Ontology Design for Scientific Theories That Make Probabilistic Predictions

David Poole, *University of British Columbia*

Clinton Smyth and Rita Sharma, *Georeference Online Ltd.*

A multidimensional ontology design paradigm based on Aristotelian definitions provides knowledge structure necessary for testing scientific theories that make probabilistic predictions.

Imagine having a number of expert systems that provide predictions—for example, diagnoses of what is wrong with patients based on their symptoms, or predictions of whether there will be a landslide at some particular location. Which of these predictions should we believe most? Apparently, many of Google's queries are people typing in symptoms and wanting diagnoses. Google's ranking system, based on page rank, essentially measures popularity. Other recommender systems base their predictions explicitly on some measure of how authoritative sources are. Scientists (and the rest of us) should be suspicious of both answers. We would prefer the prediction that best fits the available evidence. To this end, semantic science can provide a way to have explicit theories that make predictions together with the data upon which to test the predictions.

To enable meaningful results (and avoid what is known as "garbage in"), we need to use consistent vocabulary for the data and the predictions. We don't want a semantic mismatch between the data and the predictions. Users need to know what vocabulary to use for the new cases. Thus we need some sort of ontology to enable terms to be used consistently (or made consistent).

The work on expert systems that peaked in the 1980s has given rise to two seemingly separate fields. One is concerned with uncertainty and (statistical) learning that typically uses features or random variables. The other concentrates on ontologies and rich representations of knowledge with individuals and relations, but has essentially ignored uncertainty. This article is part of an endeavor to put these together, building on the advances in both.

The aim of semantic science is to have machine-interpretable scientific knowledge. There have been considerable advances in developing ontologies and using them to describe data and processes.[1,2]

We are advocating adding the publication of scientific theories that make predictions. Thus, the main components of our conception of semantic science are data about observations of the world, theories that make predictions about the data, and

1541-1672/09/$25.00 © 2009 IEEE
Published by the IEEE Computer Society

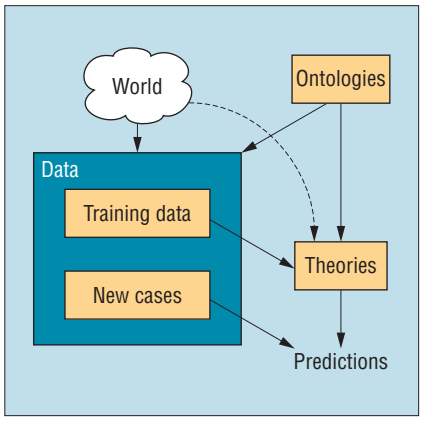

Figure 1. Ontologies, data, and theories in semantic science. The data depend on the world and the ontology. The theories depend on the ontology, indirectly on the world (if a human is designing the theory), and directly on some of the data. Given a new case, we can use a theory to make a prediction.

ontologies that describe the vocabulary used by the data and the theories.[3] The ontologies must define both the vocabulary needed to express application domains and the vocabularies of data and theories themselves. By publishing ontologies, data, and theories, researchers can use new data to evaluate existing theories, and new theories can be evaluated against existing data. Theories can be used to make predictions for new cases, and the predictions can be justified by reference to the empirical evidence.

This article is about how to define ontologies to represent observations and scientific theories that make (probabilistic) predictions. These predictions can be used to evaluate the theories on available data and can be used for new cases. We are not trying to encompass all of the activities of science, but rather add one more desideratum to the design of ontologies—namely, taking into account future uses of the ontologies for developing theories that make probabilistic predictions.

This semantic-science framework can also be motivated by starting with machine learning. We assume that the theories make predictions about individuals and relations, not just features, and thus are part of what has been called statistical relational learning.[4,5] The data and the learned theories are assumed to be persistent. The theories are built using prior knowledge and multiple heterogeneous data sources, and can be compared with other theories. When a theory is used, the data upon which it is based are available for scrutiny. The theories and the data refer to formal ontologies to allow for semantic interoperability. We expect to have the highest standards for evaluation of theories, with declarations of which data were used for training; so, there is a clean separation of training and test data.

We also assume that probability is the appropriate form of prediction for scientific theories.[6,7] Probabilistic predictions minimize the prediction error for most error measures and are what is required (along with utilities) to make decisions.

We are building systems in two domains of earth sciences: minerals exploration and geohazards (predicting landslide susceptibility). In both domains, there are multiple theories (models), and users are interested in asking what predictions different models make about a particular piece of land or about which land area best fits a model.

We base our ontologies on OWL, the W3C recommendation for representing ontologies.[8] We see OWL as an "assembly language" for ontologies. This article describes a high-level design pattern for ontologies that is suitable for designing the rich hypothesis space needed for probabilistic reasoning and shows how the resulting ontologies can be represented in OWL DL (one of the species of OWL).

Semantic-Science Overview

As we mentioned in the introduction, our semantic-science framework consists of ontologies, data, and theories.

Ontologies specify the meaning of the vocabulary. These evolve slowly and are built by communities. We expect that, through a process of natural selection, a particular community will converge on useful ontologies that interoperate. For example, the geology community is actively working on what symbols to use for rocks, minerals, and so on. (See http://onegeology.org, www.cgi-iugs.org, and www.seegrid.csiro.au for international efforts to share information and to develop standardized vocabulary.) Shared ontologies are important for semantic interoperability.

The data about observations are written using the vocabulary of the ontology. In practice, this means that data sets are published with reference to the ontologies they use, so that we can recognize when different data sets are about the same or related phenomena. For example, in the geology domain, the observations might be of the rocks and minerals (and their spatial relation with other land features such as rivers) found at a particular location of the earth. The observations do not include probabilities.

Scientific theories make predictions about (potentially) observable features or outcomes. These theories are often called hypotheses, laws, or models, and we do not distinguish between these terms. In the realm of semantic science, a distinction that depends on how well established theories (or laws or hypotheses) are is redundant because we can access the relevant data to determine how much they should be believed. There, of course, might be other reasons to distinguish these terms. Theories specify what data they can make predictions about, they make predictions that can be checked against all of the relevant data, and they can be applied to new cases. As we mentioned before, we expect these theories to make probabilistic predictions.[6,7] Again, these probabilistic theories refer to ontologies. For example, we are developing theories in the geology domain that make predictions on where minerals are more likely to be found, and theories that make predictions about where various forms of landslides are likely to occur, in terms of emerging standards of the vocabulary of earth sciences. The ontologies allow for the interoperation of the data and the theories.

Figure 1 shows the relationship between ontologies, data, and theories. The data depend on the world and the ontology. The theories depend on the ontology, indirectly on the world (if a human is designing the theory), and directly on some of the data (because we would expect that the best theories would be based on as much data as possible). Given a new case, we can use a theory to make a prediction. The real situation is more complicated, because there are many theories, ontologies, and heterogeneous data sets, and they all evolve in time. The same piece of data can act in the role of training data for one theory and in the role of a new case for another theory, and perhaps in both roles for theories that make multiple predictions (but we have to be careful not to judge a prediction by the data it was trained on). Often a prediction will rely on multiple theories (for example, in a diagnostic situation, there might be a theory that predicts whether a patient has

cancer, a theory that predicts the type of cancer, and another that predicts the severity, and all might be needed to predict the outcome for a particular patient).

The idea of "science" here is meant to be very broad. We can have scientific theories about anything. As well as traditional scientific disciplines such as geology or medicine, we could have theories about someone's preferences in real estate, theories about what companies are good to invest in, theories about how much a subway system in a city will cost to build, or theories in any domain where we can have testable predictions.

Ontologies for Semantic Science

In philosophy, ontology is the study of existence.

In AI, an ontology is a specification of the meaning of the symbols in an information system.[9] In particular, an ontology contains a commitment to what kinds of individuals and relationships are being modeled, specifies what vocabulary will be used for the individuals and relationships, and gives axioms that restrict the use of the vocabulary. The axioms have two purposes: to show that some use of the terms is inconsistent with intended interpretation, and to allow for inference to derive conclusions that are implicit in the use of the vocabulary. The simplest form of an ontology is a database schema with an informal natural language description of what the attributes and the constants mean. More formal ontologies allow machine-understandable specifications.

For example, an ontology of real estate could specify that the term "building" will represent buildings. The ontology will not define a building but give some properties that restrict the use of the term. It might specify that buildings are human-constructed artifacts, or it might give some restriction on the size of a building so that shoe boxes cannot be buildings or so that cities cannot be buildings. It might state that a building cannot be at two geographically dispersed locations at the same time (so if you take off some part of the building and move it to a different location, it is no longer a single building). Although ontologies include a number of other kinds of information, taxonomies, which are essentially naming schemes for related things according to subclass, are one of the essential building blocks of an ontology. We discuss rock taxonomies later in the article.

An ontology written in a language such as OWL specifies the vocabulary for individuals, classes, and properties. Sometimes classes and properties are defined in terms of more primitive classes and properties, but ultimately they are grounded in primitive classes and properties that are not actually defined. This can work when people who adopt an ontology consistently use the notation with its intended meaning.

The primary purpose of an ontology is to document what the symbols mean—the mapping between symbols (usually the words in an information system such as a book or a computer) and concepts. In particular, an ontology should facilitate the following tasks:

> We advocate separating definitions from predictions; the former forms the ontology, and the latter forms the theories.

- Given a symbol used in an information system, a person should be able to use the ontology to determine what the symbol means.
- The ontology should enable a person to find the appropriate symbol for a concept or determine that there is currently no appropriate symbol. Different users, or the same user at different times, should be able to find the same symbol for the same concept.
- Through the use of axioms, the system should be able to infer some implicit knowledge or determine that some combination of values is inconsistent.
- The system should be able to construct a hypothesis space over which it can put a probability distribution. Integrating this task with the other tasks is the subject of this article.

The main challenge in building an ontology is to find a structure that is simple enough for a human to comprehend, yet powerful enough to be able to represent the logical distinctions needed in the domain of interest.

This article takes a perspective on the role of the ontology and uncertainty formalisms that differs from many other recent proposals.[10,11] In particular, we do not include actual probabilities in the ontology. The ontology defines the vocabulary for a community who need to share vocabulary and the semantics of that vocabulary. As the community need not, and should not, agree on theories or probabilities, these should not be part of the ontology. The ontology should define the vocabulary to express theories, including the vocabulary to express probability. In essence, we advocate separating definitions from predictions; the former forms the ontology, and the latter forms the theories. An ontology can provide definitions that involve probabilities—for example, defining a fair coin to be one that has a 0.5 chance of landing heads—but these definitions do not make predictions until we have asserted or hypothesized that a coin is fair.

There are five reasons why the ontologies should not contain the probabilities about the domain, even though the theories might be probabilistic.

First, ontologies come logically before observational data, and probabilities come logically after. In order to have data, you need a meaning for the data. Any data come explicitly or implicitly with an ontology; otherwise they are just a sequence of bits with no meaning. In order to acquire data, we need to have some meaning associated with the data, which is the ontology. To have reasonable probabilities, we need to use as much information as possible. That is, the probabilities need to depend on the data; to make a prediction on a new case, we want to use the posterior probability based on all previous data. It is possible that someone might reinterpret some data with a different ontology, and we have to be careful not to double-count that as evidence.

Second, data that adheres to an ontology can't be used to falsify that ontology. For example, if some data adhere to an ontology that specifies that a gneiss is a metamorphic rock, then by definition, all gneisses are metamorphic rocks, so

that data cannot refute that fact. However scientific theories need to be refutable.[12] In probabilistic terms, evidence obtained from observations should change our belief in theories. This does not mean that ontologies should not change; we expect them to evolve as the requirements for representing data and theories change.

Third, to allow for semantic interoperability, a community should agree on an ontology to make sure they use the same terminology for the same things. However, as we mentioned before, a community cannot and should not agree on the probabilities, because people might have different priors and have access to different data, and the ontology should have a longer life than one data set. Also, we don't want to update an ontology after each new data set, because then we need to map between these different ontologies. We do want to update theories when new evidence becomes available.

Fourth, people should be allowed to disagree about how the world works without disagreeing about the meaning of the terms. If two people have different theories, they should first agree on the terminology (for otherwise they would not know they have a disagreement)—this forms the ontology—and then they should give their theories so that those theories can be compared.

Finally, the structure of the information a prediction depends on does not necessarily follow the structure in the ontologies. For example, an ontology of lung cancer should specify what lung cancer is, but the prediction of whether someone will have lung cancer depends on many factors that depend on particular facts of the case and not just on other parts of ontologies. Whether a person has lung cancer may depend on whether he or she worked in a bar that allowed smoking, but we wouldn't expect bars to be part of the definition of cancer, nor would we expect lung cancer to be part of the definition of bars or even smoking. As another example, the probability that a room will be used as a living room depends not just on properties of that room but also on other rooms.

In our vision of semantic science, the ontology should describe the vocabulary for any concept that needs to be shared between data and theories. In particular, we are making no claims as to the distinction between theoretical terms and observational terms.[13] People can use whatever ontology they want. This freedom means that the philosophical debate about scientific terms possibly becomes more important, but the underlying technology needs to be neutral in this debate. We advocate that people designing scientific ontologies should take into account the (future) use of these ontologies in building theories.

Representations of Ontologies

Modern ontology languages such as OWL define classes, properties, and individuals. The semantics of OWL is defined in terms of sets: a class is a set of individuals (RDF calls the individuals "resources," and individuals are also called "objects"), and a property is a set of individual-value pairs.

For semantic interoperability, a community should agree on an ontology to make sure they use the same terminology for the same things.

There are many ways to define classes in OWL. They can be defined in terms of the union, intersection, or complement of other classes or in terms of property restrictions. A class *A* can also be specified by stating it is a subclass of some other class *B*. This latter specification loses much structure that can be useful. For the rest of this section, we consider only the specification of classes that would otherwise be specified by just stating what classes they are immediate subclasses of.

The notion of a subclass is important; however, it isn't obvious that it should be primitive. Deriving the subclass property from more primitive notions exposes structure that is natural and can be exploited in probabilistic models.

An Aristotelian definition[14] of class *A* is of the form "An *A* is a *B* such that *C*," where *B* is a superclass of *A* and *C* is a condition that defines how *A* is special among the subclasses of *B*. Aristotle called *B* the *genus* and *C* the *differentia*.[15] Restricting all subclass definitions to be definitions in this form does not reduce what can be represented but provides random variables that can be exploited in probabilistic models.

Aristotelian definitions can be represented in logic and in OWL DL using what we call the *multidimensional design pattern*, where the conditions in the differentia are built from properties that form local dimensions. To define a class, first choose a superclass that will form the genus, then consider what values of what properties distinguish this class from the other subclasses of the genus. Each of these properties defines a (local) dimension. The domain of each property should be the most general class for which it makes sense. The subclass is then defined as equivalent to the superclass conjoined with the restrictions on the values of the properties defining the dimensions. Thus, in the multidimensional design pattern, a class is never just stated to be a subclass of another class. There are still subclasses; the subclass relation is just derived from more primitive constructs. Following the multidimensional design pattern does not restrict what can be represented.

Geologists have traditionally defined rocks along three major dimensions: genesis (sedimentary, igneous, or metamorphic), composition, and texture. When depicted in a taxonomy, the rocks are typically classified using first genesis, then texture, then composition. Particular rocks, such as granite and limestone, are defined as having particular values in each dimension (or some subset of the dimensions). There have been attempts to build rock taxonomies by splitting on the dimensions in order, as in the British Geological Survey Rock Classification System.[16] However, these produce taxonomies that are difficult to use because they have to commit to an order in which subclass splits are made—grain size before composition, for example, or composition before grain size. If the former order is chosen, as in the case of the British Geological Survey system, it is difficult, if not impossible in a single word, to refer to all rocks of a particular composition, irrespective of grain size. This problem is well documented.[17] A multidimensional approach to representing taxonomies solves these problems, makes the ontologies more amenable to modern computer-reasoning capabilities, and, we would argue, provides for more accurate scientific research.

Stephen Richard and his colleagues define nine dimensions in which to distinguish earth materials.[18] One dimension is the *consolidation degree*, which specifies whether some earth material is consolidated or unconsolidated. Rock is consolidated earth material. Volcanic ash is unconsolidated earth material. Another dimension is *fabric type*—the pervasive feature of a rock that specifies the directionality of the particles that are visible in it. This dimension is defined only for rocks (that is, for earth material that is consolidated). One value for fabric type is *foliated*, which means that the rock consists of thin sheets. Particular rocks are defined by their values on the dimensions.

> **Example 1.** *Richard and his colleagues define a gneiss as a metamorphic rock where the fabric type is foliated, the particle type is crystal, and the grain size is phaneritic (large enough to be seen by the human eye).*[18]

When there are rocks that have similar descriptions (such as gneisses and schists), geologists decide whether one is a subclass of the other or what features distinguish these rocks, perhaps needing to invent new dimensions.

There is not a fixed set of dimensions that distinguish all individuals. Rather, different dimensions come into existence at different levels of abstraction. For example, the dimensions of size and weight might appear for physical individuals but are not applicable for abstract concepts. This idea can be traced back to Aristotle:

> If genera are different and coordinate, their differentiae are themselves different in kind. Take as an instance the genus "animal" and the genus "knowledge." "With feet," "two-footed," "winged," "aquatic," are differentiae of "animal"; the species of knowledge are not distinguished by the same differentiae. One species of knowledge does not differ from another in being "two-footed."[15]

Note that "coordinate" means that neither genus is subordinate to the other.

Multidimensional-Ontology Assumptions

In this section we will be more formal in the assumptions behind multidimensional ontologies. We do this to show how the multidimensional structure can give us random variables with which we can define probabilistic models. To keep the discussion simple, we will ignore classes that are defined in terms of intersection, union, complement, or cardinality. Such classes are important but complicate the discussion.

In a multidimensional ontology,

- dimensions are defined by functional properties or by each value of a nonfunctional property,
- classes are defined in terms of values on properties, and
- the domain of a property that defines a dimension is the most general class on which the property makes sense.

Defining classes or subclasses only in terms of properties has advantages over trying to specify the subclass relation directly.

Assuming that subclasses are defined only in terms of their values on properties does not restrict what can be represented. An explicitly stated subclass relationship can induce a Boolean property that is true on the subclass and is false otherwise. That is, if all you know is that A is a subclass of B, you can always invent a new Boolean predicate *is_A* with domain B, and define A to be equivalent to $B \wedge is\_A$. For example, if someone states humans are a subclass of animals, this induces a property *is_human* that is true of members of the humans class and is false otherwise. Part of this article is arguing that there are advantages in explicitly representing the predicates that define classes.

In particular, we make three assumptions about the ontology. First, the top class, *Thing*, is predefined.

Second, classes are one of two types:

- *Enumeration classes* are predefined sets of values. For example, in geology, *FabricTypeValue* could be defined as the set of constants {*aplitic*, *biogenic*, *foliated*, ...}.
- *Nonenumeration classes* are made up of individuals in the world of the application domain and are defined in terms of values of properties. That is, a class A is defined as $A \equiv B \wedge C$, where B, the genus of A, is a class and C, the differentia, is a Boolean formula of property restrictions.

Finally, there is a total ordering of classes and properties such that

- *Thing* is first in the order,
- the genus of a class must be before the class in the total ordering,
- the domain and the range of a property must be before the property in the total ordering,
- the properties that define a class must come before the class in the total ordering, and
- this total ordering ensures that there are no cyclic definitions. For example, saying a *flat* is an *apartment* and an *apartment* is a *flat*, without saying what either one is, violates the acyclic condition.

(Suppose there is a cyclic set of definitions: $A_1 \equiv A_0 \wedge C_1$, $A_2 \equiv A_1 \wedge C_2$, …, $A_k \equiv A_{k-1} \wedge C_k$, $A_0 \equiv A_k \wedge C_0$. This implies that all of the A_i are equivalent and imply all of the C_i. Such a cyclic representation is very misleading and should be avoided. If there is a set of equivalent classes, this can be represented as having a canonical representation for the classes.)

Under this interpretation, a nonenumeration class can be seen as a set of property restrictions. (Each genus that is not *Thing* can be reduced to its genus and a set of property restrictions, and this can be done recursively.)

Defining classes or subclasses only in terms of properties has four advantages over trying to specify the subclass relation directly (or even trying to impose a tree structure over the abstraction hierarchy).

First, it is easy to specify, compute, and explain subclasses in terms of the dimensions, even though the induced subclass relationship might be very complex to depict.

Second, a concept does not need to specify values for all dimensions. Overlapping

```
EquivalentClasses(FabricTypeValue
                  ObjectOneOf(aplitic biogenic foliated))
DifferentIndividuals(aplitic biogenic foliated)
```

Figure 2. The OWL functional specification of the enumerated class *FabricTypeValue*. *FabricTypeValue* is an enumeration class that is equivalent to the collection {*aplitic, biogenic, foliated*}.

```
FunctionalObjectProperty(fabricType)
ObjectPropertyDomain(fabricType Rock)
ObjectPropertyRange(fabricType FabricTypeValue)
```

Figure 3. The OWL functional specification of the fabric dimension. This is equivalent to saying that rocks have a property called "fabricType" whose value must come from the set of values called "FabricTypeValue."

```
EquivalentClasses(Gneiss
  ObjectIntersectionOf(
    Rock
    ObjectHasValue(geneticCategory metamorphic)
    ObjectHasValue(fabricType foliated)
    ObjectHasValue(particleType crystal)
    ObjectHasValue(grainSize phaneretic)))
```

Figure 4. Defining the class `Gneiss` in the OWL functional syntax. A gneiss is a metamorphic, foliated, crystal, phaneretic rock.

concepts can specify values for different sets of dimensions.

Third, it is often difficult to decide on which attribute to split a hierarchy. Different splits might be applicable for different purposes. The multidimensional splitting means that you don't have to make this (often arbitrary) choice.

Finally, this approach is important for probabilistic reasoning where the dimensions create random variables (see the section "From Ontologies to Possible Worlds and Random Variables"). This provides a way to have probabilistic models (and utility models) over complex objects described using complex ontologies.

OWL and the Multidimensional Design Pattern

OWL was designed to allow for the specification and translation of ontologies. OWL allows for the specification of classes, properties, and individuals and relations between them.

It is possible to use OWL to specify ontologies using the multidimensional design pattern. It is interesting to note that we could find no tutorials or material for teaching or learning OWL that use this design pattern.

We divide the object properties into two classes:

- A *discrete property* is an object property whose range is an enumeration class.
- A *referring property* is an object property whose range is a nonenumeration class (that is, the value is an individual in the world).

The dimensions of a multidimensional ontology are defined in terms of discrete properties.

> **Example 2.** *Consider representing a gneiss, as outlined in Example 1, using a multidimensional ontology in OWL. Suppose we already have rock defined (as an earth material with the consolidation degree of consolidated). We need to say a gneiss is a rock in which the genesis is metamorphic, the fabric type is foliated, the particle type is crystal, and the grain size is phaneritic. To represent this we do the following:*
>
> *First, we create the class* FabricTypeValue *(see Figure 2; we show only three values to keep it simple; Richard lists six values,*[18] *but the earth science community may recognize additional values in the future).*
>
> *Then, we create a functional property* fabricType *whose domain is* Rock *and whose range is* FabricTypeValue *(see Figure 3).*
>
> *Similarly, we create functional properties for* geneticCategory, particleType, *and* grainSize, *each with the domain* EarthMaterial *or* Rock, *as appropriate, and a range that is an enumeration class. (Some of these enumeration classes have a hierarchical structure. This can be achieved by having subclasses of enumeration classes and using OWL's facility for a class to have some values or all values of a property in some class. A description of how to do this is beyond the scope of this article.)*
>
> *Finally, we define* Gneiss *as a rock with the appropriate values on the properties (see Figure 4).*

We claim that this multidimensional ontology fulfills the two main purposes of an ontology: given a concept, find the appropriate terminology or determine that one does not exist, and, given a symbol, determine what it means. To find the terminology for a concept, start at the top (at *Thing*) and find the value for each property that is defined. A user will never encounter a question that does not make sense. Given a symbol in the ontology, the ontology will specify what values it has on the properties that define it.

From Ontologies to Possible Worlds and Random Variables

Although we have tried to argue that the multidimensional ontology is important in its own right, a main motivation is to use it as a foundation for specifying probabilistic models. The general idea is that the dimensions form random variables for each individual.

For this article, we assume there is no uncertainty about the existence or identity of individuals. We assume that we are given a finite set of uniquely identifiable individuals in the world.

A possible world specifies, for each do-

main individual, a value for each property that is legal (is consistent with the ontology) for that individual in the possible world. In particular, the individual must be in the domain of the property and must fulfill the cardinality and other restrictions in that world.

Although this construction gives finitely many possible worlds, the number of worlds grows like $O(e^{dn}n^{in})$, where n is the number of individuals, e is the (maximum) size of the enumeration classes, d is the number of discrete properties, and i is the number of referring properties. To specify a probability distribution explicitly over such possible worlds is not feasible. Rather, we can describe the worlds in terms of random variables. The structure of random variables lets us concisely state probabilities using parameter sharing, by treating individuals about which we have the same information identically and by making explicit independence assumptions.

A natural specification is to have a Boolean random variable for each individual-property-value triple. There is, however, more structure that can be exploited for functional properties. For functional properties, there can be a random variable for each individual-property pair, where the domain of the random variable is the range of the property.

For example, in the geology domain, there are variables for the consolidation for each individual of type *EarthMaterial* and variables for the genetic category for each rock.

Given such random variables defined by properties, we can define random variables that specify the type of the individual. The type of an individual is a deterministic function of the genus and the properties that define the differentia.

This random-variable formulation is complicated by the fact that the random variables defined in terms of individual-property pairs are not all defined in all worlds. In particular, a variable is defined only if the individual is of the type of the domain of the property that defines the variable. Thus, the existence of some random variable might be dependent on the value of other variables. For example, suppose you are uncertain whether an earth material is a gneiss, and the fabric type is defined only when the object is consolidated (a rock). In terms of possible-worlds semantics, in any possible world where a particular individual is a rock, the fabric type is defined. In a possible world where the individual is not consolidated (or not earth material), the fabric type for that individual is not defined. Thus, when you talk about the fabric type of some object, you are implying it is a rock. This is reminiscent of context-specific independence,[19] but instead of one variable being irrelevant given some value of another, one variable is not defined given the value of the other.

Given an ontology made up of Aristotelian definitions, we define a possible-worlds semantics as follows. Note that the possible worlds can be heterogeneous, each with different random variables defined, so we have to be careful to refer to a random variable only in a context where it is defined. We can procedurally define what individual-variable-value triples are defined and what value they have in each world. Each different choice in the following description will give a different possible world.

> The possible worlds can be heterogeneous, each with different random variables defined, so we refer to a random variable only in a context where it's defined.

For each individual i, and for each property p, enumerated using the total ordering assumed for the acyclicity of the hierarchy, if the individual i is in the domain of the property p in the world (and by the total ordering, this only depends on values already chosen), we follow this procedure:

- If p is functional, choose a value v in the range of p that satisfies all of the other properties of p. We will say that the individual-property-value triple $\langle i, p, v\rangle$ is true in this world and that $\langle i, p, v'\rangle$ for all values $v' \neq v$ is false in this world.
- If p is not functional, for each value v, choose either true or false for the value of $\langle i, p, v\rangle$ in this world, making sure the other constraints specified by the ontology are satisfied.

An individual-property-value triple that is not assigned in the previous procedure is undefined.

To interpret any formula made up of values of global variables and of individual property-value triples, we use the standard logical connectives. However, we add a third truth value, undefined ($\perp$), interpreted as follows: for any operation op, $\perp op \perp \equiv \perp$, true $\wedge \perp \equiv \perp$, false $\wedge \perp \equiv$ false, true $\vee \perp \equiv$ true, false $\vee \perp \equiv \perp$, $\neg\perp \equiv \perp$. This logic was first introduced by Jan Łukasiewicz in 1920.[20]

For example, $\langle i_7, fabricType, foliated\rangle$ will have the value $\perp$ in any world where $\langle i_7, type, Rock\rangle$ is false. The formula $\langle i_7, type, Rock\rangle \wedge \langle i_7, fabricType, foliated\rangle$ will be true or false in all worlds where $\langle i_7, type, Rock\rangle$ is true.

We define a probability measure over the possible worlds and define conditional probabilities in the standard way. The probability of a hypothesis h, given evidence e, is the measure of the set of the worlds where $h \wedge e$ is true, divided by the probability of the measure of the worlds where e is true.

We say that conditional probability $P(h|e)$ is *well defined* if e is true in some possible worlds and $h \wedge e$ does not have the value $\perp$ in any possible world where e is true.

We can prove the following proposition:

Proposition 1. *For $P(\langle i,prop,val\rangle|\alpha)$ to be well defined, α must logically imply that i is in the class that is the domain of prop. That is, the formula that defines the class that is the domain of prop is true for i.*

Proof. *If α doesn't imply that the domain of prop is true for i, then there is a possible world where α is true and the domain of prop is false for i, but then $\langle i,prop,val\rangle \wedge \alpha$ has the value $\perp$ in that possible world, so the conditional probability is not well defined.*

For example,

$$P(\langle i,fabricType, foliated\rangle | \langle i,particleType,crystal\rangle)$$

is not well defined because the conditions do not imply that i is in the domain of *fab-*

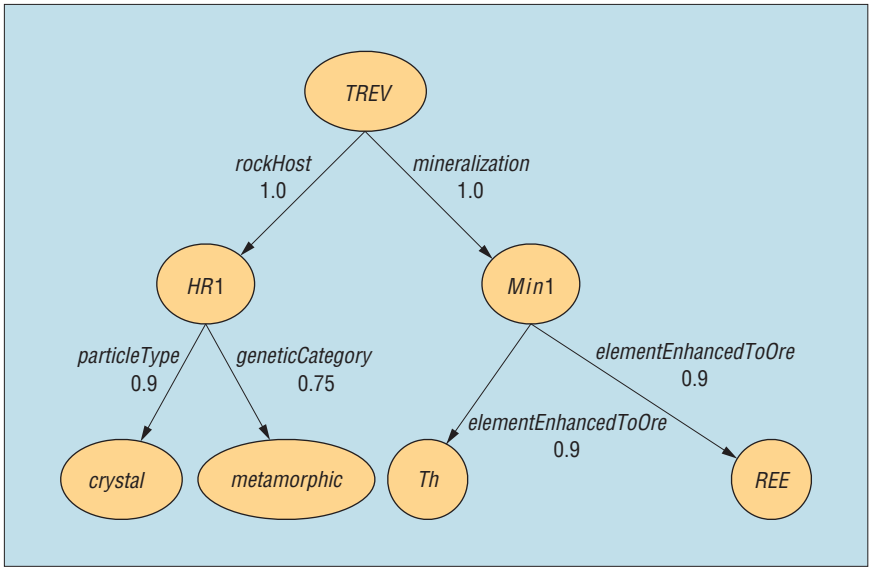

Figure 5. Part of the Thorium-Rare-Earth Vein (TREV) model. The TREV model predicts a rock host that is crystal and usually metamorphic, and contains mineralization of thorium (Th) and rare-earth elements (REE), with associated probabilities.

ricType, which is *Rock* (which requires the consolidation degree to have the value *consolidated*, and sand has crystal particles but is not consolidated).

Given this restriction, we can specify probability distributions over the types and properties of individuals. Note that we end up with conditional probabilities over triples. Although it is possible to reify such statements[11] so that they can be represented in RDF, we use quintuples that include probabilities and the providence of the triples. A standardized language for such statements will need to be developed when we have more experience in building diverse collections of theories.

An Example in Geology

Geological surveys publish descriptions of the geology of various locations in their jurisdictions. There is much work on developing ontologies to allow the interpretation of these data sets. Various models are also published, typically in natural language, that can be seen as theories that make predictions. In our applications, we have represented these ontologies, observations, and theories in order to make predictions. For example, given a model of where thorium might occur, we can predict which location is most likely to be a candidate to contain thorium. Given a particular location and multiple models, we can ask about which models best fit this location and so make predictions about that location.

Here we sketch part of a multidimensional ontology, some observational data, and part of a model. We call a description of an observation of a set of interacting individuals in the domain an *instance*.

We can define instances in terms of individuals and properties using RDF triples. (Our application does not use RDF triples because we also want to be able to specify nonexistence—that no object in the world satisfies some description. Although this can be represented in RDF by reifying the statement, it isn't very natural.) For example, La Esperanza is a mineral occurrence in Argentina. Part of its description, in a functional syntax, is

```
Age(LaEsperanza
    Precambrian)
MineralEnhancedToOre(LaEspe
    ranza muscovite)
RockHost(LaEsperanza rh1)
rdf:type(rh1 schist)
RockHost(LaEsperanza rh2)
rdf:type(rh2 gneiss)
```

The mineral occurrence is hosted in two rocks: a schist and a gneiss.

One model that can make a prediction on this mineral deposit is the US Geological Survey's Thorium-Rare-Earth Vein (TREV) model (http://pubs.usgs.gov/bul/b2004/html/bull2004thorium_rareearth_veins.htm). Here we explain a small part of this model, shown in Figure 5. This is like a semantic network in that the nodes are objects (in roles) or values and the arcs are properties. It is like a Bayesian belief network in that it defines the probability of a property value and the probability of the existence of an object that fills a role, conditioned on all of its ancestors. It represents a naive Bayesian model in that the properties are independent of each other given the roles assigned by the parents. The roles are implicit in the diagram. We describe the syntax and semantics elsewhere.[21]

We do not allow arbitrary probabilities of first-order formulas but here use a simple language that gives a naive Bayesian model of the existence of objects that fill roles and of the properties of these objects.

The knowledge represented in Figure 5 represents conditional probabilities of the form

$$P(\exists HR1\ \mathrm{r1}(HR1) \wedge rockHost(TREV,HR1) \mid thoriumRareEarthModel(TREV)) = 1.0$$

$$P(particleType(HR1) = crystal \mid r_1(HR1) \wedge rockHost(TREV,HR1) \wedge thoriumRareEarthModel(TREV)) = 0.9$$

$$P(geneticCategory(HR1) = metamorphic \mid r_1(HR1) \wedge rockHost(TREV,HR1) \wedge thoriumRareEarthModel(TREV)) = 0.75$$

$$P(elementEnhancedToOre(Min1) = Th \mid r_2(Min1) \wedge mineralization(TREV,Min1) \wedge thoriumRareEarthModel(TREV)) = 0.9$$

where r_1 is true of the object that satisfied the role represented by the node labeled *HR*1 and where r_2 is true of the object that satisfied the role represented by the node labeled *Min*1.

From the complete model, we wish to compute the probability that La Esperanza will have ore-grade thorium. What is important for this article is noticing that both the model and the instance refer to the same ontology. The instance can be built without any knowledge of any (probabilistic) models. Similarly, the model can be built without knowing about mineral occurrences in Argentina. The ontology enables the model to make predictions about the instance. These predictions can then be used by exploration geologists to make decisions. The predictions of various models

can also be used to evaluate the theories.

Finding good languages for theories and instances is an ongoing research activity. We need to define ontologies that support a wide variety of theories. Multidimensional ontologies are a good candidate for this.

Designing ontologies is difficult. There are many objectives that need to be simultaneously considered. For building scientific ontologies, we have suggested that the ability to use the ontology for defining probabilistic theories is essential. We have outlined a way that this can be done in a straightforward manner that should not distract ontology designers from the other issues that need to be considered.

The multidimensional design pattern provides more structure than stating the subclass relation directly. We argue that it is more natural and show how it can be used for probabilistic modeling.

This interaction between ontologies and probabilistic reasoning forms the foundation of applications we are building in minerals exploration and landslide prediction. This article considers only one aspect of the problem. Another aspect is, given descriptions of theories and individuals in the world at various levels of abstraction and detail, how to use them to make coherent decisions, which will also involve modeling utilities. A further aspect is that the assumption that we know the correspondence between individuals in the world and the model is not generally applicable. We need to determine which model individuals correspond to which individuals in the world (that is, which individuals fill the roles in the model). We also need to model and reason about existence and nonexistence. These are ongoing research topics that build on the foundations given in this article.

With respect to other efforts on the Semantic Web, semantic science seems to be an area where the bootstrapping problem might be the least difficult: scientists and their funders want their results to be as widely used as possible. There are large efforts going on to define ontologies in the sciences. The way that science can most fruitfully be applied is to have the theories be used for new predictions. We want a user to be able to ask, "What does the best science predict in this case?" Finally, this work directly addresses the issue of trust, which is the current top layer of the Semantic Web. We don't believe that appeal to traditional authority is the most appropriate basis for trusting a conclusion. We advocate that a user should be able to say "Show us the evidence" and ask "How well does this predictor actually work, compared to the alternatives?" There is still a long way to go to bring this vision to fruition, but the prize seems to be worth the effort. ■

THE AUTHORS

David Poole is a professor of computer science at the University of British Columbia. He's known for his work on knowledge representation, default reasoning, assumption-based reasoning, diagnosis, reasoning under uncertainty, combining logic and probability, algorithms for probabilistic inference, and representations for automated decision making. He's coauthor of a forthcoming AI textbook (Cambridge University Press, 2009), coauthor of *Computational Intelligence: A Logical Approach* (Oxford University Press, 1998), and coeditor of the *Proceedings of the Tenth Conference in Uncertainty in Artificial Intelligence* (Morgan Kaufmann, 1994). Poole received his PhD from the Australian National University. He's former associate editor and on the advisory board of the *Journal of AI Research* and is an associate editor of *AI Journal*. He's the secretary of the Association for Uncertainty in Artificial Intelligence and is a fellow of the AAAI. Contact him at poole@cs.ubc.ca.

Clinton Smyth is president of Georeference Online Ltd., a private software development and earth sciences consulting company, and vice president of exploration for Durango Capital Corp., a public minerals-exploration company. He's active in the development of ontologically based software systems for problem solving in the earth sciences and in exploration for copper and gold. Clinton received his MSc in computer science from Imperial College London and his MSc in geochemistry from the University of Cape Town. He's a member of the Society of Economic Geologists and the Geological Society of South Africa. Contact him at cpsmyth@georeferenceonline.com.

Rita Sharma is a research scientist with Georeference Online Ltd., a private software development and earth sciences consulting company. Her main research interest is AI, including inference and learning in probabilistic graphical models (Bayesian networks), Semantic Web technologies, planning and decision making, machine learning, and pattern recognition. Sharma received her PhD in computer science from the University of British Columbia. Contact her at rita@georeferenceonline.com.

References

1. J. Hendler, "Science and the Semantic Web," *Science*, vol. 299, no. 5606, 2003, pp. 520–521; www.sciencemag.org/cgi/content/full/299/5606/520?ijkey=1BUgJQXW4nU7Q&keytype=ref&siteid=sci.
2. P. Fox et al., "Semantically Enabled Large-Scale Science Data Repositories," *Proc. 5th Int'l Semantic Web Conf.* (ISWC 06), LNCS 4273, Springer, 2006, pp. 792–805; www.ksl.stanford.edu/KSL_Abstracts/KSL-06-19.html.
3. D. Poole, C. Smyth, and R. Sharma, "Semantic Science: Ontologies, Data and Probabilistic Theories," *Uncertainty Reasoning for the Semantic Web I*, P.C. da Costa et al., eds., LNAI/LNCS 5327, Springer, 2008; www.cs.ubc.ca/spider/poole/papers/SemSciChapter2008.pdf.
4. L. Getoor and B. Taskar, eds., *Introduction to Statistical Relational Learning*, MIT Press, 2007.
5. L. De Raedt et al., eds., *Probabilistic Inductive Logic Programming*, Springer, 2008.
6. E.T. Jaynes, *Probability Theory: The Logic of Science*, Cambridge Univ. Press, 2003; http://omega.albany.edu:8008/JaynesBook.html.
7. C. Howson and P. Urbach, *Scientific Reasoning: The Bayesian Approach*, 3rd ed., Open Court, 2006.
8. P.F. Patel-Schneider, P. Hayes, and I. Horrocks, *OWL Web Ontology Language: Semantics and Abstract Syntax*, World Wide Web Consortium (W3C) Recommendation, Feb. 2004; www.w3.org/TR/owl-semantics.
9. B. Smith, "Ontology," *Blackwell Guide to the Philosophy of Computing and Information*, L. Floridi, ed., Blackwell, 2003, pp. 155–166; http://ontology.buffalo.edu/smith/articles/ontology_pic.pdf.
10. T. Lukasiewicz, "Expressive Probabilistic Description Logics," *Artificial Intelligence*,

vol. 172, nos. 6–7, 2008, pp. 852–883.
11. P.C.G. da Costa, K.B. Laskey, and K.J. Laskey, "PR-OWL: A Bayesian Ontology Language for the Semantic Web," *Proc. ISWC Workshop Uncertainty Reasoning for the Semantic Web*, 2005; http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS//Vol-173.
12. K. Popper, *The Logic of Scientific Discovery*, Basic Books, 1959.
13. C. Dilworth, "On Theoretical Terms," *Erkenntnis* [Cognition], vol. 21, no. 3, 1984, pp. 405–421.
14. B. Smith, "The Logic of Biological Classification and the Foundations of Biomedical Ontology," *Invited Papers from the 10th Int'l Conf. Logic Methodology and Philosophy of Science*, Elsevier North-Holland, 2003, pp. 190–201; http://ontology.buffalo.edu/bio/logic_of_classes.pdf.
15. Aristotle, *Categories*, E.M. Edghill, trans.; www.classicallibrary.org/aristotle/categories.
16. M.R. Gillespie and M.T. Styles, *BGS Rock Classification Scheme, Vol. 1: Classification of Igneous Rocks*, research report RR 99-06, British Geological Survey, 1999; www.bgs.ac.uk/bgsrcs.
17. L. Struik et al., *A Preliminary Scheme for Multihierarchical Rock Classification for Use with Thematic Computer-Based Query Systems*, Current Research 2002-D10, Geological Survey of Canada, 2002; http://daks.ucdavis.edu/~ludaesch/289F-SQ06/handouts/GSC_D10_2002.pdf.
18. S. Richard et al., "Lithology Categories Vocabulary," SEE GRID community, 2008; www.seegrid.csiro.au/twiki/bin/view/CGIModel/LithologyCategories.
19. C. Boutilier et al., "Context-Specific Independence in Bayesian Networks," *Proc. 12th Ann. Conf. Uncertainty in Artificial Intelligence* (UAI 96), Morgan Kaufmann, 1996, pp. 115–123.
20. J. Łukasiewicz, "On Three-Valued Logic" (in Polish), *Ruch Filozoficzny* [Philosophical Movement], vol. 5, 1920, pp. 170–171. English translation in *Jan Łukasiewicz Selected Works*, L. Borkowski, ed., North-Holland and Polish Scientific Publishers, 1970.
21. R. Sharma, D. Poole, and C. Smyth, *A Framework for Ontologically Grounded Probabilistic Matching*, tech. report, Dept. of Computer Science, Univ. of British Columbia, 2008; http://cs.ubc.ca/~poole/papers/ProbMatching.pdf.

For more information on this or any other computing topic, please visit our Digital Library at www.computer.org/csdl.