Review

Advancing ecological research with ontologies

Joshua S. Madin[1,2], Shawn Bowers[3], Mark P. Schildhauer[1] and Matthew B. Jones[1]

[1] National Center for Ecological Analysis and Synthesis, University of California, Santa Barbara, CA 93101, USA
[2] Department of Biological Sciences, Macquarie University, New South Wales 2109, Australia
[3] UC Davis Genome Center, University of California, Davis, CA 95616, USA

Ecology is inherently cross-disciplinary, drawing together many types of information to address questions about the natural world. Finding and integrating relevant data to assist in these analyses is crucial, but is difficult owing to ambiguous terminology and the lack of sufficient information about datasets. Ontologies provide a formal mechanism for defining terms and their relationships, and can improve the location, interpretation and integration of data based on its inherent meaning. Ontologies have assisted other disciplines (e.g. molecular biology) in unifying and enriching descriptions of data, and ecology can benefit from similar approaches. We review ontology efforts in ecology, and describe how these can benefit research by enhancing the location and interpretation of relevant data for confronting crucial ecological questions.

> *'Without concepts it is impossible to work scientifically. The price for this, however, is that the concepts determine the ways and methods in which we perceive nature. Critical examination of the concepts of their field is therefore part and parcel of every scientist's obligations. This is particularly the case in ecology: nearly all ecological concepts are hotly contested (e.g. the concepts 'competition' and 'density dependence'), but the contested point is the role that these concepts actually play in nature – the concepts themselves are relatively clear and simple to define.'* [1]

Ambiguous scientific terminology

Ecological science progresses by testing hypotheses based on fundamental conceptualizations of processes and patterns in the real world (such as biodiversity). These conceptualizations range in complexity, but typically derive from simpler, underlying facts or measurements directly linked to real-world phenomena. For example, the terms 'productivity', 'community', 'competition', 'restoration' and 'ecosystem' reflect concepts that are often subject to broad interpretation [1–3]. A lack of formalization of such concepts has led to numerous controversies in ecology, including the well known debate on the relationship between ecosystem complexity and stability [1,3–5]. This lack of formalization has also negatively impacted the ability of researchers to find and incorporate relevant data into broader-scale ecological studies [including synthetic and meta-analyses (see Glossary)] [6], where crucial ecological information is often left undocumented or presented in ambiguous and context-sensitive ways (e.g. whether a 'biomass' measure can be used to calculate 'productivity' or a 'biodiversity' measure is limited to 'herbaceous shrubs'). This situation extends to other fields as well, where the need to address problems of human impacts upon natural systems (such as threats of climate change, and species invasion and extinction) [7–9] increasingly requires access to multi-disciplinary information, including chemical, behavioral, geological, meteorological and sociological data. Whereas use of ambiguous terms can stimulate discussion and lead to multiple operational definitions [5,10], terminological ambiguity slows scientific progress, leads to redundant research efforts, and ultimately impedes advances towards a unified foundation for ecological science [4,11,12].

Modern technologies, such as the Internet and electronic data catalogs, enable researchers to exchange

Glossary

Automated reasoning: algorithms and software systems for automating the computation of logical inferences (typically refers to procedures for deductive reasoning).
Controlled vocabulary: a predefined, authorized set of terms, as opposed to natural language vocabularies where there is no restriction on the vocabulary that can be used.
Domain (of discourse): a set of semantic terms that is specific to any one focal area, such as ecology.
Glossary: a list of terms and their definitions in a particular domain of knowledge.
Mathematical logic: investigates and classifies the structure of statements and arguments through the study of formal systems of inference.
Ontology: a formal model that uses mathematical logic to clarify and define concepts and relationships within a domain of interest (e.g. behavioral ecology).
Precision and recall: within information retrieval, precision refers to the fraction of results retrieved that are relevant, whereas recall refers to the fraction of all possible relevant results successfully retrieved.
Relational database model: the set of relation (table) names, attribute (column) names and types (e.g. integer or string valued), and constraints (e.g. primary and foreign keys) in a database.
Semantic annotation: capturing the mapping of data to classes in an ontology.
Semantic web initiative: a broadly scoped effort led by the world wide web consortium (W3C) to enable software systems to easily find, analyze, share and integrate web content. The W3C has created many technology specifications for extending web content, including via ontologies using the web ontology language (OWL).
Synthetic and meta-analyses: synthetic analyses are informed by multiple types or scales of data, typically from different scientific disciplines, while meta-analyses are based on data or results from previously independent studies.
Thesaurus: a list of terms and their similar (synonym), related (broader or narrower) or opposite (antonym) relationships with one another.

Corresponding author: Madin, J.S. (madin@nceas.ucsb.edu).

0169-5347/$ – see front matter doi:10.1016/j.tree.2007.11.007

Box 1. What is an ontology?

Concepts and relationships

\Ontologies are formal models that define concepts and their relationships within a scientific domain such as ecology. Analogous to mathematical set theory, an ontological 'concept' (i.e. set) denotes a collection of 'instances' that share common characteristics. The backbone of ontologies is the 'is-a' relationship, which states that all instances of a sub-concept (i.e. subset) are also members of a super-concept and, therefore, inherit all characteristics of the super-concept (Figure I). For example, Tree would generally be defined as a sub-concept of Plant. There are other commonly used relationships that describe how concepts interact, including 'part-of' (or, conversely, 'has-part'), 'equivalence' and 'disjoint' relations. In a 'part-whole' (i.e. 'part-of' or 'has-part') relationship, the instances of one concept (e.g. Tree Branch) are components of instances of another concept (e.g. Tree). These relationships are constrained by the number of instances enabled in the relationship using cardinality restrictions (e.g. a Tree Branch can only be 'part-of' one Tree). In an 'equivalence' relationship, two concepts denote the same set of instances (e.g. Animals and Metazoans), whereas in a 'disjoint' relationship, the instances of the two concepts are mutually exclusive (e.g. Plants and Animals). Relationships and cardinality restrictions are inherited through 'is-a' relationships; for example, instances of the Deme concept have two or more Organism instances as parts, because Deme is a sub-concept of Population.

Formal representation

Ontology modeling languages, such as the Web Ontology Language (OWL) [30] for the Semantic Web [31], are based on a subfamily of mathematical logic called 'description logic' [22]. The formal underpinnings of these languages offer advantages over less formal approaches, such as controlled vocabularies, thesauri and concept maps. For example, ontology languages enable precise expressions of the meaning of a scientific assertion that can be checked for consistency and compared with other formal assertions. Through automated reasoning techniques, it is possible to automate the process of determining whether an ontology is internally consistent and to infer new relationships between concepts (beyond those explicitly given in the ontology). For example, in Figure I although Barnacles have Biological Parts (i.e. Barnacle 'is-a' Animal, Animal 'is-a' Organism and Organism 'has-part' Biological Part), and Tree Branches are Biological Parts (i.e. Tree Branch 'is-a' Biological Part), Barnacles cannot have Tree Branches because Animals are 'disjoint' from Plants. Although these relationship implications might be obvious to scientists, ontologies enable computers to deduce the implications of long chains of these formal assertions.

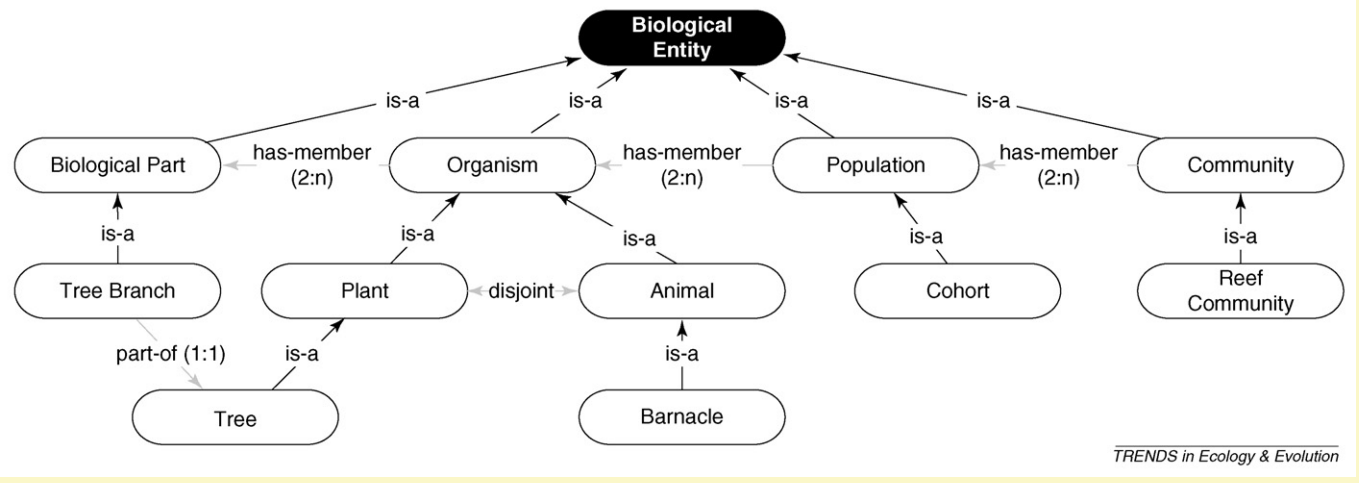

Figure I. An ontology fragment representing some Biological-Entity concepts and their relationships. In this graphical notation, ellipses denote concepts, arrows denote relationships and cardinality restrictions are given in parentheses. For example, any instance of Tree Branch is a part of one and only one (i.e. 1:1) instance of a Tree; but, conversely, an instance of Tree has at least two or more (i.e. 2:n) parts that are instances of Biological Part, because 'has-part' relationships and cardinality are inherited from super-concepts. This ontology represents only one interpretation of the domain Biological Entity, where other interpretations can similarly be described and possibly interrelated using different ontologies.

ideas and information with greater rapidity and ease than ever before. Effectively locating information that is directly relevant to a study, and then integrating it for analysis, requires seamless access to datasets and improved mechanisms for describing their information content [13–16]. Most of the current approaches for describing scientific data, however, rely on the *ad hoc* use of user-supplied keywords, and do not ensure that such terms are defined and used consistently. Similarly, to search for relevant data, users supply their own search terms, which are then matched against keywords assigned to datasets. Not surprisingly, these systems are not effective for finding relevant data (they have both low precision and recall [17]) or for determining similarities and differences between datasets (e.g. for integration). These problems are especially prevalent in fields such as ecology in which many terms are used, often having multiple interpretations and variable meanings in different contexts.

Ontologies provide a formal representation of the terminology and concepts in a scientific domain, which can be used to clarify the relationships among those terms and concepts. The use of ontologies has proliferated in recent years in the molecular biology and biomedical communities, providing benefits to those disciplines by facilitating closer collaboration and better synthetic analyses owing to precise and unified descriptions of their fields' data contents [18,19]. For example, an initial goal of the gene ontology consortium (http://www.geneontology.org) was to unify the terminologies from different model-organism domains (fruit fly, mouse and yeast) for improved querying across their major database efforts [19]. Ecology stands to benefit in similar ways by developing ontologies to control and clarify terms, and thereby enhance data-sharing capabilities [15,20,21].

In this article, we first describe problems stemming from the lack of formalized concept definitions in ecology

and discuss how ontologies can help. We then review existing ontology-development efforts and conclude by describing how ontologies can be used within a broader infrastructure for improving ecological information management.

How ontologies can help

Ontologies are sets of terms and concepts interrelated using mathematical logic [22–26] or other similar languages, and are therefore considered formal representations of the knowledge in a domain of discourse (Box 1). A familiar example of an ontology is the Linnaean classification system, which sorts and relates living organisms based on shared phenotypes [27]; this ontology imposes a strict hierarchical structure on terms with well specified rules for how to place objects within it, despite the persistent ambiguity between taxonomic names and concepts [28]. Although ontologies might seem an esoteric topic for today's ecologist, biology has a strong ontological tradition dating back over two millennia to Aristotle, who introduced several concepts familiar to ecologists, including 'quality', 'genus', 'category' and 'hypothesis' [26,29]. Ontologies hold great promise as a unifying mechanism for representing knowledge because they are interpretable by both humans and computer applications [30,31], and subsequently facilitate the use of automated reasoning [22] for helping with arduous data management tasks that scientists deal with on a daily basis.

Although large amounts of ecological data are increasingly available for research, finding data relevant to a given study is still difficult. For instance, as discussed in the previous section, most search engines rely on straightforward text-matching algorithms, which return large numbers of results that have to be manually examined to determine their relevance [17]. Ontologies can streamline these approaches by enabling users to submit queries that can be used to obtain more relevant and precise results (Box 2). Obtaining benefits from ontologies, how-

Box 2. Finding data using ontologies

Finding relevant data and integrating them are challenging tasks: the definitions of data variables can be ambiguous, their relationships unspecified and the context of data collection is often undocumented. For example, column labels 'wt', 'bm' and 'LL' in separate datasets from studies of ecosystem productivity might all refer to a measure of 'biomass of leaf litter', but this biologically meaningful information is not exposed in the abbreviated labels, or, if accompanied with metadata, is not described in a uniform or consistent way (i.e. using only uncontrolled natural-language terms and abbreviations). By 'semantically' annotating data with concepts from ontologies (Figure I, green and blue arrows), it becomes possible to clarify the implicit, or hidden, meaning of data. Furthermore, ontologies can be used to construct precise queries and to infer relationships between terms as part of the search process, leading to a greater number of relevant search results.

For example, we might define the concept Dry Weight as a sub-concept of Biomass, and Biomass as a sub-concept of Weight in our ontology (Figure I). A conventional keyword search for 'biomass' would return only those datasets with variables directly labeled 'biomass' (or whose metadata directly contains this term). By annotating data to ontology concepts, the search can automatically be expanded to include additional terms, in this case, searching for 'biomass' would also find variables annotated with 'dry weight' and 'wet weight' [58–60], regardless of the actual variable name (e.g. 'LL' or 'wt'). Queries can also be constructed explicitly using ontology concepts. For instance, rather than searching simply for 'leaf litter', a scientist can search via an implicit concept such as 'leaf litter biomass in Southern California'. This query represents a complex concept definition built from other concepts (Leaf Litter, Biomass, Southern California) and relationships (e.g. 'located-in'), which can be further expanded (e.g. to include Dry Weight) by an ontology-enabled search engine [22,32]. Furthermore, if the spatial and temporal context and semantic nature of data have been sufficiently described (including measurement standards and units; Figure I), compatible parts of resulting datasets could be automatically merged through ontology reasoning systems [6,33,61].

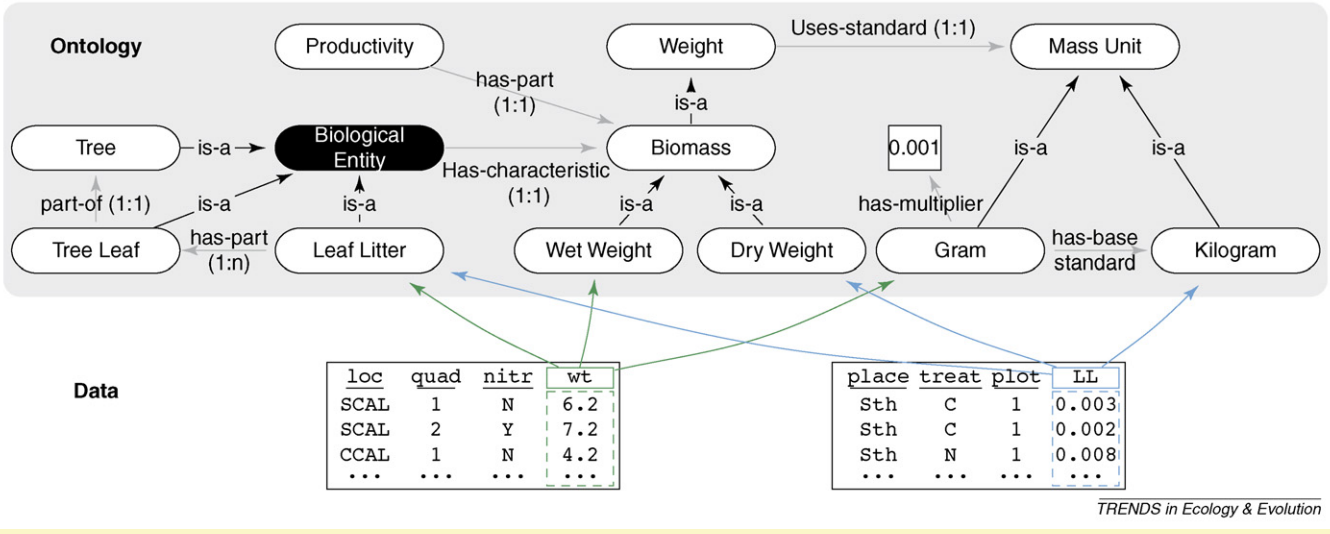

Figure I. Semantic annotations create a mapping between data and an ontology (shown using the same graphical notation as for Box 1). Data points in columns from different datasets are semantically annotated with ontology concepts (the black ellipse, Biological Entity, is the same concept as in Figure I in Box 1). For example, the green arrows show that the value 7.2 in the dataset on the left is associated with three concepts: Leaf Litter (the entity measured), Wet Weight (the measured characteristic of the entity) and Gram (the standard used to measure the characteristic). Blue arrows show similar annotations for the dataset on the right. By traversing relationships in the ontology (e.g., 'is-a', 'part-of', and 'has-characteristic'), a query for 'biomass' could now find data from both datasets, because measurements that are instances of Wet Weight and Dry Weight are also instances of Biomass. Moreover, the information needed to integrate the data points is represented, which in this case would involve conversion of measurement units from grams to kilograms.

Box 3. Building consistent ontologies

Developing meaningful and consistent ontologies can be challenging. Like most modeling tasks, ontological modeling requires an understanding of the domain, clearly defined assumptions (e.g. concerning the scope of the domain being modeled) and a plan for how domain concepts should be represented in the model. Formal ontology languages (e.g. description logic) provide modeling constructs (Box 1), which when combined in particular ways can lead to subtle changes to the model. As the number of ontology efforts increases, guidelines to help with ontology construction [62–64] become essential for ensuring well defined and consistent ontologies.

Figure I outlines some common 'pitfalls' in ontology creation identified in the OntoClean methodology that was developed in response to the creation of inconsistent ontologies in the past [62]. For example, distinctions between concept membership and sub-concept relations are often misinterpreted. When considering unique specimens of an oak *Quercus lobata*, the concept *Q. lobata* might be considered a sub-concept of all Trees (Figure Ia, left). However, when considering the concept *Q. lobata* as a species according to a taxonomic opinion, it is more naturally modeled as an instance of a Species concept (Figure Ia, right). More precisely, in one model *Q. lobata* is a concept, whereas in another model it is an instance of a different concept.

Another common mistake involves using sub-concepts to represent part-whole relationships. For instance, although Trees have Branches, particular Branches are not Trees (Figure Ib). A similar problem concerns over-constraining sub-concepts owing to corresponding 'part-of' relationships. For instance, not all Branches are Tree Branches; some reef corals, for example also have branches. In general, correctly modeling the desired part-whole relationships can be difficult. However, considerable work has been done on identifying the kinds of part-whole relationships commonly used [65–67] and their representations in formal ontology languages [32].

Another common pitfall concerns the difference between sub-concepts and 'constitution' [62], which involves determining whether instances of concepts are whole. For example, the instances of concepts Wood and Water are quantities (or amounts), and are therefore not recognizable as whole objects, whereas instances of the concept Tree are. Thus, the concept Tree is not a sub-concept of Wood, because Tree instances cannot be Wood instances (Figure Ic).

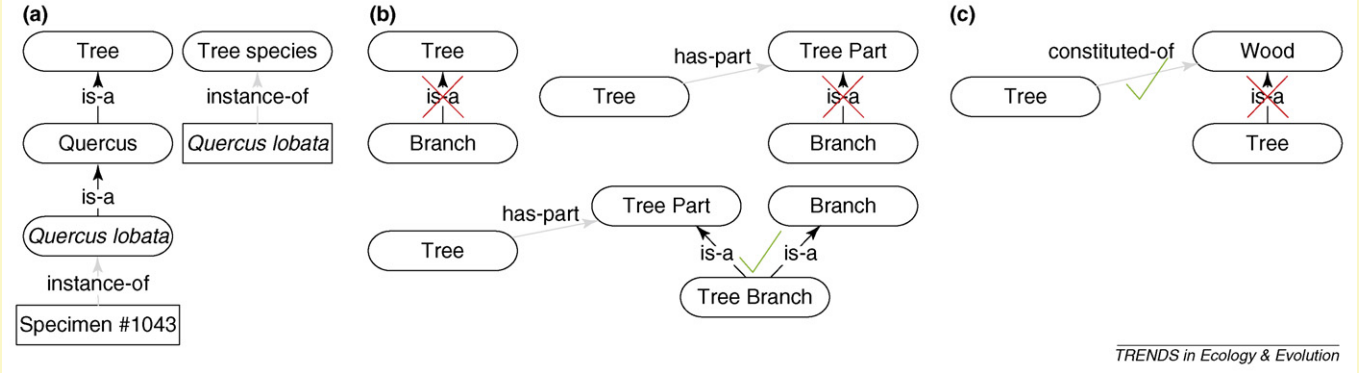

Figure I. Ontology fragments illustrating common mistakes when constructing ontologies. Part **(a)** compares different uses of sub-concept and instance relations. Part **(b)** shows an example where sub-concepts are incorrectly used to represent part-whole relations (top) as well as a more appropriate model (bottom). Part **(c)** highlights a common pitfall concerning the use of sub-concepts to denote so-called 'constitution' relations.

ever, requires that ecological concepts be formally represented. For example, the ecological concept 'community' has multiple interpretations [12] and by formally defining these different usages (e.g. as a 'group of two or more interacting populations of species' as in Box 1, or 'the assemblage of interacting organisms in an area typified by a dominant taxon', such as a kelp-forest community), scientists and computer applications can begin to resolve these differences. The process of ontology construction can be challenging however (Box 3), and should involve collaboration between ecologists and computer scientists to build topical and logically consistent ontologies.

Advanced information discovery becomes possible when ontology concepts are mapped to data points in datasets (Box 2) – a process often referred to as 'semantic annotation' – which requires scientists to spend time describing the content of their data. Through such annotations, ontologies provide additional, hitherto implicit, information about data, which both improve data visibility to search engines and enable greater levels of automation of common data transformation, sub-setting, summarization and integration tasks [6,32–34]. Furthermore, the flexibility of the annotation approach enables multiple interpretations of data. By contrast, traditional database integration approaches typically require that data be transformed to fit a single relational database model, which can lead to semantically incompatible data being discarded from the integrated dataset [15,35]. For example, different scientists can provide different annotations of the same dataset to make it simultaneously useful for separate lines of inquiry (e.g. using leaf litter as a measure of productivity or as a defoliation measure following a hurricane) or to capture differences of opinion about what the data represent (measures of net primary productivity or for determining nutrient-recycling rates). These benefits require the existence of logically consistent ontologies and corresponding software applications to use them.

Currently, there are several efforts within the ecological community to build ontologies that are useful for describing data. These efforts are community driven and informed by specialists in different areas of ecology working closely with computer scientists familiar with the logical formalisms of ontologies (e.g. http://seek.ecoinformatics.org and http://spire.umbc.edu). However, progress is slower than in fields such as molecular biology, where ontology and supporting application development are widespread and well funded activities. The limited success of ontology-building efforts in ecology, when compared to other disciplines, also stems from the broad and interrelated topics spanned by

ecological research, where data are often collected for a targeted purpose without consideration for how data might be reused for broader projects and analyses.

A trend toward ontologies in ecology

Interest in developing ontologies for describing ecological data is growing, because newer synthetic approaches to ecological analyses increasingly rely on streamlined access to a broad range of cross-disciplinary data sources and larger-scale monitoring studies [6,7,36]. Here, we provide an overview of selected ontologies that are key to ecology, organized into three broad categories: (i) domain-specific ontologies that focus on capturing terminology used in specialized scientific disciplines or communities; (ii) framework ontologies that define general concepts and relationships that others can extend when building domain ontologies; and (iii) other less formal approaches that identify and describe terms and concepts that are relevant to ecological research.

Some ontology efforts described in this section are aligned with the Semantic Web initiative [31], which is a program to standardize languages and technologies for improving information exchange over the Internet. Many of the approaches described here represent ontologies using the W3C Web Ontology Language (OWL; Box 1) and use standard software systems for editing [e.g. Protégé (http://protege.stanford.edu)] and processing [e.g. Pellet (http://pellet.owldl.com)] OWL documents. Benefits of using OWL accrue from it being a well defined, formal language for expressing ontology concepts and relationships, which facilitates exchanging ontologies between groups or software systems using a single representation format, and leads to growing support within a variety of software systems.

Domain-specific ontologies

Williams *et al.* [37] describe a set of OWL domain ontologies for modeling food-web interaction networks [38]. In food-web modeling, data are collected and integrated from a variety of studies of individual species, including direct observations of feeding interactions and stomach contents. Because of the difficulty of directly observing all components of the diet of a species, food-web models are often incomplete and inaccurate [37]. The ontologies of Williams *et al.* were developed to address these issues by providing concepts and relationships to test the consistency of, and infer missing information in, food webs. For example, some species in a food web might be annotated as being herbivores, a concept defined in the ontology as an organism that consumes plants. From the definition of herbivore, it becomes possible using software tools that employ automated reasoning to ensure that all items consumed by the herbivorous species are members of the plant kingdom. Similarly, missing taxonomic information about prey can be automatically inferred, for example that the prey is a plant, when the prey's species is unknown or only a common name is provided.

Domain ontologies are also being developed to broadly catalog information about plant and animal taxa. For example, the ETHAN ontologies (http://spire.umbc.edu/ont/ethan.php) focus on biological taxonomies and associated natural history characteristics. A primary aim of ETHAN is to provide access to categorical and summary data needed for comparative biology, and to provide ontologies for data collected on populations or individuals of species. The ETHAN Evolutionary Tree ontology is an OWL-based representation of the Animal Diversity Web data (http://animaldiversity.ummz.umich.edu), which also can be used to represent the evolutionary relationships among organisms from the Integrated Taxonomic Information System [ITIS (http://www.itis.gov)]. All species and higher taxonomic levels are represented as ontology concepts (e.g. species *Canis lupis* is defined as a sub-concept of its corresponding genus *Canis*). The ETHAN Keywords ontology is a separate OWL ontology that uses sub-concept relationships (Box 1) to associate behavioral and natural history characteristics with taxonomic concepts. For example, *Acropora hyacinthus* (a reef-building coral species) would be represented as a sub-concept of the ETHAN concept Reef Dwelling Thing. Numerical characteristics also can be given for entire taxon groups (e.g. elevation and depth ranges can be assigned for habitat dimensions, or mass and length ranges for physical descriptions).

Of the more widely used domain ontologies are those developed under the Semantic Web for Earth and Environmental Terminology (SWEET) project (http://sweet.jpl.nasa.gov). The goal of SWEET is to define general ontological concepts and relationships that can be used to both describe and help find earth-science data and information. The SWEET ontologies are represented using OWL, and are used by software systems and research projects for improved data integration and access capabilities [39]. The SWEET ontologies were initially constructed by defining formal relationships between terms in the Global Change Master Directory (GCMD) keyword list (see 'Other approaches'), and have since been expanded and subdivided into so-called 'facet' and 'phenomena' ontologies. The facet ontologies represent orthogonal sets of concepts (e.g. living and non-living things, and space and time), whereas the phenomena ontologies define concepts derived from facet-ontology concepts (e.g. hurricane is defined by high wind speeds and rainfall, and low barometric pressure). Taken together, the SWEET ontologies include numerous terms for describing properties, structures and processes of the Earth, as well as terms for units, numeric quantities, sensors and human activities.

Additional examples of domain ontologies that are relevant to ecology include: EngMath [40], a precursor to SWEET that formalizes physical quantities, dimensions and units; so-called 'upper ontologies' [26,41], which attempt to define the universal concepts that underlie all domains; and highly specialized ontologies for describing ecological niches [29,42,43] and other specific terms used in ecology [4,44–46]. Many of these approaches also have been incorporated into domain ontologies specifically targeted at ecology. For example, the EcoLingua ontology [46] further extends EngMath to include various ecological concepts.

Framework ontologies

Framework ontologies are designed to interconnect existing domain-specific ontologies while providing a consistent foundation for future ontology-building efforts. Several

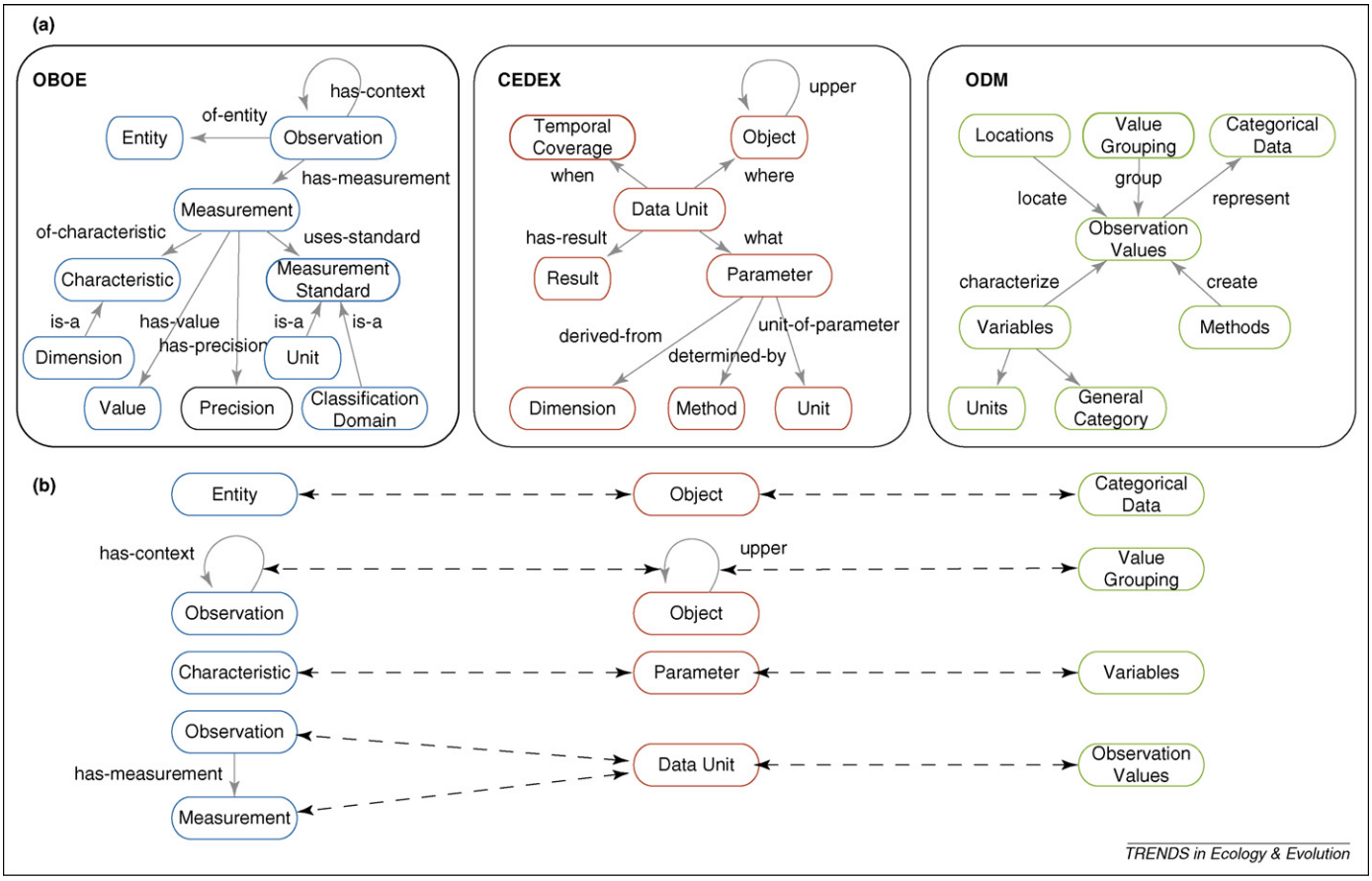

Figure 1. Framework ontology development in ecology and the environmental sciences. Part **(a)** shows fragments of three independently developed ontologies for describing scientific observations and measurements: the Extensible Observation Ontology (OBOE) is shown on the left in blue, the Classes for Environmental Data Exchange (CEDEX) ontology is shown in the middle in red and the Observations Data Model (ODM) is shown on the right in green. Similar to the figures of Boxes 1 and 2, ontology concepts are drawn as labeled ellipses and properties between classes are shown as labeled arrows. For example, Entity and Observation denote concepts in OBOE, where the 'of-entity' property states that instances of observations are related to instances of entities (that is, each observation is related to an observed entity in OBOE). There is significant overlap between these three ontology models as described in part **(b)** by the high-level correspondences (denoted by dashed lines) between concepts and relationships. For example, Entity, Object and Categorical Data are similar concepts in OBOE, CEDEX and ODM, respectively.

groups are developing 'framework' ontologies for describing ecological and environmental data. Because observations and measurements are the primary information expressed within most scientific data [6], these ontologies explicitly focus on modeling these concepts (Figure 1). Framework ontologies can support generic and flexible data management systems for diverse and non-standardized ecological and environmental datasets.

Datasets are viewed as collections of observations that provide information about the entity observed, the measurement taken of the entity (e.g. its weight in kilograms) and the context of the observation (e.g. when and where the observation was taken, who took the observation, and so on). Thus, the concept of observation in these approaches is the primary structure for storing scientific data, instead of less constrained approaches used in spreadsheets or tabular files. Systems that support framework ontologies typically either require tabular data to be mapped and stored explicitly within these models[1] or use lightweight approaches based on semantic annotations to enable tabular data to be viewed (but not explicitly stored) according to observations and measurements [6].

Figure 1a shows the main concepts and relationships of three independently developed framework ontologies for use in ecology. Several key concepts that are similar across these ontologies are illustrated in Figure 1b.

The Extensible Observation Ontology (OBOE) [6] (Figure 1, blue) is being developed under the Science Environment for Ecological Knowledge (SEEK) project. This framework ontology provides basic concepts and relationships for describing observational datasets, including field, experimental, simulation and monitoring data. OBOE is designed to enable detailed semantic annotation of datasets (Box 2), to be compatible with and supplement the Ecological Metadata Language (EML) [14], and to be easily extended with domain-specific concepts. In particular, new ontologies can be created from OBOE by extending (through sub-concept relations) the core OBOE concepts (namely Entity, Characteristic and Measurement Standard) and existing ontologies such as SWEET can be linked to OBOE through these concepts. To illustrate, the OBOE concept of characteristic (i.e. the traits of ecological entities being measured) can be extended to include properties such as color and moisture, as well as basic dimensions such as length and mass, and their sub-concepts (e.g. for height, distance, wing length, biomass, etc.). OBOE provides the ability to define various

[1] Many of these approaches use so-called 'triple-stores' [56,57] to store and query large numbers of ontology concepts, relationships and instances.

types of measurement standards including units (e.g. meter, gram, meter per second), 'semantic' units (e.g. grams of carbon per liter of seawater), and categorical and naming standards (e.g. color codes and geographic names), as well as different kinds of ecological entities (e.g. populations, organisms, locations, plots, replicates, etc.). OBOE also allows contextual relationships among observations to be specified (e.g. for representing spatial and temporal contexts as well as experimental treatment and sampling designs). OBOE is currently being developed to provide improved data search capabilities and tools for automating data integration tasks, and is expressed in the OWL language.

Another example is the Classes for Environmental Data Exchange (CEDEX; Figure 1, red) ontology [47]. This framework ontology is used to describe a broad array of monitoring and experimental data produced by the ALTER-Net [the European consortium of long-term ecological research stations (http://www.alter-net.info)]. CEDEX serves as the model for comprehensively storing and managing the environmental data for ALTER-Net. Similar to OBOE, the CEDEX ontology provides basic concepts for representing measurements of data and for data-sampling strategies. Concepts in CEDEX can be extended to define domain-specific ontologies. Unlike OBOE, which can independently provide additional information about datasets via semantic annotation, the CEDEX ontology and domain extensions explicitly store data within the ALTER-Net's database system, and mechanisms are provided to view and retrieve data through CEDEX concepts.

The Observations Data Model [ODM (http://water.usu.edu/cuahsi/odm/); Figure 1, green], under development by the Consortium of Universities for the Advancement of Hydrologic Science (CUAHSI), is designed to facilitate data sharing and reuse within and across hydrologic projects, environmental observatories and observing networks. Similar efforts to ODM also exist in the geospatial [in particular, Open Geospatial Consortium's Observations and Measurements (O&M) model (http://portal.opengeospatial.org/files/?artifact_id=17038)] and atmospheric [39] research communities. ODM is used to store and manage hydrologic observations and their associated metadata. In addition to defining a core observation model, ODM provides mechanisms for defining and using new concepts to describe data. Unlike OBOE, new concepts in ODM are restricted to lists of controlled terms in which sub-concept and part-whole relations are not supported (Box 1). ODM has yet to adopt OWL as their underlying interchange and representation format, and the model is implemented as a relational database.

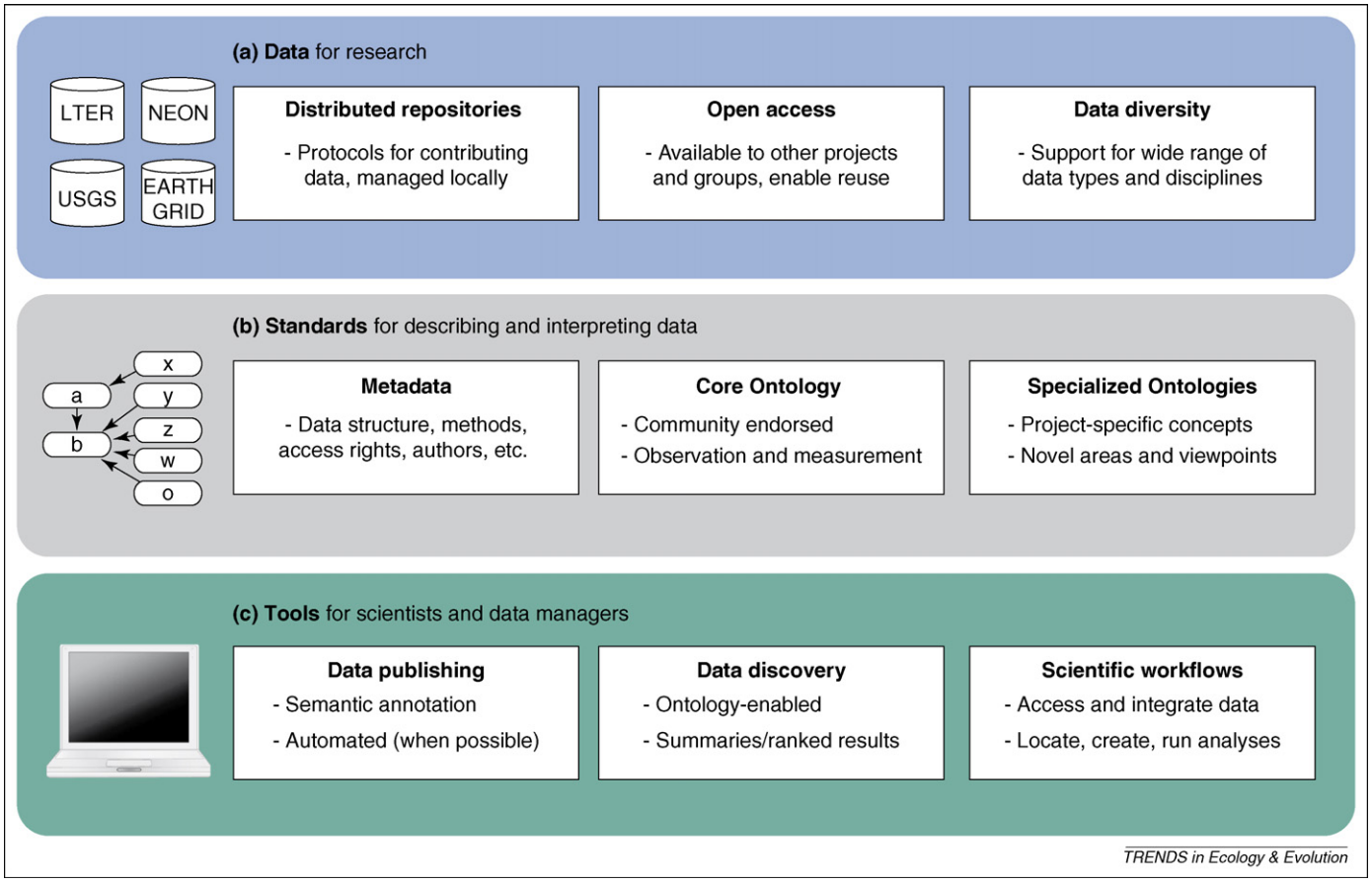

Figure 2. A proposed high-level architecture for ecological and environmental data management is shown consisting of three primary levels. Data stored within distributed data repositories **(a)** is mediated by standard metadata and ontologies **(b)** to power software tools used by scientists and data managers **(c)**. Software applications use community-endorsed ontologies and metadata standards from the middle level to provide tools that are more effective for publishing, querying, integrating and analyzing data. Ontologies are separated into framework ontologies and domain-specific extensions, enabling contributions from multiple research groups, disciplines and individuals. Cross-disciplinary data are maintained in local repositories, but made accessible to the broader research community through distributed systems based on shared, open protocols (such as Metacat). Example repositories include the LTER network, National Ecological Observatory Network, United States Geographical Survey and SEEK's EarthGrid.

Table 1. Summary of software-supported capabilities useful for sharing ecological and environmental data, and the corresponding additional information needed to enable the capability

Capability	Description	Requirements
Text query	Find datasets based on text string matches	Controlled or uncontrolled terms relevant to the dataset
Semantic ('smart') query	Find datasets or records in datasets based on ontology concepts and relationships	Well defined and consistent ontologies of concepts, with 'is-a' relations, and 'part-of' relations and constraints
Unit conversion	Standardize data values based on their measurement standards and units	Controlled lists of units and mappings to dimensionally compatible units
Taxonomic resolution	Determine taxonomic 'matches' among data	Structured model of names and taxonomic concepts
Dataset merge	Align, standardize and join datasets based on observations and measurements	Ontology of observations, measurements and context (e.g. space and time). Also, unit and taxa models
Sensible summarization	Obtain automated and useful summarizations of datasets	Semantic model of observations, measurements, context and data variables (e.g. discrete and continuous)
Statistical modeling	Distinguish appropriate analyses based on sampling design and data variables	Semantic model of statistical and ecological model including inputs and outputs

Other approaches

Several less-formal approaches than ontologies have also been used for capturing ecology-related concepts. Some of these have been used to start developing formal ontologies for ecology and environmental science, and are supported within applications and tools related to finding and sharing ecology data. One popular approach, the controlled vocabulary, provides a set of selected terms (or keywords) that each denote a distinct concept. In principle, controlled vocabularies can eliminate problems associated with term ambiguity (e.g. synonyms and homonyms) and other complications of using natural language (e.g. multiple meanings, misspellings, etc.). Extensions of controlled vocabularies include glossaries, which provide natural-language definitions of terms, and thesauri, which provide additional relationships between terms (e.g. 'broader-term' and 'narrower-term').

Examples of less-formal ontology approaches relevant for ecology include the Biocomplexity Thesaurus from the National Biological Information Infrastructure (http://thesaurus.nbii.gov), the General Multilingual Environmental Thesaurus [GEMET (http://www.eionet.europa.eu/gemet)] from the European Environment Information and Observation Network, the GCMD keyword list (http://gcmd.nasa.gov), and the United States Geological Survey thesaurus (http://www.usgs.gov/science/about). Although these approaches can provide benefits over otherwise 'unconstrained' keywords (e.g. by restricting the terms used in describing data and for searching and browsing), their informal nature limits overall applicability for supporting data searching and integration tasks that rely on automated reasoning. In addition, the modeling problems raised in Box 3 are often evident in existing controlled vocabularies and thesauri [48], which do not have associated tools to detect such problems. The capabilities that can be facilitated by formal ontology approaches within software systems are described further in the next section.

Incorporating ontologies in ecological information management

Ontologies can play a pivotal role in helping scientists both utilize and publish ecological and environmental data (Figure 2). Publishing data to repositories (e.g. to share data with other researchers) benefits from the use of metadata standards, which provide descriptive information ranging from who created and can use a particular dataset to details about the structure of data and methods of collection. For example, the Metacat data-management system [49] is a distributed repository (Figure 2a) of raw data and metadata (using EML [14]) used by organizations such as the Ecological Society of America and the US Long Term Ecological Research network. Ontologies can enhance the descriptive power of metadata (Figure 2b), enabling important software capabilities in addition to semantic ('smart') queries already described in Box 2, including unit conversion, taxonomic resolution, dataset merging (integration), sensible summarization and statistical modeling (Figure 2c and Table 1) [6].

In particular, framework ontologies such as OBOE can provide, with the help of semantic annotations, the detailed information needed to translate between the units or taxa from different datasets. Framework ontologies additionally provide information to help automate the merging of datasets by clarifying and relating their logical structures (i.e. their entities and attributes). Software applications can use this information to determine if two data attributes are compatible for a particular analytical purpose (e.g. they might be different types of size measurement but of the same 'thing' – e.g. tree height or trunk diameter), and then generate and execute the appropriate steps for merging the data.

Ontologies can also enable sensible summarization of data by exposing important contextual information about when, where and how measurements were taken. This observation context enables software to interpret the sampling design used in a series of measurements, and determine which variables represent (e.g. nesting or blocking factors) and summarize these appropriately. For example, when plots represent replicated observations in a site it would be sensible to average leaf-litter biomass by plot within a site, but not sensible to summarize leaf-litter biomass by plot across all sites. This is because the plot labels across sites are arbitrary and are not assigned to a common experimental unit (e.g. mean biomass for plot '1'

across sites would be non-sensible) (Box 2). Ontologies enable computers to interpret and utilize these experimental design implications.

Finally, ontologies can be used in statistical analysis and modeling tools, such as workflow-automation systems, which provide graphical environments for creating and running analyses, visualizing results and tracking data provenance [6,50]. In workflow systems, ontologies can assist researchers in finding relevant analyses, and indicate the data transformation and integration tasks required for analysis [51].

These approaches that leverage ontologies will be most effective when the concepts used to describe and interrelate datasets are standardized and shared among a broad community of scientists. Thus, community-driven development and endorsement of ecological ontologies, similar to efforts in the molecular biology community [18,19], are crucial to the success of information-management architectures, such as in Figure 2.

Concluding remarks

Despite the potential for ontologies to enhance ecological information management, there are few examples of such systems in use. There are several reasons why this is probably the case.

First, ecological data are still typically collected by small groups of individuals for use within their respective projects. Traditionally, data owners are the only intended users, and information about their data regarding its structure, content and appropriate usage is often not recorded. This situation is no longer tenable, because as ecological research becomes holistic and integrative, better approaches are needed for locating and interpreting all relevant data that can inform a topic [52,53]. Second, current data practices in ecology are not particularly amenable to data sharing and re-use. The prevalent 'spreadsheet' model and even sophisticated database frameworks typically lack the requisite information to facilitate effective long-term preservation and interpretation of data. However, although software products are available and used by scientists to manage data using these approaches, software is still under development for similar ontology-based approaches. Thus, the adoption of ontologies is hindered both by the familiarity of current practices and the lack of tools to readily migrate to improved practices. Third, developing comprehensive and consistent ontologies is challenging, especially within ecology, which is a complex and multidisciplinary field with concepts spanning many spatiotemporal scales, and multiple levels of organization and processes.

The need to incorporate effective semantics into data-storage systems has long been recognized [54,55], and becomes crucial given the promise of the Internet for successfully placing vast amounts of relevant information in the hands of researchers. Formal ontologies provide a mechanism to address the drawbacks of terminological ambiguity in ecology, and fill an important gap in the management of ecological data by facilitating powerful data discovery based on rigorously defined, scientifically meaningful terms. By clarifying the terms of scientific discourse, and annotating data and analyses with those terms, well defined, community-sanctioned, formal ontologies based on open standards will provide a much-needed foundation upon which to tackle crucial ecological research while taking full advantage of the growing repositories of data on the Internet.

Acknowledgements

We thank the reviewers for comments that greatly improved this paper. We also thank participants on the Knowledge Representation/Semantic Mediation Systems, and Biodiversity and Ecological Analysis and Modeling Working Groups, of the SEEK project. This work was supported by grants from the National Science Foundation (0225676, 0533368, 0630033, 0612326), including the SEEK project, and by the National Center for Ecological Analysis and Synthesis, a Center funded by NSF (0553768), the University of California, Santa Barbara and the State of California.

References

1 Grimm, V. and Wissel, C. (1997) Babel, or the ecological stability discussions: an inventory and analysis of terminology and a guide for avoiding confusion. *Oecologia* 109, 323–334
2 Callicott, J. *et al.* (1999) Current normative concepts in conservation. *Conserv. Biol.* 13, 22–35
3 Pimm, S.L. (1984) The complexity and stability of ecosystems. *Nature* 307, 321–326
4 Jax, K. (2006) Ecological units: definitions and applications. *Q. Rev. Biol.* 81, 237–257
5 Lehman, C.L. and Tilman, D. (2000) Biodiversity, stability, productivity in competitive communities. *Am. Nat.* 156, 534–552
6 Madin, J. *et al.* (2007) An ontology for describing and synthesizing ecological observation data. *Int. J. Ecol. Informatics* 2, 279–296
7 Worm, B. *et al.* (2006) Impacts of biodiversity loss on ocean ecosystem services. *Science* 314, 787–790
8 Graham, C.H. *et al.* (2004) New developments in museum-based informatics and applications in biodiversity analysis. *Trends Ecol. Evol.* 19, 497–503
9 Kerr, J.T. *et al.* (2007) The macroecological contribution to global change solutions. *Science* 316, 1581–1584
10 Scheiner, S.M. and Willig, M.R. (2005) Developing unified theories in ecology as exemplified with diversity gradients. *Am. Nat.* 166, 458–469
11 Pickett, S. *et al.* (1994) *Ecological Understanding: The Nature of Theory and The Theory of Nature,* Academic Press
12 Jax, K. *et al.* (1998) The self-identity of ecological units. *Oikos* 82, 253–264
13 Parr, C.S. and Cummings, M. (2005) Data sharing in ecology and evolution. *Trends Ecol. Evol.* 20, 362–363
14 Fegraus, E. *et al.* (2005) Maximizing the value of ecological data with structured metadata: an introduction to Ecological Metadata Language (EML) and principles for metadata creation. *Bull. Ecol. Soc. Am.* 86, 158–168
15 Jones, M.B. *et al.* (2006) The new bioinformatics: integrating ecological data from the gene to the biosphere. *Annu. Rev. Ecol. Evol. Syst.* 37, 519–544
16 Michener, W.K. *et al.* (1997) Non-geospatial metadata for the ecological sciences. *Ecol. Appl.* 7, 330–342
17 Grossman, D.A. and Frieder, O. (2004) *Information Retrieval: Algorithms and Heuristics,* (2nd edn), Springer
18 Bard, J.B.L. and Rhee, S.Y. (2004) Ontologies in biology: design, applications, and future challenges. *Nat. Rev. Genet.* 5, 213–221
19 Ashburner, M. *et al.* (2000) Gene ontology: tool for the unification of biology. *Nat. Genet.* 25, 25–29
20 Soldatova, L.N. and King, R.D. (2005) Are the current ontologies in biology good ontologies? *Nat. Biotechnol.* 23, 1095–1098
21 Smith, B. *et al.* (2005) Relations in biomedical ontologies. *Genome Biol.* 6, R46
22 Baader, F. *et al.* (2003) *The Description Logic Handbook: Theory, Implementation, and Applications,* Cambridge University Press
23 Kelly, J. (1996) *The Essence of Logic,* Prentice Hall
24 Guarino, N. (1995) Formal ontology, concept analysis, and knowledge representation. *Int. J. Hum. Comput. Stud.* 43, 625–640
25 Chandrasekaran, B. *et al.* (1999) What are ontologies, and why do we need them? *IEEE Intell. Syst.* 14, 20–26

26 Sowa, J.F. (2000) *Knowledge Representation: Logical, Philosophical, and Computational Foundations,* Brooks Cole
27 Schuh, R.T. (2003) The Linnaean system and its 250-year persistence. *Bot. Rev.* 69, 59–78
28 Kennedy, J. *et al.* (2005) Scientific names are ambiguous as identifiers for biological taxa: their context and definition are required for accurate data integration. In *Proceedings of the 2nd International Workshop on Data Integration in the Life Sciences, Springer Lecture Notes in Computer Science* (Vol. 3615), pp. 80–95, Springer
29 Smith, B. (2001) Objects and their environments: from Aristotle to ecological ontology. In *The Life and Motion of Socio-Economic Units* (Frank, A. *et al.*, eds), pp. 79–97, Taylor and Francis
30 McGuinness, D.L. and Van Harmelen, F. (2004) OWL Web ontology language overview. W3C Recommendation (http://www.w3.org/TR/owl-features)
31 Berners-Lee, T. *et al.* (2001) The semantic web: a new form of web content that is meaningful to computers will unleash a revolution of new possibilities. *Sci. Am.* 284, 34–43
32 Borgida, A. (1995) Description logics in data management. *IEEE Trans. Knowl. Data Eng.* 7, 671–682
33 Ludäscher, B. *et al.* (2001) Model-based mediation with domain maps. In *Proceedings of the 17th International Conference on Data Engineering*, pp. 81–90, IEEE Computer Society
34 Bowers, S. *et al.* (2004) On integrating scientific resources through semantic registration, In *Proceedings of the 16th International Conference on Scientific and Statistical Database Management*, pp. 349–352, IEEE Computer Society
35 Haas, L.M. *et al.* (2002) Data integration through database federation. *IBM Syst. J.* 41, 578–596
36 Costanza, R. *et al.* (1997) The value of the world's ecosystem services and natural capital. *Nature* 387, 253–260
37 Williams, R.J. *et al.* (2006) Ontologies for ecoinformatics. *J. Web Sem.* 4, 237–242
38 Yoon, I. *et al.* (2005) Interactive 3D visualization of highly connected ecological networks on the WWW. In *Proceedings of the 2005 ACM Symposium on Applied Computing* (Haddad, H. *et al.*, eds), pp. 1207–1212, ACM Press
39 Fox, P. *et al.* (2006) Semantically-enabled large-scale science data repositories. In *Proceedings of the International Semantic Web Conference, Lecture Notes in Computer Science 4273* (Cruz, I. *et al.*, eds), pp. 792–805, Springer
40 Gruber, T.R. and Olsen, G.R. (1994) An ontology for engineering mathematics. In *Proceedings of the 4th International Conference on Principles of Knowledge Representation and Reasoning* (Doyle, J. *et al.*, eds), pp. 258–269, Morgan Kaufmann
41 Gangemi, A. *et al.* (2002) Sweetening ontologies with DOLCE. In *Knowledge Engineering and Knowledge Management, Lecture Notes in Computer Science* (Vol. 2473), pp. 166–181, Springer
42 Smith, B. and Varzi, A. (1999) The niche. *Nous* 33, 198–222
43 Smith, B. and Varzi, A. (1999) The formal structure of ecological contexts. In *Modeling and Using Contexts: Proceedings of the 2nd International Interdisciplinary Conference* (Bouquet, P. *et al.*, eds), pp. 339–351, Springer
44 Keet, C.M. (2005) Factors affecting ontology development in ecology. In *Data Integration in the Life Sciences, Lecture Notes in Computer Science* (Ludäscher, B. and Raschid, L., eds), pp. 46–62, Springer
45 Mabee, P.M. *et al.* (2007) Phenotype ontologies: the bridge between genomics and evolution. *Trends Ecol. Evol.* 22, 345–350
46 Brilhante, V. (2004) An ontology for quantities in ecology. In *Proceedings of the Brazilian Symposium on Artificial Intelligence, Lecture Notes in Artificial Intelligence* (Vol. 3171), pp. 144–153, Springer
47 Schentz, H. and Mirtl, M. (2003) MORIS: a universal information system for environmental monitoring. In *Environmental Software Systems* (Schimak, G.P. *et al.*, eds), pp. 60–68, Springer
48 Doerr, M. (2001) Semantic problems of thesaurus mapping. *J. Digital Information* 1, article 52 (http://hdl.handle.net/2249.2/jodi-35)
49 Jones, M.B. *et al.* (2001) Managing scientific metadata. *IEEE Internet Comput.* 5, 59–68
50 Ludäscher, B. *et al.* (2006) Scientific workflow management and the Kepler system. *Concurr. Comp. Pract. E* 18, 1039–1065
51 Berkley, C. *et al.* (2005) Incorporating semantics in scientific workflow authoring. In *Proceedings of the 17th International Conference on Scientific and Statistical Databases* (Frew, J., ed.), pp. 75–78, Lawrence Berkeley Laboratory
52 Thompson, J. *et al.* (2001) Frontiers of ecology. *Bioscience* 51, 15–24
53 Greene, J. *et al.* (2005) Complexity in ecology and conservation: mathematical, statistical, and computational challenges. *Bioscience* 55, 501–510
54 Chen, P.P. (1976) The entity-relationship model – toward a unified view of data. *ACM Trans. Database Syst.* 1, 9–36
55 Hammer, J. and McLeod, D. (1998) On the resolution of representational diversity in multidatabase systems. In *Management of Heterogeneous and Autonomous Database Systems* (Elmagarmid, A.K. *et al.*, eds), pp. 91–118, Morgan Kaufmann
56 Sachs, J. *et al.* (2006) Using the semantic web to support ecoinformatics. In *Proceedings of the AAAI Fall Symposium on the Semantic Web for Collaborative Knowledge Acquisition* (Honavar, V. and Finin, T., eds), pp. 56–61, AAAI Press
57 Stuckenschmidt, H. *et al.* (2004) Index structures and algorithms for querying distributed RDF repositories. In *Proceedings of the International Conference on World Wide Web* (Feldman, S. *et al.*, eds), pp. 631–639, ACM Press
58 Moldovan, D.I. and Mihalcea, R. (2000) Using WordNet and lexical operators to improve Internet searchers. *IEEE Internet Comput.* 4, 34–43
59 Jarvelin, K. *et al.* (2001) ExpansionTool: concept-based query expansion and construction. *Inform. Retrieval* 4, 231–255
60 Voorhees, E.M. (1994) Query expansion using lexical-semantic relations. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Croft, W.B. and van Rijsbergen, C.J., eds), pp. 61–69, Springer
61 Bowers, S. and Ludäscher, B. (2004) An ontology-driven framework for data transformation in scientific workflows. In *Data Integration in the Life Sciences, Springer Lecture Notes in Computer Science* (Rahm, E., ed.), pp. 1–16, Springer
62 Guarino, N. and Welty, C.A. (2002) Evaluating ontological decisions with OntoClean. *Commun. ACM* 45, 61–65
63 Rector, A. *et al.* (2004) OWL pizzas: practical experience of teaching OWL-DL: common errors and common patterns. In *Engineering Knowledge in the Age of the SemanticWeb* (Vol. 3257), pp. 63–81, Springer
64 Pinto, H.S. and Martins, J.P. (2004) Ontologies: how can they be built? *Knowl. Inf. Syst.* 6, 441–464
65 Wand, Y. *et al.* (1999) An ontological analysis of the relationship construct in conceptual modeling. *ACM Trans. Database Syst.* 24, 494–528
66 Winston, M.E. *et al.* (1987) A taxonomy of part-whole relations. *Cogn. Sci.* 11, 417–444
67 Smith, B. (1996) Mereotopology: a theory of parts and boundaries. *Data Knowl. Eng.* 20, 287–303