzes the changes and determines which annotated resources are affected. It creates a new version of annotations for those resources and notifies the user of the change.

5. EXAMPLE COLLABORATION-CENTERED SOLUTION

Our team has started investigating technical solutions to the challenge of defining user-friendly, semi-automated processes to distill disciplinary knowledge into formal ontologies. Our goal is to accomplish this with the least possible amount of difficulty for the user and transparent, non-obtrusive involvement of the knowledge base. The approach that we are taking is design of interacting systems for knowledge base development and management, community-based ontology interaction, and multiple knowledge-based applications (Figure 2.).

The ThinkCap Collaborative Knowledge Portal is a prototype web application still under development that provides user interfaces over a remote, multi-ontology knowledge base, designed to meet the needs of both non-technical and technical users (http://ecoinformatics.uvm.edu/technologies/thinkcap.html). It aims to allow remote users of diverse disciplines and technical levels to develop shared conceptualizations that are automatically formalized into OWL or RDFS ontologies.

The paradigm of knowledge elicitation being implemented in ThinkCap uses a knowledge engineer in an asynchronous way; by decoupling the formal knowledge base from the "arena" of user discussion full concurrency of the editing of both is made possible. The process is assisted by a full-text search engine that

indexes OWL concept descriptions as well as user-provided documentation (such as web pages or academic papers).

We are currently extending ThinkCap to help such a diverse community of users negotiate the rigorous, streamlined axioms in an OWL knowledge space. A new collaborative portal in ThinkCap will use a reasoner-assisted process and an upper ontology to define different views of an OWL knowledge base. These simplified views will allow applications to show only the level of semantic complexity necessary for the immediate task. Views will include conversion of ontologies to topic maps (www.topicmaps.org). Topic maps reflect the knowledge in the ontology base in ways that are much friendlier to the user community and much easier to operate on concurrently. The portal will provide a web-based whiteboard environment for collaborative topic map editing. A reasoner-assisted listener process will analyze user changes to the topic map and provide suggestions to a knowledge engineer about possible relevance to the underlying OWL axioms. Once a prototype has been tested with users, we will design additional interfaces.

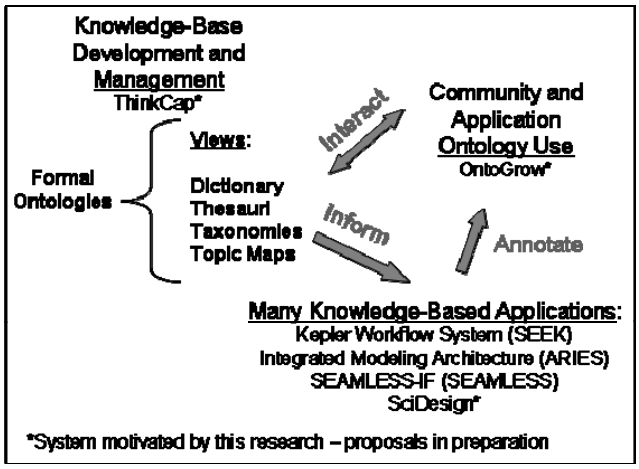

Figure 2. Interacting systems that make use and inform development of multiple ontologies.

OntoGrow is an interface to ThinkCap that is currently under design. OntoGrow will provide functionality for communities to interact with ThinkCap and can either be accessed directly or indirectly through an application add-in. OntoGrow has three objectives: 1) provide community feedback/critique of ontologies, 2) recommend a term for an ontology, and 3) map semantics between a resource and one or more ontologies. The multiple views of ThinkCap will allow OntoGrow to provide wizards that step a user through these processes in more intuitive ways. For instance, to recommend a new term, the user could first be asked to provide its definition through the dictionary view, find a related term with a thesauri search, place the term in a hierarchy by using a taxonomy to expose the context around the related terms that the user has selected, and relate the term on a topic map generated from the portion of the ontology that includes that hierarchical element. Thus, the user can be stepped through the task of ontology navigation by using their choices at each point to simplify the choices at the next level of complexity.

The applications that are currently being developed by our projects will each be able to make use of OntoGrow as an add-in or through remote calls, providing a uniform mechanism of interaction with the knowledge base. In addition, we are designing a new system, SciDesign, that is envisioned to provide the semantic-driven functionality described in our use cases. SciDesign will provide an interface for knowledge-based scientific discourse, resource discovery, exploration, and management, and research design. As scientists and information managers make use of SciDesign for individual or collaborative efforts, their actions will be captured and analyzed by the system and used to inform ontology development. Technical designs for SciDesign, OntoGrow, and ThinkCap are currently being developed under the second, implementation phase of complex problem solving that follows idea generation.

6. KNOWLEDGE ELICITATION-CENTERED PROCESSES AND SOLUTIONS

We present four approaches that our KR team has used to acquire scientific knowledge, beginning with the least demanding for the participants and ending with the most collaboration-intensive. Each is followed by suggestions for incorporation of these tasks into the proposed system.

6.1 Text mining

In science, the knowledge representation method of choice has historically been written texts (publications) or conference presentations with accompanying figures and tables. These approaches are highly expressive and have worked well for sharing scientific knowledge for generations. A wealth of information about scientific concepts is locked up in textbooks and publications. Effective mechanisms for mining these sources provide abundant information for ontology development with no additional effort on scientists' parts. The downside of this approach is that structure or presentation of knowledge within a text represents the perspective of one or a few scientists, and does not necessarily capture the perspective of the broader community. It may not provide a knowledge model for which there can be widespread ontological commitment [8]. Therefore, text mining approaches are dependent on extensive collaborative review of the results.

The knowledge representation team is exploring different ways of extracting knowledge from a popular ecological textbook [3] for use in the OBOE framework. The team is quantifying the strength of association among key ecological terms using various measures of proximity. For example, the term "population" is strongly associated with "individual" and also "community"; however, the association between "individual" and "community" is considerably weaker. Moreover, the proximity of different sets of prepositions and verbs to coupled ecological terms is being used as a mechanism to determine the most likely type of relationship between terms. For example, when "individual" and "population" are in close proximity, words like "in", "part" and "contain" are often also in close proximity suggesting a part-of relationship between these terms. The team is also using book chapter, section, and subsection headings to help structure the nested ecological terms, which helps distill broader concepts in the textbook domain (e.g., "competition" or "ecosystem").

There are many mechanisms for incorporating text mining into the hypothetical system. This functionality could be provided to knowledge engineers within ThinkCap. Text mining could be integrated into SciDesign as an aid for scientific literature search and review. Substantial time is dedicated by scientists to following the literature in their own discipline. Increasingly the boundaries between disciplines must be crossed and scientists

must search for relevant literature in disciplines that are less well known to them. Visual analytics is a new approach that mines semantic content across many potential resources and provides tools for visual content analysis [25]. Linking visual analytics with text mining would provide scientists with functionality to more easily, effectively, and comprehensively conduct literature searches. Providing computer support that enables this task would create an environment where it is to the scientist's benefit to use the system while providing valuable semantic information for ontology development. In a given literature search, selection of multiple resources from different disciplines, journals, websites, and other online sources provides evidence that these content sources are semantically related in some way. Combining source-specific semantic keywords with the choices and actions of many scientists equates to other forms of social tagging prevalent in Web 2.0. The system should be equipped to analyze these choices, mine the relevant texts, and both suggest other literature that might be relevant to the scientist and in parallel, propose terms and relationships to the knowledge engineer.

6.2 Keywords and controlled vocabularies

Scientist's regularly apply keywords to textbooks, publications, and datasets. Traditionally these are uncontrolled, though controlled vocabularies are becoming more common (i.e. for computer science publications IEEE and ACM share a definite tree-structured list of terms). Additionally, the titles they choose provide information about important terms. Mining titles and keywords for concepts and relationships provides a pathway for acquiring views on scientific knowledge that requires little effort from scientists, but does require collaboration with information managers who know how to access these on their systems.

Separate from our projects, LTER information managers conducted a mining project on network datasets and publications in order to develop a controlled vocabulary [21]. A list was generated by compiling all words appearing in metadata titles, keywords, and attributes, and in publication titles and keywords. The resultant list contained 21,153 terms. The list was filtered for 'of,' 'the,' and similar definite articles and prepositions. Terms were then rated in importance based on a number of usage criteria. The information managers are continuing to work with this list to develop a controlled vocabulary for use in tagging datasets and publications. They provided this list to our KR team, who were able to incorporate these terms into ontology development. The intention of both groups is to ultimately link the information managers' controlled vocabularies to the ontology such that controlled keywords applied to any resource are automatically annotated to the ontology, the ontology can be used to suggest terms that are not available in the controlled vocabulary, and the process of users applying new keywords can inform continued development of both the controlled vocabulary and the ontology.

In the proposed system, support for information management activities could be embedded in SciDesign. One of the above use cases explicitly addresses supporting construction and management of local vocabularies. There are many other information management activities that could be supported. Sometimes these activities are conducted by information managers, but there are many scientists who work independently and must conduct these activities themselves. Even when information managers are employed, they must work closely with scientists. Design of functionality to assist information management needs can be leveraged to support the activities of the scientists.

For instance, information managers collectively invest much effort in designing databases, developing normalized schemas, standardizing keywords, and developing standards for metadata. They have their own knowledge arena that combines both generic data management concepts and how those concepts are best applied to a particular domain of interest. Separate ontologies should be constructed to capture this knowledge. Rule-bases could be constructed that link to those ontologies and can be used to guide data management efforts. For instance, in designing a new table for collection of a particular kind of field data, the system could use an ontology and rules about database design to provide expert advice and best practices, mine available data to find and show examples of datasets that meet those guidelines and are semantically equivalent to the data the scientist intends to collect, and suggest one or more table designs.

6.3 Concept mapping

Concept mapping is an approach that the KR team has used that provides direct input for ontology development from a number of scientists while they are engaged in an activity that is useful to them. Concept mapping is a representation mechanism that has been developed to support a constructivist notion of learning [18]. Concept maps are a form of directed graph that captures associations (links) between concepts (nodes; Figure 3). Concept mapping provides maximum flexibility for conceptualization of a domain of interest, and any kind of association can be mapped. From a collaborative perspective, concept maps provide visual representation of disparate conceptual frameworks including the most important terms from a particular view, and places those terms in context with one another for rapid understanding.

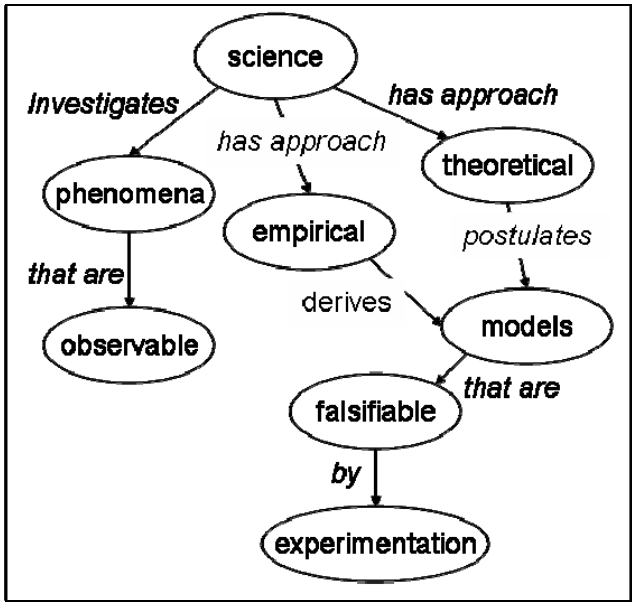

Figure 3. Example concept map showing relationships between terms related to the scientific method.

The utility of concept maps as a mechanism for enabling interdisciplinary discussion has been demonstrated [11, 13]. In cross-disciplinary problem solving efforts, colleagues with differing conceptual frameworks often have limited ability to comprehend each other [7, 28, 13]. The degree to which comprehension is limited depends on the conceptual proximity of

relevant conceptual frameworks - hence, two physical scientists are more readily able to collaborate than a physical and social scientist, or a life scientist and a computer scientist. Enabling cross-disciplinary collaboration is therefore a problem of representing disciplinary concepts in a way that enables rapid comprehension and learning by those outside of that discipline such that integrative problems can be solved.

The process of concept mapping is analogous in many ways to social tagging systems. The content, in this case, is an unrepresented concept in the mind of the scientist. A node in a concept map represents that concept. Two scientists may use different terms in the node that describes that concept, essentially tagging that concept differently. Links between nodes specify that a relationship of some sort exists between those concepts. This is roughly equivalent to inferring implicit semantic links between Web content. Two scientists drawing concept maps about the same research area will each have their own map using the same or different terms and relationships, but they are tagging the same semantic content. During scientific discourse, these disparate concept spaces may become partially aligned. Hence, concept maps from multiple scientists build a participatory ecosystem of content that can provide important vocabulary, indicate synonyms, show informal associations between terms, and provide hierarchical relationships. These semantic tags require structuring by the KR team and subsequent review and editing for clearance, cohesion, and soundness. However, the benefit of using concept maps is that it engages the scientific community in supplying knowledge for ontology development in a way that has other direct and immediate benefits to them, such that they are more likely to participate.

In the proposed system, concept maps and other diagrammatic forms are expected to be an important part of SciDesign. Scientists draw many sorts of diagrams and frequently find that mode of expression useful while discussing complicated cross-disciplinary subjects. Process diagrams, flow diagrams, project diagrams – there are an unlimited number of uses of diagrams. The system should provide flexible, intuitive diagramming tools that can be collaboratively constructed and shared, plus easily extracted and converted to publication-quality diagrams. If the nodes on the diagrams are linked to ontologies they can provide an individual "view" of the knowledge base, allowing each scientist to maintain his own conceptual perspective without compromising the collective formal structure. We have found that it is important to the scientists to be able to express their individual view with no constraints, and that the underlying subsumption hierarchy is much less important to them [20]. Science is, after all, about investigating areas of our understanding where there is not agreement, and understanding linkages across hierarchies rather than within hierarchies. Much of our analysis involves providing mechanisms for online and collaborative construction of concept maps and other scientific diagrams that facilitate working with different 'views' of a set of ontologies based on individual perspectives and choices about representation.

6.4 Meeting with scientists

The utility of ontologies has been introduced to scores of ecologists during a week-long training workshop on ecoinformatics that the SEEK project holds each January. The participants in this training are 20 new faculty and postdoctoral associates selected from on average 60-80 applicants from around the US. The selected participants represent the most technically-savvy of young ecologists tackling problems that require computational approaches. During the workshop, one full day is spent covering ontologies. Over the four years that the training has been conducted, the ontology portion has been constantly modified based on feedback from students, and many new approaches have been tried. In general, the students are exposed to exercises that highlight the semantic issues in ecological datasets and the requirements for resolving those issues. They construct ontologies for their research interests on paper. We demonstrate ontology editors and touch graph visualizations. They step through portions of ontology editing exercises such as CO-ODE's pizza ontology [12]. The ontology portion of the training is always the most difficult to present, and often receives criticism in post-training surveys. Even though participants understand the semantic issues and recognize that ontologies might be useful for addressing them, they do not think that it is important for them to understand ontologies. In the most recent training (January 2007) survey feedback indicated that 50 percent of participants, when asked what one thing they would change about the training, thought the ontology portion should be removed. This is a clear indication that direct exercises with ontologies is an obscure task for ecological scientists and more gentle tools are needed for communicating semantic models.

The KR team has attempted to engage groups of scientists in ontology development through working meetings where they are asked to talk about their research, explain terms, brainstorm hierarchies, and provide lists of terms. Generally, their level of interest in these activities fades rather rapidly, mirroring the response from the training activities. Additionally, the hierarchical structures that they propose are often unusable in our ontologies due to logical errors. Most importantly, those who are willing to participate are typically new faculty who are under substantial pressure to produce research results quickly in order to obtain tenure. Their modus operandi is to only get involved in activities that will quickly lead to publication. Few obtain any short-term professional benefit for assisting in the development of ontologies; hence, few can remain engaged at the level needed.

Given all of these issues, the KR team has to be creative about finding other ways to obtain their input. The hypothetical system as a whole represents a new approach to "meeting with the scientists." This new approach is virtual rather than physical, and focuses on linking user-centered task support with knowledge development task needs. It combines "pulling" ontology development through analysis of the way semantics are used by the community with "pushing" ontology development with easy mechanisms for reviewing and suggesting changes during task performance. It is an attempt to solve the problems of disparity of work and benefit, critical mass, and Prisoner's dilemma [10] that are prevalent in collaborative ontology development projects. It does that by bridging the gap between formal and informal semantic approaches in ways that reduces workload and provide benefits for all participants.

7. COLLABORATIVE ONTOLOGY DEVELOPMENT MODEL

Developing semantic systems that depend on and enable group sharing of resources differ in fundamental ways from developing software that supports individuals and large organizations [10]. One clear difference is that in both of the latter, the tasks to be supported are well-defined in advance by product managers or in-

house IT experts, respectively. In contrast, semantic tasks may be understood for the work of the KR team but are poorly defined for any new community that is to be supported. For instance, much work has been conducted on semantic tasks of online shoppers and therefore systems that support and make use of these activities are becoming common place. Those tasks are not necessarily analogous in any way to the semantic tasks of a completely different group such as scientists. The semantic tasks must be understood before they can be supported. A second difference is that the introduction of systems that drastically change work patterns require corresponding investments in dealing with social and political factors that go along with change management. These issues are largely absent in development of single-user software. They are strongly present in organizational settings where there is also an infrastructure in place to provide training, restructure work, and provide leadership. Our semantic systems for scientists bring about all of the challenges of changing work processes with little of the supporting infrastructure. This is a common reason for failure of new groupware solutions. For these reasons and many others it is essential that collaborative knowledge development teams become strategic in their activities. Unfortunately, there are few models available to guide strategic choices.

We propose the following model for development of semantic systems that depend on collaboration between knowledge representation specialists and the communities that they aspire to support. System development should be explicitly divided into two phases: an idea generation phase and an implementation phase (Figure 4). The idea generation phase can be conceived of as product development on steroids. It is separated out to emphasize that this is a lengthy, time-consuming process that may require as much resource investment as the implementation phase.

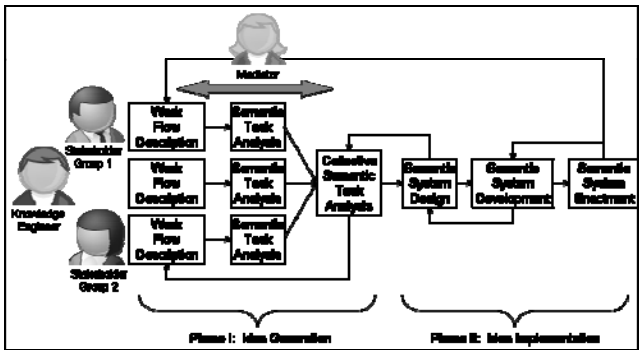

Figure 4. Model of collaboration on semantic systems. The idea generation phase is made explicit and involves needs analysis across the collective stakeholders rather than a single user group. Every step in the model is iterative and involves feedback from other steps.

Idea generation is an iterative process that has the goal of discovering linkages between semantic tasks of the collective group of participants that can be leveraged by system design. In its simplest form, it consists of learning about the workflow of each participating stakeholder group, analyzing those in terms of semantic tasks, then analyzing the collective set for tasks that can be linked in some way. In practice this involves a rather chaotic period of interaction between different participants and the KR team as they learn about each other's perspectives and search for common ground. These interactions are difficult because they depend on overcoming the very semantic barriers that semantic systems target. Developing cross-disciplinary understanding is the first step towards the truly interdisciplinary perspective that is required for effective idea generation. While there are few theories about enabling interdisciplinary interaction, social science research on boundaries, boundary crossing, and boundary spanners point to the importance of constructing shared artifacts, facilitated by an individual whose is explicitly tasked with mediating between the groups [24, 13, 30, 2]. The role of a mediator in any sort of groupware development is currently unspecified but in the semantic system case, could include soft system analysis of the KR team, domain specialists, and the broader community.

8. CONCLUSIONS

This paper describes interactions that have taken place between a knowledge representation team, natural scientists, and information managers, and uses those to drive a set of use cases for design of systems that enable better collaboration on ontology development. Previous interactions have been stymied by the lack of community understanding of ontologies and willingness to dedicate time towards ontology development. These problems reflect the lack of direct, immediate benefit for the participant. Our experience leads us to believe that formal ontology development could be more effectively informed by constructing tools that capture semantic decisions that are made in the course of the community's everyday work. Our community of interest regularly semantically tags the artifacts used in the conduct of science – datasets, publications, and models, and makes use of them in ways that capture semantic linkages. Design and development of systems that capture these semantic decisions and effectively make use of them to inform ontology development has been initiated but is in its infancy. Ultimately, we hope to have prototype systems and showcase applications that use those systems to demonstrate the collective benefits of ontology-based systems and applications.

The ideas that are generated through this process are not a complete set. They represent one or a few of many possible integrated approaches to linking semantic tasks. As the ideas are implemented and enacted within the broader community, other ideas will emerge. It is extremely important that any strategy taken explicitly account for feedbacks throughout the entire process including providing mechanisms to incorporate the views of the broader community in long-term system development.

9. ACKNOWLEDGMENTS

This work was funded through National Science Foundations grant 0225665 for the SEEK project, grant DBI 0640837 for the ARIES project, and European Union grant 010036-2 for SEAMLESS. We would like to recognize the many relevant discussions with the rest of the SEEK and ARIES teams, along with valuable comments by anonymous reviewers that led to restructuring of this paper and considerable sharpening of content.

10. REFERENCES

[1] Athanasiadis, IN (2007). Towards a virtual enterprise architecture for the environmental sector, In: (Protogeros, N, Ed.) *Agent and Web Service Technologies in Virtual Enterprises*. Idea Group Inc.

[2] Baker, KS, Jackson, SJ, and Wanetick, JR (2005). Strategies supporting heterogeneous data and interdisciplinary collaboration: Towards an ocean informatics environment,

Proceedings of the 38th Hawaii International Conference on system Sciences.

[3] Begon, M, Townsend, C, and Harper, JL (2006). *Ecology*, Blackwell Publishing, 752 pp.

[4] Berkley, C, Bowers, S, Jones, M, Ludaescher, B, Schildhauer, M, and Tao, J (2005). Incorporating semantics in scientific workflow authoring, *Proceedings of the Statistical and Scientific Database Management (SSDBM) 2005*.

[5] Bowers, S, and Ludaescher, B (2004). An ontology driven framework for data transformation in scientific workflows, *Proceedings of Data Integration for Life Sciences (DILS) 2004*.

[6] Bowers, S, Thau, D, Williams, R, and Ludaescher, B (2004). Data procurement for enabling scientific workflows: On exploring inter-and parastism, *Proceedings of Semantic Web and Databases (SWDB) 2004*.

[7] Daily, GC and Ehrlich, PR (1999). Managing earth's ecosystems: an interdisciplinary challenge, *Ecosystems* 2:277-280.

[8] Davis, R, Shrobe, H, and Szolovits, P (1993). What is a knowledge representation? *AI Magazine* 14(1):17-33.

[9] DiCastri, F (2000). Ecology in a context of economic globalization, *BioScience* 50(4):321-332.

[10] Grudin, J (1994). Groupware and social dynamics: eight challenges for developers, *Communications of the ACM* 37(1): 92-105.

[11] Heemskerk, M, Wilson, K, and Pavao-Zuckerman, M (2003). Conceptual models as tools for communication across disciplines, *Conservation Ecology* 7(3):8-17.

[12] Horridge, H, Knublauch, H, Rector, A, Stevens, R, and Wroe, C (2004). *A Practical Guide To Building OWL Ontologies Using the Protégé-OWL Plugin and CO-ODE Tools, Edition 1.0*. Cooperative Ontologies Program tutorial, 118 pp. Available at http://www.co-ode.org/resources/tutorials/ProtegeOWLTutorial.pdf.

[13] Jeffrey, P (2003). Smoothing the waters: observations on the process of cross-disciplinary research collaboration, *Social Studies of Science* 33(4):539-562.

[14] Löber, A, Schwabe, G, Grimm, S (2007). Audio vs. chat: The effects of group size on media choice. *Proceedings of the 40th HICCS Hawaii International Conference on System Sciences*.

[15] Madin, J, Bowers, S, Schildhauer, M, Krivov, S, Pennington, D, and Villa, F (in review). An ontology for describing and synthesizing ecological observation data. Submitted to International Journal of Ecological Informatics.

[16] Michener, WK, Beach, JH, Jones, M.B, Ludaescher, B, Pennington, DD, Pereira, RS, Rajasekar, A, and Schildhauer, M, (2007). A knowledge environment for the biodiversity and ecological sciences. *Journal of Intelligent Information Systems* DOI 10.1007/s10844-006-0034-8 available online at url: http://www.springerlink.com/content/e252n818242783g4/.

[17] Newell, B, Crumley, CL, Hassan, N, Lambin, EF, Pahl-Wostl, C, Underdal, A, Wasson, R (2005). A conceptual template for integrative human-environment research, *Global Environmental Change* 15:299-307.

[18] Novak, JD, and Wurst, M (2005). Collaborative knowledge visualization for cross-community learning, In: (Tergan, S and Keller, T Eds.) *Knowledge and Information Visualization, Lecture Notes in Computer Science* 3426:95-116, Berlin Heidelberg: Springer-Verlag.

[19] Noy, NF, Sintek, M, Decker, S, Crubezy, M, Fergerson, RW, and Musen, MA (2001). Creating semantic web content with Protégé-2000, *Intelligent Systems* 16(2):60-71.

[20] Pennington, D (2006). Representing the dimensions of an ecological niche. *Proceedings 5th International Semantic Web Conference (ISWC'06) Workshop: Terra Cognita 2006 – Directions to the Geospatial Semantic Web*, November 6, 2006, Athens, Georgia. Available online: http://www.ordnancesurvey.co.uk/oswebsite/partnerships/research/research/terracognita.html.

[21] Porter, J (2006). Improving data queries through use of a controlled vocabulary, *DataBits: An Electronic Newsletter for Information Managers*, Spring 2006. Available online: http://intranet.lternet.edu/archives/documents/Newsletters/DataBits/06spring/.

[22] Rizzoli, AE, Donatelli, M, Athanasiadis, IN, Villa, F, and Huber, D (accepted). Semantic links in integrated modeling frameworks, *Mathematics and Computers in Simulation*.

[23] Smith, B (2003). Ontology: An introduction. In: (Floridi, L ed.), *Blackwell Guide to the Philosophy of Computing and Information*. Oxford:Blackwell, pp. 155-166.

[24] Star, SL (1990). The structure of ill-structured solutions: boundary objects and heterogeneous distributed problem solving. In: (L. Gasser and EMN Huhns, Eds.) *Distributed Artificial Intelligence, Vol. 2*. London: Morgan Kaufmann Publishers, Inc., pp. 35-54.

[25] Thomas, JJ and Cook, KA (2006). A visual analytics agenda, *IEEE Computer Graphics and Applications* 26(1):10-13.

[26] Villa, F, and Athanasiadis, IN (submitted). Modelling with knowledge: Emerging semantic approaches to ecological modeling, *Ecological Modelling*.

[27] Vincent, AS, Decker, BP, and Mumford, MD (2002). Divergent thinking, intelligence, and expertise: A test of alternative models, *Creativity Research Journal* 14(2):163-178.

[28] Wear, DN (1999). Challenges to interdisciplinary discourse, *Ecosystems* 2:299-301.

[29] Welp, MA, de la Vega-Leinert, A, Stoll-Kleemann, S, and Jaeger, CC (2006). Science-based stakeholder dialogues: Theories and tools, *Global Environmental Change* 16:170-181.

[30] Williams, P (2002). The competent boundary spanner, *Public Administration* 80(1):103-124.