Ontology Design for Scientific Theories That Make Probabilistic Predictions

Article *in* Intelligent Systems, IEEE · January 2009

DOI: 10.1109/MIS.2009.15 · Source: DBLP

3 authors:

David Poole
University of British Columbia - Vancouver
238 PUBLICATIONS **9,070** CITATIONS
SEE PROFILE

Clinton Smyth
Georeference Online Ltd
18 PUBLICATIONS **110** CITATIONS
SEE PROFILE

Rita Sharma
FINCAD
11 PUBLICATIONS **90** CITATIONS
SEE PROFILE

Ontology Design for Scientific Theories that make Probabilistic Predictions

David Poole
Department of Computer Science,
University of British Columbia,
http://www.cs.ubc.ca/spider/poole/

Clinton Smyth
Georeference Online Ltd.,
http://www.georeferenceonline.com/

Rita Sharma
Georeference Online Ltd.,
http://www.cs.ubc.ca/spider/rsharma/

December 11, 2008

Abstract

Scientific theories that make predictions about observable quantities can be evaluated by their fit to existing data and can be used for predictions on new cases. Our goal is to publish such theories along with observational data and the ontologies needed to enable the inter-operation of the theories and the data. This paper is about designing ontologies that take into account the defining properties of classes. We show how a multi-dimensional design paradigm, based on Aristotelian definitions, is natural, can easily be represented in OWL, and can provide random variables which provide structure for theories that make probabilistic predictions. We show how such ontologies can be the basis for representing observational data and probabilistic theories in our primary application domain of geology.

1 Introduction

Imagine having a number of expert systems that provide predictions, e.g., diagnoses of what is wrong with a patient 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? If the systems were chosen by current search engines[1], the most popular would be chosen. Other recommender systems base their predictions on some measure

[1] Apparently many of Google's queries are people typing in symptoms and wanting diagnoses. Goggle's ranking system, based on pagerank, essentially measures popularity. These diagnostic queries typically result in authoritative sites.

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). This paper is about part of delivering this vision; how to define ontologies to enable (probabilistic) predictions on observational data and new cases.

The work on expert systems which peaked in the 1980's has given rise to two seemingly separate fields; one of which is concerned with uncertainty and (statistical) learning that typically uses features or random variables. The other concentrates on ontologies and rich representations with individuals and relations, but has essentially ignored uncertainty. This paper is part of an endeavour to put these together, building on the advances in both. As part of putting them together, we don't expect to incorporate all of the sophistication that has been developed in either field.

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 [Hendler, 2003; Fox et al., 2006].

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 ontologies that describe the vocabulary used by the data and the theories. The ontologies need to define the vocabulary needed to express application domains and the vocabularies of data and theories themselves. By publishing ontologies, data and theories, new data can be used 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 paper 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 adding one more desiderata to the design of ontologies, namely taking into account future used 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 [Getoor and Taskar, 2007; De Raedt et al., 2008]. 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 is available for scrutiny. The theories and the data refer to formal ontologies to allow for semantic inter-operability. We expect to have the highest standards for evaluation of theories, with declarations of which data was used for training, and 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 [Jaynes, 2003; Howson and Urbach, 2006]. 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, one in minerals exploration and one in geohazards (predicting landslides). 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 [Patel-Schneider et al., 2004]. We see OWL as an "assembly language" for ontologies. This paper 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.

2 Semantic Science Overview

The purpose of semantic science is to have machine accessible scientific knowledge. There is much knowledge that a scientist has that could be usefully represented; we concentrate on representing data about observations of the world and theories that make predictions on data [Poole et al., 2008]. This semantic science framework has three main components:

- Ontologies that specify the meaning of the vocabulary. These evolve slowly and are built by communities. Through a process of natural selection, we expect that a particular community will converge on useful ontologies that inter-operate. For example, the geology community is actively working on what symbols to use for rocks, minerals, etc.[2] Shared ontologies are important for semantic interoperability.
- 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 it can be recognized when different data sets are about the same or related phenomenon. For example, in the geology domain, the observations may 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[3] make predictions about (potentially) observable features or outcomes. In particular, the theories specify what data they can make predictions about, make predictions that can be checked against all of the relevant data, and can be applied to new cases. We expect these theories to make probabilistic predictions [Jaynes, 2003; Howson and Urbach, 2006]. Again, these probabilistic

[2] See http://onegeology.org/, http://www.cgi-iugs.org and http://www.seegrid.csiro.au for international efforts to share information, and the development of standardized vocabulary.

[3] These 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 they are is redundant as we can access the relevant data to determine how much they should be believed. There, of course, may be other reasons to distinguish these terms.

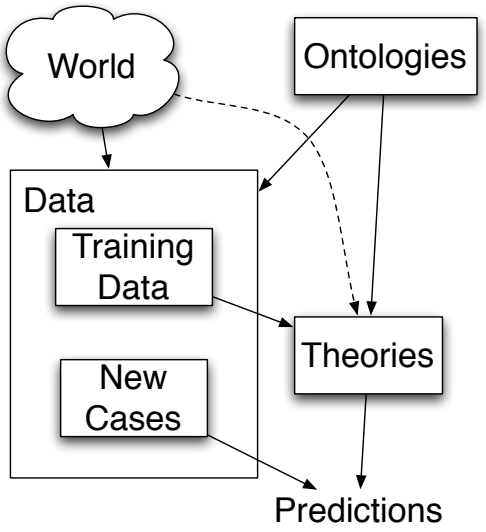

Figure 1: Ontologies, Data and Theories in Semantic Science

models refer to ontologies. For example, we are developing theories in the geology domain that make predictions on where minerals can be found, and theories 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 inter-operation of the data and the theories.

Figure 1 shows the relationship between ontologies, data and theories. The data depends 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 (as we would expect that the best theories would be based on as much data as possible). Given a new case, a theory can be used to make a prediction. The real situation is more complicated, as there are many theories, many ontologies, and lots of 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 both for theories that make multiple predictions (but we have to be careful not to judge a prediction on data it was trained on). Often a prediction will rely on multiple theories (e.g., in a diagnostic situation, there may 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 may 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 someones 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.

3 Ontologies for Semantic Science

In philosophy, *ontology* is the study of existence.

In AI, an *ontology* [Smith, 2003a] is a specification of the meaning of the symbols in an information system. In particular, an ontology contains commitment to what kinds of individuals and relationships are being modelled, 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 the term. It may specify that buildings are human-constructed artifacts, or it may give some restriction on the size of a building so that shoe boxes cannot be buildings or that cities cannot be buildings. It may 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 below.

An ontology written in a language such as OWL [Patel-Schneider et al., 2004] specifiesthe 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 out into 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 (in an information system such as a book or a computer) and concepts. In particular, an ontology should facilitate the following tasks:

- 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, allow inference or determine that some combination of values is inconsistent.
- to facilitate the construction of a hypothesis space over which someone can put a probability distribution. Integrating this task with the other tasks is the subject of this paper.

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 distinctions needed in the domain of interest.

This paper takes a different perspective on the role of the ontology and uncertainty formalisms from many other recent proposals [Lukasiewicz, 2008; da Costa et al., 2005]. 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 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 it is asserted or hypothesized that a coin is fair.

There are a number of reasons that the ontologies should not contain the probabilities about the domain, even though the theories may be probabilistic:

- 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 comes explicitly or implicitly with an ontology; otherwise it is 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 may reinterpret some data with a different ontology, and we have to be careful not to double count that as evidence.

- Data can't be used to falsify an ontology. For example, if some data adheres to an ontology that specifies that a *gneiss* as a metamorphic rock, then by definition all of the gneiss are metamorphic rocks so that data cannot refute that fact. However scientific theories need to be refutable [Popper, 1959]. 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.

- 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, a community cannot, and we argue should not, agree on the probabilities, as people may 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, as then we need to map between these different ontologies. We do want to update theories when new evidence becomes available.

- 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 they can be compared.

- The structure of the information a prediction depends on does not necessarily follow the structure of the ontology. 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 depends on particular facts of the case and not just on other parts of ontologies (e.g., whether they have another form of cancer and whether they worked in a bar that allowed 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 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 [see e.g., Dilworth, 1984]. People can use what ever 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.

4 Representations of Ontologies

Modern ontology languages such as OWL [Patel-Schneider et al., 2004] 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.

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 only consider this case; the specification of classes that would otherwise be specified by just stating what class(es) they are immediate subclasses of.

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

An Aristotelian definition [Smith, 2003b] of class A is of the form "An A is a B such that C", where B is a super-class of A and C is a condition that defines how A is special amongst the subclasses of B. Aristotle [350 B.C.] called the B the *genus* and C the *differentia*. 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 *multi-dimensional 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 super-class conjoined with the restrictions on the values of the properties defining the dimensions. Thus in the multi-dimensional 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 multi-dimensional 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 [Gillespie and Styles, 1999]. However, these produce taxonomies that are difficult to use because they have to commit to a 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 [Struik et al., 2002]. 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.

Richard [2008] defines nine dimensions in which to distinguish earth materials. One dimension is *consolidation degree* that 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 a rock. This dimension only makes sense for rocks (i.e., 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 [2008] defines 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).

When there are rocks that have similar descriptions (such as *gneiss*s and *schist*s), 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 size and weight may appear for physical individuals, but are not applicable for abstract concepts. This idea can be traced back to Aristotle:

> "If genera are different and co-ordinate, 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'." [Aristotle, 350 B.C.]

Note that "co-ordinate" means that neither is subordinate of the other.

4.1 Multi-dimensional Ontology Assumptions

In this section we will be more formal in the assumptions behind multi-dimensional ontologies. We do this to show how the multi-dimensional 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 multi-dimensional ontology:

- Dimensions are defined by functional properties or by each value of a non-functional property.
- Classes are defined in terms of values on properties.
- The domain of a property that defines a dimension is the most general class on which the property makes sense.

Assuming that subclasses are only defined 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 paper is arguing that there are advantages in explicitly representing the predicates that define classes.

In particular, we make the following assumptions about the ontology:

- The top class, *Thing* is predefined.
- Classes are either:
  - *enumeration classes* that are predefined sets of values. For example, in geology, *FabricTypeValue* could be defined as the set of constants $\{aplitic, biogenic, foliated, \ldots\}$.
  - *non-enumeration classes*, that are made up of individuals in the world of the application domain, 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.
- 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.

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.[4]

Under this interpretation, a non-enumeration 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 (sub)classes only in terms of properties has a number of advantages over trying to specify the subclass relation directly (or even trying to impose a tree structure over the abstraction hierarchy):

- It is easy to specify, compute and explain subclasses in terms of the dimensions, even though the induced subclass relationship may be very complex to depict.
- A concept does not need to specify values for all dimensions. Overlapping concepts can specify values for different sets of dimensions.
- It is often difficult to decide on which attribute to split a hierarchy. Different splits may be applicable for different purposes. The multi-dimensional splitting means that you don't have to make this (often arbitrary) choice.
- It is important for probabilistic reasoning where the dimensions create random variables (see Section 5). This provides a way to have probabilistic models (and utility models) over complex objects described using complex ontologies.

4.2 OWL and the Multi-dimensional Design Pattern

OWL [Patel-Schneider et al., 2004] is the W3C recommendation for representing ontologies. It 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 multi-dimensional design pattern. It is interesting to note that none of the tutorials or material used for teaching or learning OWL we could find use that 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 non-enumeration class (i.e., the value is an individual in the world).

The dimensions of a multi-dimensional ontology are defined in terms of discrete properties.

Example 2 Consider representing a *gneiss*, as outlined in Example 1 using a multi-dimensional ontology in OWL. Suppose we already have *rock* defined (as an earth material with consolidation degree of *consolidated*). We need to say a *gneiss* is a rock with genesis of metamorphic, fabric type is *foliated*, the particle type is *crystal*, and the grain size is *phaneritic*. To represent this we do the following:

[4] Suppose there is a cyclic set of definitions: $A_1 \equiv A_0 \wedge C_1, A_2 \equiv A_1 \wedge C_2, \ldots A_k \equiv A_{k-1} \wedge C_k, A_0 \equiv A_k \wedge C_0$. 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 are a set of equivalent classes, this can represented as having a canonical representation for the classes.

- We create a class *FabricTypeValue* that is an enumeration class that is equivalent to the collection {*aplitic*, *biogenic*, *foliated*}. (We just show these three values to keep it simple; Richard [2008] used 6 values).

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

- We create a functional property *fabricType*, whose domain is *Rock* and whose range is *FabricTypeValue*.:

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

Figure 2: OWL functional specification of the fabric dimension

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

Figure 3: Defining the class *Gneiss* in OWL functional syntax

- We construct a functional property *fabricType* whose domain is *Rock* and whose range is {*aplitic*, *biogenic*, *foliated*, ...}. This is shown in Figure 2.
- Similarly, we create functional properties for *geneticCategory*, *particleType* and *grainSize*, each with the domain *EarthMaterial* or *Rock* as appropriate, and a range which is an enumeration class[5].
- We define *Gneiss* as a rock with the appropriate values on the properties. This is shown in Figure 3.

We claim that this multi-dimensional 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.

5 Ontologies to possible worlds and random variables

Although we have tried to argue that the multi-dimensional ontology is important in its own right, a main motivation is to use it as a foundation for specifying probabilis-

[5]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 paper.

tic models. The general idea is that the dimensions form random variables for each individual. In this section, we first define the possible worlds, then a notion of random variables that have existence prerequisites, and finally we specify sufficient and necessary conditions to ensure that conditional probabilities are well-defined.

For this paper, 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 domain 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 only defined if the individual is of the type of the domain of the property that defines the variable. Thus the existence of some random variable may 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 only defined 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 when 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 implicitly conjoining that it is is a rock. This is reminiscent of context-specific independence [Boutilier et al., 1996], 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 only refer to a random

variable in a context where it is defined. We can define what individual-variable-value triples are defined and what value they have in each world, procedurally. Each different choice in the description below will give a different possible world:

- 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):
  - 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 $\langle i,p,v'\rangle$ 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 above procedure is undefined.

We interpret any formula made up of values of global variables and of individual-property-value triples, using the standard logical connectives, but with a third truth value, undefined ($\bot$), interpreted as follows: $\bot op \bot \equiv \bot$, for any operation op, $true \wedge \bot \equiv \bot$, $false \wedge \bot \equiv false$, $true \vee \bot \equiv true$, $false \vee \bot \equiv \bot$, $\neg\bot \equiv \bot$. This logic was first introduced by Łukasiewicz in 1920.

For example, $\langle i_7, fabricType, foliated\rangle$ will have value $\bot$ 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 an 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 $\bot$ 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 value $\bot$ in that possible world, and so the conditional probability is not well defined.

For example, the following is not well defined:

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

as the conditions do not imply that i is in the domain of *fabricType*, 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. While it is possible to reify such statements [da Costa et al., 2005] so 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.

6 An Example In Geology

In geology, geological surveys publish descriptions of the geology of various locations in the jurisdictions. The 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 multi-dimensional 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[6]. For example, La Esperanza is a mineral occurrence in Argentina. Part of its description, in a functional syntax, is:

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

The mineral occurrence is hosted in two rocks, one of which is a schist and the other is a gneiss.

One model that can make a prediction on this mineral deposit is the USGS Thorium-Rare-Earth Vein (TREV) Model[7]. Here we explain a small part of this model, shown in Figure 4. 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 give the roles assigned by the parents. The roles are implicit in the diagram. The syntax and semantic is described by Sharma et al. [2008].

[6]Our application does not use RDF triples as we also want to be able to specify non-existence, that no object in the world satisfies some description. While this can be represented in RDF be reifying the statement,

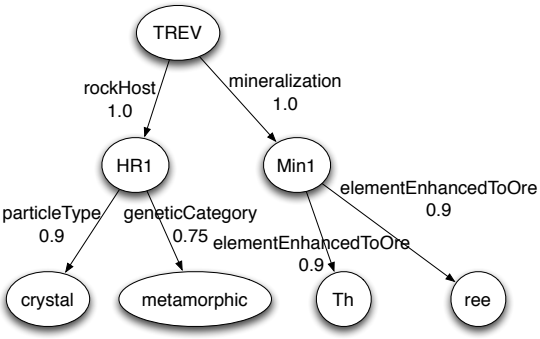

Figure 4: Part of a Thorium-Rare-Earth Vein (TREV) Model

We will not explain our syntax, but note that we do not allow arbitrary probabilities of first order formulae, but have a simple language that gives a naive Bayes model of the existence of objects that fill roles, and properties of these objects.

This represents conditional probabilities of the form:

$$P(\exists HR1\ r_1(HR1) \wedge rockHost(TREV, HR1) \\ |thoriumRareEarthModel(TREV)) = 1.0$$
$$P(particleType(HR1) = crystal | r_1(HR1) \wedge rockHost(TREV, HR1) \\ \wedge thoriumRareEarthModel(TREV)) = 0.9$$
$$P(geneticCategory(HR1) = metamorphic | r_1(HR1) \wedge rockHost(TREV, HR1) \\ \wedge thoriumRareEarthModel(TREV)) = 0.75$$
$$P(elementEnhancedToOre(Min1) = Th | r_2(Min1) \\ \wedge mineralization(TREV, Min1) \wedge thoriumRareEarchModel(TREV)) = 0.9$$

where r_1 is true of the object that satisfied the role represented by the node labeled $HR1$. r_2 is true of the object that satisfied the role represented by the node labeled $Min1$.

From the complete model, we wish to compute the probability that La Esperanza will have ore-grade Thorium. What is important for this paper 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 prediction 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. Multi-dimensional ontologies are a good candidate for this.

---

it isn't very natural.

[7] http://pubs.usgs.gov/bul/b2004/html/bull2004thorium_rareearth_veins.htm

7 Conclusion

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 multi-dimensional 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 modelling.

This interaction between ontologies and probabilistic reasoning forms the foundations of applications we are building in minerals exploration and landslide prediction. This paper only considers one aspect of the problem. Another aspect is, given descriptions of theories and individuals in the world at various levels of abstraction and detail, to use them to make coherent decisions. Another 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 (i.e., which individuals fill the roles in the model). We also need to model and reason about existence and non-existence. These are ongoing research topics that build on the foundations given in this paper.

With respect to other efforts on the semantic web, semantic science seems to be an area where the bootstrapping problem may 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 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.

References

Aristotle. *Categories*. Translated by E. M. Edghill, http://www.classicallibrary.org/Aristotle/categories/, 350 B.C.

C. Boutilier, N. Friedman, M. Goldszmidt, and D. Koller. Context-specific independence in Bayesian networks. In E. Horvitz and F. Jensen, editors, *UAI-96*, pages 115–123, Portland, OR, 1996.

P. C. G. da Costa, K. B. Laskey, and K. J. Laskey. PR-OWL: A Bayesian ontology language for the semantic web. In *Proceedings of the ISWC Workshop on Uncertainty Reasoning for the Semantic Web*, Galway, Ireland, Nov 2005. URL http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS//Vol-173/.

L. De Raedt, P. Frasconi, K. Kersting, and S. H. Muggleton, editors. *Probabilistic Inductive Logic Programming*. Springer, 2008.

C. Dilworth. On theoretical terms. *Erkenntnis*, 21(3):405–421, 1984.

P. Fox, D. McGuinness, D. Middleton, L. Cinquini, J. Darnell, J. Garcia, P. West, J. Benedict, and S. Solomon. Semantically-enabled large-scale science data repositories. In *5th International Semantic Web Conference (ISWC06)*, volume 4273 of *Lecture Notes in Computer Science*, pages 792–805. Springer-Verlag, 2006. URL http://www.ksl.stanford.edu/KSL_Abstracts/KSL-06-19.html.

L. Getoor and B. Taskar, editors. *Introduction to Statistical Relational Learning*. MIT Press, Cambridge, MA, 2007.

M. R. Gillespie and M. T. Styles. BGS rock classification scheme, volume 1: Classification of igneous rocks. Research Report (2nd edition) RR 99-06, British Geological Survey, 1999. URL http://www.bgs.ac.uk/bgsrcs/.

J. Hendler. Science and the semantic web. *Science*, 299(5606):520 – 521, January 2003. URL http://www.sciencemag.org/cgi/content/full/299/5606/520?ijkey=1BUgJQXW4nU7Q&keytype=ref&siteid=sci.

C. Howson and P. Urbach. *Scientific Reasoning: the Bayesian Approach*. Open Court, Chicago, Illinois, 3rd edition, 2006.

E. T. Jaynes. *Probability Theory: The Logic of Science*. Cambridge University Press, 2003. URL http://omega.albany.edu:8008/JaynesBook.html.

T. Lukasiewicz. Expressive probabilistic description logics. *Artificial Intelligence*, 172 (6-7):852–883, 2008.

P. F. Patel-Schneider, P. Hayes, and I. Horrocks. OWL web ontology language: Semantics and abstract syntax. W3C Recommendation 10 February 2004, W3C, February 2004. URL http://www.w3.org/TR/owl-semantics/.

D. Poole, C. Smyth, and R. Sharma. Semantic science: Ontologies, data and probabilistic theories. In P. C. da Costa, C. d'Amato, N. Fanizzi, K. B. Laskey, K. Laskey, T. Lukasiewicz, M. Nickles, and M. Pool, editors, *Uncertainty Reasoning for the Semantic Web I*, LNAI/LNCS. Springer, 2008. URL http://www.cs.ubc.ca/spider/poole/papers/SemSciChapter2008.pdf.

K. Popper. *The Logic of Scientific Discovery*. Basic Books, New York, NY, 1959.

S. Richard. Vocabulary of lithology categories for review and comment. SEE GRID community website, 25 June 2008. URL https://www.seegrid.csiro.au/twiki/bin/view/CGIModel/LithologyCategories.

R. Sharma, D. Poole, and C. Smyth. A framework for ontologically-grounded probabilistic matching. *International Journal of Approximate Reasoning*, submitted, 2008.

B. Smith. Ontology. In L. Floridi, editor, *Blackwell Guide to the Philosophy of Computing and Information*, pages 155—166. Oxford: Blackwell, 2003a. URL http://ontology.buffalo.edu/smith/articles/ontology_pic.pdf.

B. Smith. The logic of biological classification and the foundations of biomedical ontology. In D. Westerståhl, editor, *Invited Papers from the 10th International Conference in Logic Methodology and Philosophy of Science*. Elsevier-North-Holland, Oviedo, Spain, 2003b. URL http://ontology.buffalo.edu/bio/logic_of_classes.pdf.

L. Struik, M. Quat, P. Davenport, and A. Okulitch. A preliminary scheme for multihierarchical rock classification for use with thematic computer-based query systems. Current Research 2002-D10, Geological Survey of Canada, 2002. URL http://daks.ucdavis.edu/~ludaesch/289F-SQ06/handouts/GSC_D10_2002.pdf.

Author Biographies

David Poole is a Professor of Computer Science at the University of British Columbia. He has a Ph.D. from the Australian National University. He is 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 is a coauthor of a forthcoming AI textbook (Cambridge University Press, 2009), a coauthor of an older AI textbook, Computational Intelligence: A Logical Approach (Oxford University Press, 1998), and co-editor of the Proceedings of the Tenth Conference in Uncertainty in Artificial Intelligence (Morgan Kaufmann, 1994). He is former associate editor and on the advisory board of the Journal of AI research, is an associate editor of AI Journal. He is the secretary of the Association for Uncertainty in Artificial Intelligence, and is a Fellow of the Association for the Advancement Artificial Intelligence (AAAI).

Clinton Smyth is President of Georeference Online Ltd, a private software development and earth sciences consulting company, and Vice President (Exploration) for Durango Capital Corp, a public minerals exploration company. He is active in the development of ontologically-based software systems for problem-solving in the earth sciences, and in exploration for copper and gold. Clinton has an MSc in Computer Science from Imperial College (London), and an MSc in Geochemistry from the University of Cape Town. He is a member of the Society of Economic Geologists, and of the Geological Society of South Africa.

Rita Sharma is Research Scientist with Georeference Online Ltd, a private software development and earth sciences consulting company. Rita has PhD and MSc in Computer Science from the University of British Columbia, Vancouver Canada. Her main research interests are artificial intelligence including Inference and learning in probabilistic graphical models (Bayesian Networks), semantic web technologies, planning and decision-making, machine learning, and pattern recognition.