Indirectly driven knowledge modelling in ecology

Deana D. Pennington*

Department of Biology,
University of New Mexico,
MSC03 2020, Albuquerque,
New Mexico 87131-0001, USA
E-mail: dpennington@LTERnet.edu
*Corresponding author

Ioannis N. Athanasiadis

Istituto Dalle Molle di Studi,
sull'Intellienza Artificiale,
Manno, Lugano, Switzerland
E-mail: ioannis@idsia.ch

Shawn Bowers

University of California at Davis Genome Center,
451 Health Sciences Drive,
Davis, California 95616-8816, USA
E-mail: sbowers@ucdavis.edu

Serguei Krivov

University of Vermont,
617 Main St., Burlington,
Vermont 05405, USA,
E-mail: serguei.krivov@uvm.edu

Joshua Madin

Department of Biological Sciences,
Macquarie University,
N.S.W. 2109, Australia
E-mail: jmadin@bio.mq.edu.au

Mark Schildhauer

National Center for Ecological Analysis and Synthesis,
University of California at Santa Barbara,
735 State Street, Suite 300,
Santa Barbara, California 93101, USA
E-mail: schild@nceas.ucsb.edu

Ferdinando Villa

University of Vermont,
617 Main St., Burlington, Vermont 05405, USA
E-mail: ferdinando.villa@uvm.edu

Abstract: We describe collaborative efforts among a group of Knowledge Representation (KR) experts, domain scientists, and scientific information managers in developing knowledge models for ecological and environmental concepts. The development of formal, structured approaches to KR used by the group (i.e., ontologies) can be informed by evidence marshalled from unstructured approaches to KR and semantic tagging already in use by the community.

Keywords: indirectly-driven knowledge modelling; ecological knowledge models; ecological metadata; scientific observation; ontologies.

Reference to this paper should be made as follows: Pennington, D.D., Athanasiadis, I.N., Bowers, S., Krivov, S., Madin, J., Schildhauer, M. and Villa, F. (2008) 'Indirectly driven knowledge modelling in ecology', *Int. J. Metadata, Semantics, and Ontologies*, Vol. 3, No. 3, pp.210–225.

Biographical notes: D.D. Pennington received her PhD in Geography from Oregon State University in 2002. She is currently research faculty at University of New Mexico, studying knowledge ecosystems, conceptual spaces, and models of cross-disciplinary collaboration.

I.N. Athanasiadis is a researcher with the Dalle Molle Institute for Artificial Intelligence, in Lugano, Switzerland, since 2005. He holds a Diploma (MSc) and PhD in Electrical and Computer Engineering, both from the Aristotle University of Thessaloniki, Greece. His research interests include software engineering for ecoinformatics, ontologies and the semantic web, intelligent systems and software agents, agent-based social modelling and simulation, decision support systems and machine learning, and data and knowledge engineering

S. Bowers is a Computer Scientist at the UC Davis Genome Center and a member of the Data and Knowledge Systems Lab where he conducts research in conceptual data modelling, data integration, and scientific workflows. He is also an active member of the Kepler Scientific Workflow project, where he has contributed to the design and development of Kepler extensions for managing complex scientific data, capturing and exploring data provenance, and ontology-based approaches for organising and discovering workflow components. He holds a PhD and a MSc in Computer Science from the OGI School of Science and Engineering at OHSU, and a BSc in Computer and Information Science from the University of Oregon.

S. Krivov obtained his PhD from the Intercultural Open University, The Netherlands. He has been working in the field of ecoinformatics with a focus on KR technologies. Currently, he is developing a Bayesian network package for probabilistic modelling of ecological systems and environmental decision making.

J. Madin is a Computational Ecologist in the Department of Biological Sciences at Macquarie University in Sydney, Australia. His PhD is in Coral Reef Ecology and Biomechanics, from James Cook University in Australia. His research interests include statistical modelling of uncertainty, large-scale and long-term ecological trends, and techniques for finding, integrating and storing data.

M. Schildhauer is Director of Computing at the National Center for Ecological Analysis and Synthesis. He has a PhD in marine and behavioural ecology from the University of California, Santa Barbara. His interests are in scientific computing, particularly ecoinformatics and the challenges of interpreting and analysing ecological data.

F. Villa has a PhD in Ecology from the University of Parma. He is currently at the Gund Institute for Ecological Economics at the University of Vermont, where he leads the ARIES project (Artifical Intelligence for Ecosystem Services). His interests include interdisciplinary research at the interface of policy, ecology and economics, artificial intelligence approaches to environmental decision making, and natural system assessment and valuation.

1 Introduction

Understanding and solving global environmental problems requires a new kind of science: one that is interdisciplinary, collaborative, and responsive to the needs of decision-makers (Dicastri, 2000; Newell et al., 2005; Welp et al., 2006). Cross-disciplinary networks of scientists aim to integrate their understanding to provide scientific results that target complex problems worldwide. These 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 information rapidly increasing, the difficulty of discovering and making use of those resources 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 ecology and related disciplines, as discussed in Athanasiadis (2007), Michener et al. (2007), Rizzoli et al. (in press), and Villa et al. (in press). These efforts all include ontology-driven middleware applications that make use of formal semantic reasoning to enable integration of heterogeneous resources (Madin et al., 2008).

Ontology-based approaches often rely on eliciting shared knowledge from large communities of domain scientists and decision makers, and formally representing that knowledge within structured semantic frameworks such that automated mechanisms may be used to support

discovery and integration of semantically-related resources. The authors are part of several large-scale initiatives that are building such shared ontologies for information discovery and integration: 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) focus on automated integration of environmental, economic, and policy data with models and analytical pipelines; and the EU-funded System for Environmental and Agricultural Modelling project (SEAMLESS; http://www.seamless-ip.org), aimed at generating integrated assessment tools to understand how future alternative agricultural and environmental policies affect sustainable development in Europe. In each of these projects the need to crystallise community knowledge into formal ontologies is critical. However, each of these projects has confronted three challenges:

Lack of similar showcase applications. The complexity and diversity of concerns within ecology make it difficult to develop semantics that support a single, far-reaching 'success story'. As a case study, the Gene Ontology (GO: Ashburner et al., 2000) has probably been the most successful community ontology effort to date. The GO focuses on a circumscribed set of issues – the cellular location, molecular function, and biological processes associated with genes. The simplicity of the conceptual realm addressed by the GO is well shown by the fact that despite listing several thousand concepts, it connects them with only a few structural relationships: is-a, part-of, regulates, positively-regulates and negatively-regulates. Compared to the GO field of application, the complexity of even the simplest useful description of entities, processes and interactions associated with ecology is overwhelming. First of all, there are many complementary ways to conceptualise ecological systems, focusing for example on individuals, populations, communities, information or energy flow. This multiplicity of scales applies to the spatio-temporal realm as well: observed ecological phenomena are different if observations are made at smaller or larger spatial scales and over short or large time horizons. According to the scale of the observation, processes can be seen as entities and the other way around: for example, populations can be identified as 'things' unless their composing individuals also are, which makes the notion of a population a changing process. Populations in turn can be conceptualised as composing communities through their interactions (e.g., the food web), creating more ambiguity of description. Ecology deals with abiotic components (structural elements of the environment) as well as with biotic elements. Roles of entities are multiple and change according to the focal observation: e.g., living entities typically serve as the structural environment of others (trees for insects, insects for bacteria). As a result, a successful ecological ontology is likely to require the incorporation of notions of endurance and perdurance (Masolo et al., 2003), will need to define notions of hierarchical composition, and ultimately will never match the crisp simplicity of the GO, where all phenomena happen at the same scales within a well-understood set of processes and entities. Given all of these issues, ontology management is critical yet there is no centralised curation mechanism as with GO.

Disparity in work and benefit (Grudin, 1994). Scientists possessing the knowledge that must be captured in ontologies typically lack insight into the benefits that semantic modelling will ultimately provide them, in part because of the lack of a referential success story. Consequently, they are unwilling to engage in activities that do not provide clear, short-term benefits. Information managers, on the other hand, might have a better understanding of the long-term benefits of advanced knowledge modelling, but are also often occupied with more immediate problems and development of short-term solutions. Hence, ontology development requires "additional work from individuals who do not perceive a direct benefit" (Grudin, 1994).

Critical mass (Grudin, 1994). Ontology-driven applications can be useful for individuals, but they are far more useful when groups share their resources, requiring a "critical mass of users" (Grudin, 1994). Given the amount of initial work necessary, early adopters of semantically-enabled technologies must contribute substantial effort with no guarantee that others will follow. Grudin makes a number of relevant suggestions for addressing these problems:

- reduce the work required of non-beneficiaries and indirect beneficiaries
- design processes that create benefits for all group members
- build in incentives for use.

In this paper, we explore a new approach to ontology construction which we feel can address these challenges, by vastly simplifying the process of developing formal Knowledge Representations (KRs) for ecology. We call this approach 'indirectly-driven' knowledge modelling. The goal of the indirectly-driven approach is to augment elicitations of knowledge models through direct interactions with users with indirect, systematic gathering of scientists' semantic usages from their daily work, as reflected in books, journals, research design, and other communications. We believe that semantic patterns discernible in these constitute strong evidence for the underlying knowledge models that inform domain discourse. Additionally, we show that evidence-gathering systems can support scientists' ability to perform such work while also facilitating knowledge modelling activities, providing an incentive and a benefit for all group members. Lastly, because these activities and supporting systems build on participants' ongoing work it is easier to construct a compelling immediate vision of the usefulness of knowledge modelling that all participants can understand and anticipate, while continuing on the path to

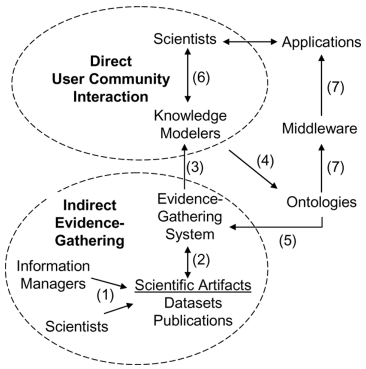
the longer-term more comprehensive solutions that we envision.

We begin with a description and investigation of the indirectly-driven knowledge modelling approach. Then, we provide a description of our experiences in knowledge modelling within the SEEK project: the participants, user communities, an example of our ontology needs, and the challenges we encountered. These sections provide an example and context for understanding the kinds and complexity of knowledge that we need to model. Next, we present a set of scenarios for semantic-based work tasks commonly undertaken by scientists and how evidence-gathering systems could support those. These scenarios and the systems supporting them are hypothetical, though we base them on our own use cases, including real examples drawn from our experiences. The objective is to generate an initial vision for evidence-gathering systems that can motivate creative thinking and debate about generic issues that confront interdisciplinary ontology development efforts.

2 Indirectly-driven knowledge modelling

Indirectly-driven knowledge modelling is an augmentation of traditional approaches to knowledge modelling rather than a replacement approach (Figure 1). Traditional approaches are usually direct: knowledge modellers engage their user community through personal interactions that range from structured meetings where each term is carefully considered to loosely structured arenas for the community to devise and edit basic conceptual frameworks in ad hoc ways. While these approaches provide information for ontology construction, the sole intent of these interactions is development of the ontologies, and users make decisions about the knowledge model itself.

Figure 1 Interaction between direct and indirect methods of knowledge elicitation (shown within dashed ellipses). Semantic-laden tasks performed by scientists and information managers (1) are captured and analysed by a system designed to marshal evidence (2) for ontology construction by knowledge modellers (3,4), which feeds back to inform ongoing tasks (5). Indirectly-driven ontologies jumpstart and extend frameworks developed through traditional interactions with the user community (6). The evidence-gathering system is a supplementary step in the development of knowledge-based solutions (7)



The complexity of ontologies and the difficulty of the knowledge modelling task presents a daunting obstacle to those who are not familiar with KR. 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 those systems exist. Hence, their personal commitment to ontology development is limited. Yet they recognise 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.

In contrast, indirect approaches engage the user in some other task that is semantic-laden, capturing and analysing their actions and providing semantic usage evidence that reflects their latent conceptual frameworks. They do not need to directly specify those frameworks. Rather, ontology construction is driven by the evidence available from usages found in typical discourse, and not by user community decision-making.

In science, there are many semantic-laden tasks from which to marshal evidence, aside from the resource discovery and integration goals of the Semantic Web. Smith (2003) suggested that philosophers often turn to science as a reliable way to learn about the things and processes operating within a given domain. Much effort in science is focused on acquiring knowledge through participating in scientific discourse, which requires some level of mastery of a specialised and interrelated terminology. This process begins during formal education but is ongoing throughout the life of a scientist, who must be able to share their own perspective while understanding those of competing explanations. These semantic perspectives are implicit in the main artifacts of science: models, datasets, and natural language communications. Creation of these artifacts involves tasks that are inherently semantic and detecting semantic patterns in these artifacts could both contribute to ontology development and be assisted by a knowledge base.

The KR method of choice in science 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. Effective mechanisms for extracting information from these sources could provide abundant information for ontology development (Cowie and Lehnert, 1996). Data collection and experimentation links scientific theories and understanding with real world objects. Datasets, therefore, capture particular conceptualisations of the objects within a scientific domain. Descriptions of datasets (e.g., metadata) provided by scientists and information managers also contain rich semantic information. Similarly, scientific models embody assumptions about how processes and objects interact, and the concepts involved are often well-defined and understood within a domain. A wealth of information about scientific concepts is thus contained in publications, datasets, and models, and scientists engage in construction of these artifacts on a daily basis. Literature-based approaches to

ontology construction are already being developed (Hahn and Schnattinger, 1998; Sicilia et al., 2003). What is lacking is a comprehensive assessment of the variety of artifacts that can constitute evidence and a formalisation of the relationship and interactions between directly-obtained and indirectly-derived ontologies.

The downside of indirect approaches is that the structure or presentation of knowledge within these artifacts represents the perspective of one or a few scientists, and does not necessarily capture the perspective of the broader community. Ontologies developed from one or a few artifacts may not provide knowledge models for which there can be widespread ontological commitment. Therefore, indirectly-derived ontologies are dependent on extensive collaborative review of the results. Conversely, the more artifacts available for evidence marshalling, the greater the ability to generate ontologies that truly represent shared conceptual frameworks.

The indirectly-driven approach represents a new way of engaging with 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, and critical mass (Grudin, 1994) that are prevalent in collaborative ontology development projects. This approach bridges the gap between formal and informal semantic approaches in ways that reduce workload and provide immediate benefits for all participants.

However, these domain ontologies must be viewed as dynamic rather than static and subject to ongoing revision. This suggests that it is important to quickly deploy semantically-based tools that will enable scientists to test and refine their knowledge models. These knowledge-based systems must be designed to accommodate such changes. After all, the goal of science is to improve our understanding and conceptual view of the observable world; the notion of a 'final' ontology is inherently at odds with the scientific enterprise. While basic domain terminology may be relatively static, periodically there are revolutions in any science that fundamentally change conceptualisations in that field (Kuhn, 1962). Ontology-based systems in science must foresee and strategically prepare for such paradigm shifts. Another common issue in collaborative projects is that terms are connoted with additional meaning, developing a project 'jargon'. This is a result of simplification or overloading of terminology. In such cases, terms employed within a project end up with broader or narrower meanings with respect to the rest of the scientific community. Ontology-based systems need to distinguish such project-specific connotations from community-sanctioned usages.

Direct and indirect approaches to ontology development must interact (Figure 1). While substantial semantic evidence can be gathered and used in many ways, advanced reasoning capabilities depend on formal semantic structuring of knowledge that someone ultimately must decide. Ultimately, knowledge modellers must make some independent decisions about how best to model the domain within a formal ontology based on the evidence available. The community must be involved in that decision-making to a greater or lesser degree, depending on the abundance of evidence and the variability shown by that evidence to achieve community commitment (Davis et al., 1993). By augmenting decision-making with evidence-gathering systems, modellers are not completely dependent on direct engagement with users. They may construct tentative ontologies based on evidence that provides a starting point for engaging the community directly, and can further refine ontologies with the formalisms and detail that would otherwise require substantial time and effort for users who may not regard provision of such detail as a high priority in their work efforts.

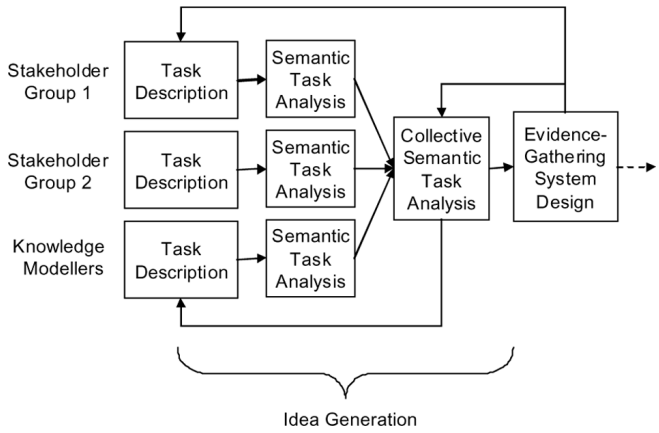
Developing evidence-gathering systems that depend on and enable group sharing of resources differs in fundamental ways from developing software that supports individuals and large organisations (Grudin, 1994). One clear difference is that in theses cases, the tasks to be supported are defined in advance by product managers or in-house IT experts, respectively. In contrast, semantic tasks are poorly defined for any new community that is to be supported. For instance, much analysis has been conducted on semantic tasks of online shoppers and therefore systems that support and make use of these activities are becoming common place (Lohse and Spiller, 1998). Those tasks are not necessarily analogous to the semantic tasks of a completely different group such as scientists.

A second difference is that introducing systems that change work patterns requires corresponding investments in dealing with sociotechnical factors that go along with change management. These issues are largely absent in development of single-user software. They are strongly present in organisational settings where there is also an infrastructure in place to provide training, restructure work, and provide leadership. Semantic systems for scientists bring about all of the challenges of changing work processes with little of the supporting infrastructure. For these reasons and 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 KR 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 2). The idea generation phase is separated out to emphasise that this may be a lengthy and time-consuming process, requiring a similar level of resource investment as the implementation phase. Idea generation is an iterative process that has the goal of understanding semantic tasks and discovering task linkages within 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, analysing those in terms of semantic tasks, then analysing the collective set for tasks that can be linked in some way. In practice this often involves a rather chaotic period of interaction between different participants as they learn about each other's perspectives and search for common ground (Olson and Olson, 2000). These interactions are difficult because they depend on overcoming the very semantic barriers that semantic systems target. In the case of engagement between two such disparate groups as knowledge modellers and ecologists, the semantic barriers can be large. Developing cross-disciplinary understanding is the first step towards the truly interdisciplinary perspective that is required for effective idea generation within semantic projects.

Figure 2 Components of an idea generation phase of development of semantic systems. Semantic tasks of all stakeholders (user community and knowledge modellers) are extracted from workflow descriptions. These are collectively analysed to identify linkages between semantic tasks of different stakeholders. These linkages are used to design systems that interact in strategic ways



3 Participants and user community

In this section and the next, we describe components of the SEEK project including stakeholders, an example of the kinds of knowledge we need to represent in these projects, and an overview of a high-level ontology we have developed to structure that knowledge. Although we focus on SEEK because it is a more mature project, analogous structures are present in the ARIES and SEAMLESS projects. We collaborate between projects such that we can leverage each others' work and create a larger, multi-disciplinary group that is more capable of critical evaluation of proposed ontologies. Our collaboration has the added benefit of leading to ontologies that will interoperate.

The SEEK Knowledge Representation (KR) group has cross-disciplinary expertise in computer science and domain science. It consists of three 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 KR. The group 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 groups' 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 developing and maintaining the ontologies in Protégé (http://protege.stanford.edu/). Another is tasked with acting as liaison to the scientific community.

The KR group 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. Some of these activities involve direct interaction for the purpose of ontology development; other activities lead indirectly to ontology development. Across the three projects, we have engaged directly with dozens of users for the purpose of ontology development. Informally, we have collected information from hundreds of users.

The KR group has attempted to directly engage groups of scientists in ontology development through small 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 such activities fades rather rapidly. Additionally, the hierarchical structures that they propose are often unusable in our ontologies due to their informal nature (e.g., they are often inconsistent). 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. In general, they only participate in activities that will quickly lead to publication. There are few short-term incentives for assisting in the development of ontologies; hence, few can remain engaged at the level needed.

Another direct source of information has been a week-long training workshop on ecoinformatics that the SEEK project held each January from 2002 to 2007. The participants in this training each year were 20 new faculty and postdoctoral associates representing the most technically-savvy of young ecologists who applied, with more than 100 participants trained. Many were tackling problems that required advanced computational approaches (Pennington et al., 2008). During the workshop, one full day was spent covering ontologies. Over the four years that the training was conducted, the ontology portion was constantly modified based on feedback from students, and many approaches were tried. In general, the students were exposed to exercises that highlighted the semantic

issues in ecological datasets and the requirements for resolving those issues. They construct ontologies for their research interests on paper. We demonstrated ontology editors and touchgraph visualisations. They stepped through portions of ontology editing exercises such as CO-ODE's pizza ontology (Horridge et al., 2004). The ontology portion of the training was the most difficult to present, and the content often received criticism in post-training surveys designed to evaluate all aspects of the workshop. Even though participants understood the semantic issues and recognised that ontologies might be useful for addressing them, they did not think semantic approaches were important for them to learn. In the most recent workshop (January 2007) survey feedback indicated that 50% 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 interaction with ontologies is an obscure task for ecological scientists and more compelling demonstrations are needed for communicating the value of semantic models.

Given all of these issues, the KR group had to be creative about finding other ways to obtain community input. We decided to rely on activities that indirectly provide evidence of the communities' ontological perspectives. For example, information managers often construct controlled vocabularies from the keywords used by scientists to describe their publications and datasets. These controlled vocabularies were further used by our group to create some simple domain ontologies. Hence, rather than directly engaging scientists and information managers we took advantage of the tasks these individuals were already performing to capture evidence regarding semantic usage for ontology development. Based on those outcomes, we were able to provide feedback to the information managers, suggesting revisions to their controlled vocabulary.

4 Ontology needs

In each of our projects, KR is tightly integrated into technical research and development. We are working toward (semi-)automated resource discovery and integration, including approaches for finding and merging heterogeneous datasets and constructing workflows that pipe data through heterogeneous computing environments (Bowers et al., 2004; Bowers and Ludaescher, 2004; Berkley et al., 2005). We are also constructing ontology-driven environmental decision support 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 help to automate many low-level tasks that the user would otherwise have to undertake manually.

Our ontology development has been two tiered:

- development of an upper-level conceptual framework for observation and measurements (core ontology)
- development of domain-specific extensions to the core ontology.

Our early work was more focused on the first task though the need for domain extensions was known. An initial version of the core observation and measurement ontology developed within SEEK is described in Madin et al. (2007) and is undergoing additional development.

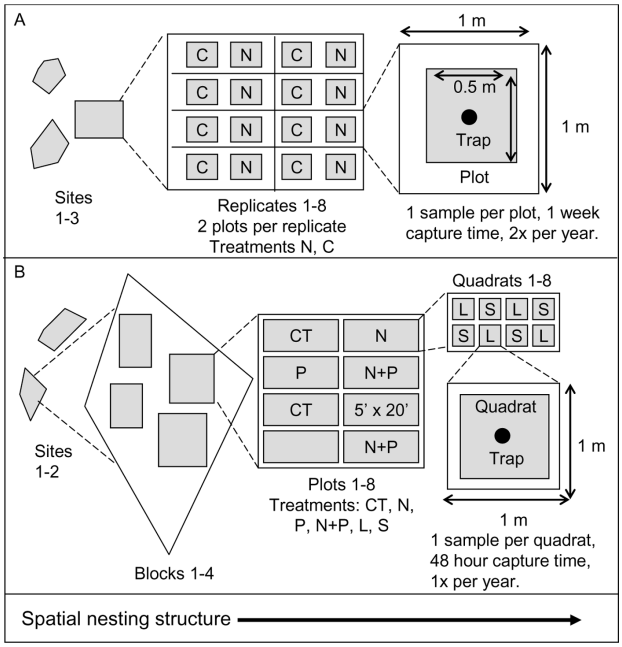
4.2 Observation, measurement, and metadata

Scientists make observations about the world that are recorded as measurements and captured in datasets (Madin et al., 2007). The design of ecological field experiments depends on the nature of the questions of interest, the context within which those are embedded, and the preferences of the scientist with regards to which factors of the environment to measure or alter, and how to accomplish this. Two scientists investigating the same phenomena may thus design very different experiments. For example, two ecologists studying changes in the quantity of aphids relative to ladybugs in the field under different chemical treatments may use quite different experimental designs (Figure 3). Although their final measurements might be identical (counts of the numbers of ladybugs and aphids), these numbers might not be directly comparable. Also, because these ecological field experiments are conducted in nature, their experimental designs might involve important spatial considerations. Unfortunately, the terminology used to describe these spatial/experimental nested structures is not consistent. The same term may refer to different parts of the experimental hierarchy.

The common characteristics of these two field experiments are that they both measured numbers of ladybugs and aphids in controlled environments where different treatments were applied. At least one of the treatments was common to both – nitrogen enrichments to the soil – and both coded that treatment with the single letter 'N'. The spatial arrangement of experimental design differed, as well as the terms that were used to refer to different parts of that design. Any attempt to integrate these observations into a single dataset must resolve these semantic discrepancies. Other discrepancies are less obvious. A capture time of one week will allow many more insects to be captured than a capture time of 84 h therefore some sort of transformation must be applied to make these semantically equivalent. The transformation may or may not be a linear relationship. Additionally, decisions must be made about how to transform data from the first experiment, conducted twice per year, into annual measurements comparable to the second dataset. Finally, in this case the smallest sampling areas were both one meter square, but this is not always the case. Samples could be from different sized areas. As with temporal differences, the transformation of values of different measurements between

different sized areas may or may not be linear depending on the phenomena being studied. For example, aphid density might scale linearly with area, but trap efficiency might not.

Figure 3 Comparison of two real field experiments, both investigating the relationship between ladybugs and aphids under different chemical treatments. Each experiment has a nested experimental structure that takes place on the ground within spatial constructs. The terminology used to describe the experiment and the spatial structuring differs between experiments. In some cases the same term means different things; in others semantically-equivalent things are given different names. Both experiments catch insects in traps, and count the number of ladybugs and aphids captured. Associated datasets code each of the nesting levels along with insect counts. A) Three 'sites' each contain eight 'replicates'. Each replicate has two 'plots'. Each with a chemical treatment. An insect trap is at the centre of each plot. B) Two 'sites' each are divided into four 'blocks', each of which has eight 'plots'. Each plot receives a different chemical treatment and is divided into eight 'quadrats'. A trap is at the centre of each quadrat

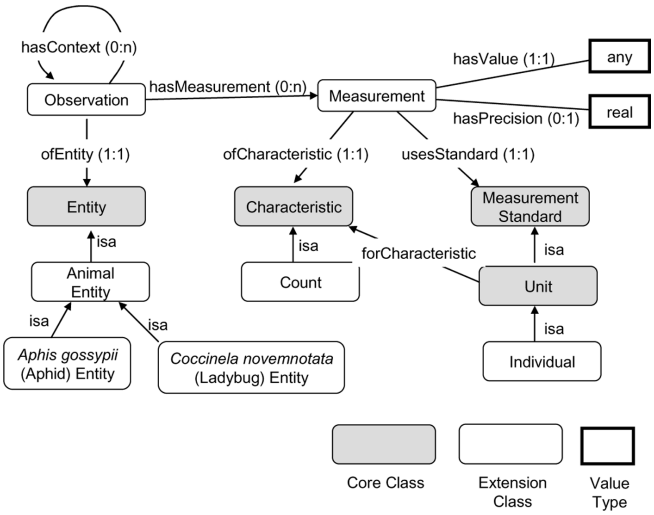


Substantial effort has been placed on developing better metadata standards for ecology that help capture and describe the full range of methods in experiments such as these so that others may make sense of, and reuse, datasets (Michener et al., 1997; Brilhante and Robertson, 2001; Fegraus et al., 2005). These efforts have focused on standardised metadata descriptions that are human readable. These rich descriptions can be further formalised through the use of ontologies to automatically discover and (at least potentially) integrate datasets from disparate yet semantically related field experiments.

4.2 Ontologies

The core ontology developed within the SEEK project is called the *Extensible Observation Ontology* (OBOE), and provides a formal and generic conceptual framework for describing the semantics of datasets that contain observations and measurements. In OBOE, an observation is about an *entity* (concept or thing), and a measurement is of a *characteristic* of the entity (Figure 4). Measurement relates a value to a *measurement standard* (e.g., a unit) as well as an estimate about the confidence level of the value (e.g., measurement *precision*). OBOE provides a structured approach for connecting domain-specific ontology terms to data 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, and linking these to specific observations and measurements.

Figure 4 Use of the Extensible Observation Ontology (OBOE) to structure field data through annotation of classes (ellipses) and properties (arrows). Field data represent counts of numbers of ladybugs and aphids in the experimental projects shown in Figure 3. Core classes represent any observable phenomena. Extension classes represent specific scientific domains, in this case organismal ecology. Modified and simplified from Madin et al. (2007)



OBOE differs from other existing ontologies related to datasets and scientific measurement that focus on the physical structure of datasets and attaching methods, pedigree, etc. to the notion of a dataset as a whole. OBOE differs by providing very detailed, conceptual descriptions of the contents of a dataset via the notions of observation, entity, characteristic, etc. OBOE also differs from a number of other similar ontology and relational approaches for describing observations by providing a flexible and generic notion of observation context, whereas most other approaches assign a standard set of contexts to observations (like space, time, etc.). Madin et al. (2008) provide a survey of these other approaches and compare them with OBOE.

In the given example, ladybugs and aphids are represented by domain entity classes. They are measured using counts (a characteristic) that have a value and a precision. The standard unit is an individual. In this case,

multiple observations must be combined to fully describe the semantics of these experiments (not shown in figure). These multiple, associated observations are often intuitively assembled by scientists into a single row or record of a table. OBOE enables a scientist to be much more explicit about the interrelationship among observations within a record. For example, nesting of the spatial entities (plots, sites, replicates, etc.) within a record can be clarified by using the *hasContext* property. Madin et al. (2007) describe in detail the use of OBOE.

Any given experiment is likely to require representation of many different kinds of observations, requiring many different domain extensions. While OBOE enforces a formal framework for describing the structure of observational data, the framework must be extended with domain ontologies. It is development of these extensions that requires novel approaches to knowledge acquisition.

5 Semantic tasks, workflows and computer support scenarios

Here we provide five scenarios of semantic-laden tasks and workflows carried out in science (specifically ecology) and a vision for computer-supported interaction mechanisms for different stakeholders. These are derived from real examples of work that our user communities engage in. The scenarios are untested 'thought experiments' which will hopefully inspire new architectural designs necessary to support such systems. We include examples of ongoing efforts that are relevant to development of such systems. The gap between this future vision and the methods we are currently able to deploy is quite large and will depend on advances across many perspectives within the KR and computer science communities, rather than just within our projects. Our hope is that these scenarios will lead to innovative thinking from many different perspectives.

5.1 Data management scenario

Karen is an information manager for a biological field site. 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 community. She is aware of the observation ontologies that are being developed, understands that they enable even better data discovery and integration than her approach. She wants to take advantage of ontologies without fully understanding the nuances of the ontology or learning ontology languages. 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 terms 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 system 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. 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. The system also tracks what changes are made, when, and who makes them. This information is carried with each term and is available for later inspection.

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 associate 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 make associations. 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 the term selected for their controlled vocabulary. 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 analyse 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 an appropriate match for their term, they can request several levels of help: tips and tricks, online documentation of annotation procedures, examples, live chat with a knowledge modeller, 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 association. Asynchronously, a knowledge modeller could 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 modeller 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 association. They will get automatic notification of the final decision made by the knowledge modeller. (Task support for the knowledge modeller is further discussed in Section 5.5.)

When Karen and her colleagues apply keywords to resources such as datasets or publications, they each apply terms from their individual controlled vocabulary. The system constructs the correct ontological associations and adds those to the metadata. The metadata therefore includes keywords from the local vocabulary and one or more ontology annotations allowing the resources to be used with ontology-driven discovery and integration tools.

Example effort. The utility of the functionality envisioned in this scenario is illustrated by a controlled vocabulary effort currently underway by LTER information managers in collaboration with the SEEK KR group (Porter, 2006), who conducted a mining project on network datasets and publications to develop a controlled vocabulary. A list was generated by compiling all words appearing in metadata titles, keywords, and attributes, and in publication titles and keywords. The resulting list contained 21,153 terms. The list was filtered for stop words such as 'of' and 'the'. Terms were then manually rated in importance based on a number of usage criteria. The information managers are continuing to collaboratively review this list to develop a controlled vocabulary for use in tagging datasets and publications. They provided this list to SEEK's KR group, who were able to incorporate these terms into domain 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.

Key needs. This scenario highlights basic functionality that must be provided by any semantic system supporting science. First, there is a need for flexible use and interaction between different semantic approaches. Thesauri, controlled vocabularies, folksonomies, ontologies and other approaches should interact seamlessly such that users can choose to work within a semantic approach that is appropriate for their own context and level of understanding yet provides opportunities to acquire functionality provided by other approaches with a reasonable level of effort. Obviously the additional functionality they can acquire will depend on the semantic approach they have chosen. Nevertheless, to the extent interactions between approaches can be enabled, they should be. Second, users should be able to use whatever term they would like to use, without being forced to change to an 'approved' term. If they make the effort to associate their term with a concept in the knowledge base the system should be able to keep track of that link such that they can continue to use their preferred term yet still make use of semantically-enabled functionality. Third, the system should track the provenance of all terms and concepts. This is important for both the users and the knowledge modeller. Knowledge of the history of a term enables review and understanding of perspectives embedded in a knowledge base. Those perspectives differ across communities and the utility of any given knowledge base for a particular individual or group depends on alignment of their perspective with the perspective represented in the knowledge base. Provenance captures the evolution of a perspective within a knowledge base. Fourth, collaboration must be appropriately supported. The efficacy of semantic approaches is closely linked to community engagement and sharing therefore collaboration must be not only facilitated but encouraged. The gains from collaborating on the system must outweigh the cost. This means incorporating highly useful collaboration functionality. It also leads to the last basic need, flexible, intuitive interfaces and a support system that provides different levels of help on demand. While this is true of any software, complex, integrated systems that are not used on a daily basis must have these in order to be approachable.

5.2 Data description and registration scenario

John, an ecologist, wants to contribute his data to a portal so that he can participate in a new collaborative project that will analyse plant species from around the globe. To do so, he must provide metadata that includes ontological annotations. He has numerous spreadsheets with similar but slightly varying schemas that he has collected over a number of years.

Information managers have previously developed a web application that walks users through the process of creating metadata for datasets. The application accesses their knowledge base, which contains 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 and online help for novices.

John creates metadata for the first dataset. He loads the dataset into the web application, which analyses 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 registration 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 domain ontologies that are extensions of OBOE (our upper-level ontology), the system prompts him to indicate 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 modeller if it requires modification of the ontology.

John has several datasets with an identical schema to the first dataset that has been described and annotated. 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 analyses 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. For example, there could be datasets of different sites (i.e., change in spatial reference), or different period (change in temporal reference). 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 recognises 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.

Example efforts. Cushing et al. (2007) are developing a database design tool for ecologists that includes domain-specific database components they call templates, which contain integrity rules. This approach allows ecologists to select the templates that match their conceptual view and automatically generate appropriate databases.

Jones et al. (2007) have developed software that uses formal, structured metadata encoded in XML to generate structured datasets and customised data entry forms. Their system is based on the Ecological Metadata Language (Fegraus et al., 2005), and output data sets that adhere to this specification. The datasets and forms can be used within handheld devices in the field as data is being collected. Validation routines generated from the metadata check for errors in data as it is entered and help maintain data integrity.

Key needs. In addition to the key needs identified above, this scenario raises the issue of levels of review. Ontology curation is a demanding task for which there are currently few resources or business models. It is likely that at least in the near-term, this task will need to be shared as much as possible. Functionality for various levels and kinds of reviewers to interact and distribute the workload is critical. This scenario also emphasises the need for ontology-enabled wizards that assist with common data management tasks. Most scientists conduct their own data management activities and may not have an information manager available to consult. Semantic systems depend on well structured datasets, and the extent to which good dataset design can be facilitated from the beginning will greatly impact their utility. Semantic systems should be able to provide generic data management support and also information regarding how those concepts are best applied in a particular domain of interest. For instance, in designing a new table for collection of a particular kind of field data, the system could use an ontology of database design to provide advice and best practices, discover available datasets that illustrate those guidelines and that are semantically equivalent to the data the scientist intends to collect, and suggest one or more table designs.

5.3 Data integration scenario

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 videoteleconferencing options. As they discuss the datasets they are able to map between these 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 analyse their choices and suggest changes to the ontology to the knowledge modeller.

Example efforts. The SEEK project is implementing a number of tool prototypes (Bowers and Ludaescher, 2004, 2005, 2006) for ontology-based semantic annotation of datasets and analytical components within the Kepler Scientific Workflow System (http://www.kepler-project.org). Tools are also being developed that use ontology annotations to infer mappings between datasets (i.e., to merge multiple datasets into a single, unified dataset) and analytical components (i.e., to help users compose workflows), as well as to enable ontology-based discovery of data and workflow components.

Goguen (2004) and Wang et al. (2007) have constructed a user-centric, semi-automatic schema matching system that multiple algorithms to both suggest mappings and to discover critical points where user input is both necessary and maximally useful. Their system can be used with any category of schema, including XML schemas, relational schemas, and ontologies. Hence, their system can be used to help users map between two dataset schemas, or between a dataset schema and an ontology.

Key needs. Semantic systems that support scientists must include functionality for data visualisation and exploration. Data discovered through semantic searches may or may not be desirable for a particular scientific effort, and no scientist will use data without any further inspection. The power of semantic systems for discovery must be combined with equally powerful mechanisms for evaluating the relevance of the data returned. Data visualisation and exploration is critical for this endeavour. Enabling (semi-)automatic mapping between resources and generation of integrated resources is one of the most fundamental needs that will drive adoption of semantic systems by scientists, who are increasingly synthesising disparate, heterogeneous information using manual approaches. This scenario also highlights the possibility of tracking user choices to inform development of ontological relationships. The combined actions of many scientists mapping between many semantically-described datasets can provide a wealth of information regarding within and across domain semantic relationships.

5.4 Research collaboration scenario

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 decide to make use of a new web application that provides collaborative concept mapping and is linked to the knowledge base. Concept maps (Novak and Wurst, 2005) are a form of directed graph that captures associations (links) between concepts (nodes). The utility of concept maps as a mechanism for enabling interdisciplinary discussion in ecology has been demonstrated (Heemskerk et al., 2003; Jeffrey, 2003).

They enter the website and rather than choose specific ontologies, they select the data 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 is not in the ontology they can suggest terms. The linkages between terms in the diagram provide information about relationships between concepts that the system tracks, analyses, and can use to suggest changes to the knowledge modeller.

Once they have each constructed their own diagrams they can collaboratively view and discuss each others work using various Web-based and audio/video collaboration tools. They can draw diagrams together representing their collective views. As they discuss the diagrams they implicitly 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 concept maps it will provide titles of datasets in the portal that are associated 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 (Michener et al., 2007). 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.

Example efforts. The Institute for Human and Machine Cognition (http://www.ihmc.us) is using techniques from logic, AI, and cognitive sciences to analyse the semantics of notations, such as mathematical diagrams and concept maps. Software tools for issue and argument visualisation are underway (Kirschner et al., 2003). These disparate efforts towards collaborative diagramming in virtual environments are not (as yet) linked with formal ontologies. The ARIES project is exploring algorithmic designs for semi-automatic alignments of concept maps. Our research in this area was inspired by existing work on ontology mapping (Kalfoglou and Schorlemmer, 2003). Salayandia et al. (2006) are developing an approach for defining workflow-driven ontologies that capture classes and relationships from domain experts and use that knowledge to support composition of services.

Key needs. A simple feature highlighted in this scenario is the ability to choose a resource and request usage of the ontologies used by that resource, rather than selecting ontologies from a list. Another desirable feature in this scenario is showing linked resources of different types from within a different component of the system. In the scenario linked datasets and publications are shown within a concept map, but another variation would be showing linked publications and concept maps while exploring and visualising data. The main functionality highlighted in this scenario is support for concept maps and other diagrammatic forms. Scientists draw many sorts of diagrams and frequently find that mode of expression useful while discussing complicated cross-disciplinary subjects (Larkin and Simon, 1987). Process diagrams, flow diagrams, project diagrams – there are an unlimited number of forms and uses of diagrams in science. 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 we consider two scientists drawing diagrams about the same research area, each will have their own diagram using the same or different terms and relationships. 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 (Pennington, 2006). Science is about investigating areas of our understanding where there is not agreement, and understanding linkages across hierarchies rather than within hierarchies. Systems must facilitate working with different views of a set of ontologies based on individual perspectives and choices about representation. During scientific discourse, these disparate concept spaces may or may not become partially aligned.

Concept maps and other diagrams 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 a knowledge modeller and subsequent review and editing for clarity, cohesion, and soundness.

5.5 Ontology review scenario

Chris is a knowledge modeller working within the ecological 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, experiment with various formal semantic constructs, make revisions and generate automated responses 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. He can also 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 provenance of all of the collective changes throughout the life of the ontology is available. The system analyses 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.

Example effort. SEEK is exploring different ways of extracting knowledge from a popular ecological textbook (Begon et al., 2006) for use as extensions of the OBOE framework. The group 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 group 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').

Key needs. The basic need in this scenario is a system that analyses all of the semantic-rich information generated in the other scenarios and makes recommendations to a knowledge modeller. It is this collective functionality that we refer to as an indirectly-driven semantic system. Other needs include the ability to flexibly explore, compare and modify multiple ontologies individually and collaboratively. As mentioned above, mechanisms to track provenance of concepts are important.

5.6 Lessons learned from the scenarios

The scenarios highlight a number of sources of indirectly-derived evidence, some of which are easily obtained now but many of which depend on research and development of an evidence-gathering system. The kinds of functionality that are needed in an evidence-gathering system include:

- useful, usable, and fully supported functionality
- individual's can work within their own, custom semantic terminology and views, while linkage to the community view provides generic functionality
- flexible migration from semantic views to views of resources linked to particular terms
- knowledge-based scientific diagramming
- easy ontology curation and levels of review
- grid approach to ontologies – ability to easily discover, access and use different ontologies either from a centralised query or by selecting a resource and choosing the ontologies that it uses
- integrated data discovery, visualisation and exploration functionality
- knowledge-driven wizards for data management and other common tasks
- enables collaborative interactions on all aspects
- tracking of user choices and ontology revision recommender functionality
- provenance tracking of all terms, concepts, and linkages
- flexible interaction between a variety of semantic approaches
- scaffolding of functionality – interactions between approaches enable greater semantic functionality than any given approach would generate.

Participants in our group are following various lines of research related to the above needs. We are investigating informal concept representation, and socio-cognitive processes for generating indirect evidence from users. We are pursuing funding for certain facets of evidence-gathering systems, particularly for collaborative design of scientific research. We are continuing research and development on ontology construction for ecology, and the use of ontologies in workflow and modelling systems. Lastly, we have a long history of research on heterogeneous data integration, and the use of ontologies to assist such efforts. These many lines of research tackle different aspects of the problem, but ultimately converge on enabling sensemaking of complex knowledge and information in ecology using emerging semantic-based approaches.

6 Conclusions

This paper describes interactions that have taken place between a KR group, ecological scientists and information managers and uses those to drive a set of scenarios for design of systems that enable better collaboration on ontology development. Previous interactions have been hindered 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 participants. 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 to conduct science – datasets, publications, and models – and makes use of them in ways that reveal 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 several 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 explicitly account for feedbacks throughout the entire process including providing mechanisms to incorporate the changing views of the broader community in long-term system development.

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 recognise the many relevant discussions with the rest of the SEEK and ARIES teams, along with valuable comments by anonymous reviewers.

References

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., and Sherlock, G. (2000) 'Gene ontology: tool for the unification of biology', *Natural Genetics*, Vol. 25, No. 1, pp.25–29.

Athanasiadis, I.N. (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., Hershey, Pennsylvania.

Begon, M., Townsend, C. and Harper, J.L. (2006) *Ecology*, Blackwell Publishing, Oxford, p.752.

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)*, University of California, Santa Barbara, 27–29 June, pp.75–78, Available online at URL: http://2005.ssdbm.org/

Bowers, S. and Ludaescher, B. (2004) 'An ontology driven framework for data transformation in scientific workflows', *Proceedings of Data Integration for Life Sciences (DILS)*, Leipzig, Germany, 25–26 March, pp.1–16.

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)*, Toronto, Canada, pp.57–63.

Bowers, S. and Ludaescher, B. (2005) 'Actor-oriented design of scientific workflows', *Proceedings of the International Conference on Conceptual Modelling* (ER), LNCS, Klagenfurt, Austria, Vol. 3716, pp.369–384.S.

Bowers, S., and Ludaescher, B. (2006) 'A calculus for propagating semantic annotations through scientific workflow queries', *Proceedings of the EDBT Workshop on Query Languages and Query Processing*, LNCS, Munich, Germany, Vol. 4245, pp.712–723.

Brilhante, V. and Robertson, D. (2001) 'Metadata-supported automated ecological modelling', *Environmental Information Systems in Industry and Public Administration,* Idea Group Publishing, Hershey, PA, USA, pp.313–332.

Cowie, J. and Lehnert, W. (1996) 'Information extraction', *Communications of the ACM*, Vol. 39, No. 1, pp.80–91.

Cushing, J.B., Nadkarni, N., Finch, M., Fiala, A., Murphy-Hill, E., Delcambre, L., and Maier, D. (2007) 'Component-based end-user database design for ecologists', Journal of Intelligent Information Systems, Vol. 29, No. 1, pp.7–24.

Davis, R., Shrobe, H. and Szolovits, P. (1993) 'What is a knowledge representation?', *AI Magazine*, Vol. 14, No. 1, pp.17–33.

DiCastri, F. (2000) 'Ecology in a context of economic globalization', *BioScience*, Vol. 50, No. 4, pp.321–332.

Fegraus, E., Andelman, S., Jones, M. and Schildhauer, M. (2005) 'Maximizing the value of ecological data with structured metadata: An introduction to ecological metadata language (EML) and principles for metadata creation', *Bulletin of the Ecological Society of America*, Vol. 86, No. 3, pp.158–168.

Goguen, J.A. (2004) 'Data, schema and ontology integration', *Workshop on Combination of Logics: Theory and Applications (CombLog'04)*, 28–30 July, Lisbon, Portugal, 9pp., Available online at URL: http://www.cle.unicamp.br/e-prints/vol_4,n_5,2004.html

Grudin, J. (1994) 'Groupware and social dynamics: eight challenges for developers', *Communications of the ACM*, Vol. 37, No. 1, pp.92–105.

Heemskerk, M., Wilson, K. and Pavao-Zuckerman, M. (2003) 'Conceptual models as tools for communication across disciplines', *Conservation Ecology*, Vol. 7, No. 3, pp.8–17.

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, pp. 118, Available at http://www.co-ode.org/resources/tutorials/ProtegeOWLTutorial.pdf

Jeffrey, P. (2003) 'Smoothing the waters: observations on the process of cross-disciplinary research collaboration', *Social Studies of Science*, Vol. 33, No. 4, pp.539–562.

Jones, C., Blanchette, C., Brooke, M., Harris, J., Jones, M. and Schildhauer, M. (2007) 'A metadata-driven framework for generating field data entry interfaces in ecology', *Ecological Informatics*, Vol. 2, No. 3, pp.270–278.

Kalfoglou, Y. and Schorlemmer, M. (2003) 'Ontology mapping: the state of the art', *Knowledge Engineering Review*, Vol. 18, No. 1, pp.1–31.

Kirschner, P.A., Buckingham Shum, S.J. and Carr, C.S. (2003) *Visualizing Argumentation: Software Tools for Collaborative and Educational Sense-Making*, Springer-Verlag, London, p.216.

Kuhn, T.S. (1962) *The Structure of Scientific Revolutions*, University of Chicago Press, Chicago.

Larkin, J.H. and Simon, H.A. (1987) 'Why a diagram is (sometimes) worth ten thousand words', *Cognitive Science*, Vol. 11, No. 1, pp.65–100.

Lohse, G.L. and Spiller, P. (1998) 'Electronic shopping: designing online stores with effective customer interfaces has a critical influence on traffic and sales', *Communications of the ACM*, Vol. 41, No. 7, pp.81–89.

Madin, J., Bowers, S., Schildhauer, M., Krivov, S., Pennington, D. and Villa, F. (2007) 'An ontology for describing and synthesizing ecological observation data', *International Journal of Ecological Informatics*, Vol. 2, No. 3, pp.279–296.

Madin, J., Bowers, S., Schildhauer, M. and Jones, M. (2008) 'Advancing ecological research with ontologies', *Trends in Ecology and Evolution*, Vol. 23, pp.159–168.

Masolo, C., Borgo, S., Gangemi, A., Guarino, N. and Oltramari, A. (2003) *Wonderweb Deliverable D18 (Ontology Library)*, Technical report, CNR, available online at URL: http://www.loa-cnr.it/Papers/D18.pdf

Michener, W.K., Beach, J.H., Jones, M.B., Ludaescher, B., Pennington, D.D., Pereira, R.S., 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/

Michener, W., James, W., Helly, J., Kirchner, T. and Stafford, S. (1997) 'Non-geospatial metadata for the ecological sciences', *Ecological Applications*, Vol. 7, pp.330–342.

Newell, B., Crumley, C.L., Hassan, N., Lambin, E.F., Pahl-Wostl, C., Underdal, A. and Wasson, R. (2005) 'A conceptual template for integrative human-environment research', *Global Environmental Change*, Vol. 15, pp.299–307.

Novak, J.D. 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*, Springer-Verlag, Berlin, Heidelberg, Vol. 3426, pp.95–116.

Olson, G.M. and Olson, J.S. (2000) 'Distance matters', *Human-Computer Interaction*, Vol. 15, pp.139–178.

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*, 6 November, 2006, Georgia, Athens, 10pp., Available online: http://www.ordnancesurvey.co.uk/oswebsite/partnerships/research/research/terracognita.html

Pennington, D.D., Michener, W.K., Katz, S., Downey, L. and Schildhauer, M. (2008) 'Transforming scientists through technical education: a view from the trenches', *Computing in Science and Engineering, Special Issue on High Performance Computing in Education*, Vol. 10, No. 28, DOI: 10.1109/MCSE.2008.125.

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

Rizzoli, A.E., Donatelli, M., Athanasiadis, I.N., Villa, F. and Huber, D. (in press) 'Semantic links in integrated modelling frameworks', *Mathematics and Computers in Simulation*, Article in press, Doi:10.1016/j.matcom.2008.01.017.

Salayandia, L., Pinheiro da Silva, P., Gates, A.Q. and Salcedo, F. (2006) 'Workflow-driven ontologies: an earth sciences case study', *Proceedings of the Second IEEE International Conference on E-Science and Grid Computing*, 4–6 December, 2006, IEEE Computer Society, Washington DC, p.17, DOI= http://dx.doi.org/10.1109/E-SCIENCE.2006.15

Smith, B. (2003) 'Ontology: an introduction', in: Floridi, L. (Ed.): *Blackwell Guide to the Philosophy of Computing and Information*, Blackwell, Oxford, pp.155–166.

Villa, F., Athanasiadis, I.N. and Rizzoli, A.E. (in press) 'Modelling with knowledge: emerging semantic approaches to ecological modelling', *Environmental Modelling and Software*.

Wang, G., Rifaieh, R., Goguen, J., Zavesov, V., Rajasekar, A. and Miller, M. (2007) 'Towards user centric schema mapping platform', *International Workshop on Semantic Data and Service Integration (SDSI) 2007*, 23 September, Vienna, Austria, 10pp., Available online at URL: http://www.dcs.bbk.ac.uk/~ptw/vldbw07/sdsi/SDSI_paper_5.pdf

Welp, M.A., de la Vega-Leinert, A., Stoll-Kleemann, S. and Jaeger, C.C. (2006) 'Science-based stakeholder dialogues: Theories and tools', *Global Environmental Change*, Vol. 16, pp.170–181.