SKIing with DOLCE: toward a Science Knowledge Infrastructure for e-Science

Boyan Brodaric[1], Femke Reitsma[2], Yi Qiang[2]
[1]Geolological Survey of Canada, Ottawa, brodaric@nrcan.gc.ca
[2]UK National e-Science Center, University of Edinburgh, femke.reitsma@ed.ac.uk, yi_qiang@hotmail.com

Abstract

We develop a general ontology of science knowledge for use in e-Science Knowledge Infrastructures (SKIo), to advance use of digitally represented scientific theories in these environments. SKIo extends the DOLCE foundational ontology with science knowledge primitives, such as science theories, models, and data. These are arranged to reflect the complex knowledge structures used in science, such as scientific ideas playing different roles within and between theories. SKIo is illustrated in UML, encoded in OWL-DL, uses the Descriptions and Situations extension, and provides defining conditions for its primitives to enable an extensible and rigorous bridge between a foundational ontology and domain science ontologies. Testing with several environmental theories confirms the suitability of its representation.

1. Introduction

Infrastructures in support of cyber-aided scientific activity, or e-Science, are being developed in many scientific domains (Hey & Trefethen, 2005). This is leading to significant scientific and societal benefits, in that larger and faster computations are occurring over more data, and the resultant predictive models are providing larger and more accurate scenarios about situations affecting human well-being. Although these early e-Science achievements are laudable and significant, they do fall short of a broader e-Science vision in which scientists not only operate over more observed data to make better predictive models, but also directly use e-Science infrastructures to find, generate and test scientific theories (i.e. networks of scientific abstractions, associated implications, and certainties). The broader vision of e-Science requires a Scientific Knowledge Infrastructure (SKI; Hars, 2001) that enables the capture, representation and use of the full spectrum of scientific knowledge for e-Science. Using SKIs, scientists should be able to annotate existing resources, such as observed data, predicted models, and ancillary products, with respect to relevant and potentially competing scientific theories, in order to enable knowledge search, knowledge evaluation, and reproduction of experimental results.

The present focus on a partial set of science knowledge primitives has some negative consequences as it deters full scientific discovery and reproducibility in e-Science infrastructures—because only some knowledge is explicitly represented in the infrastructure, while other knowledge is implicit and buried in scientists' heads and in ancillary resources such as textbooks, papers, reports and maps. An initial challenge then is the development of a formal and computable representation of a wide segment of science knowledge primitives. Foundational ontologies in particular are good candidate for representing such knowledge, not only due to their

formality, rigor and commitment to internal coherence, but also due to their generality in that, like scientific knowledge primitives, the contents of foundational ontologies are intended to be re-used across many (all) science domains. This contrasts with the numerous ontologies being developed for specific science domains, and is aligned with the few that are being developed as a general upper superstructure, but these latter general efforts are addressing only a subset of science knowledge primitives or they do not extend an existing foundational ontology, which limits their breadth of application.

In this paper we extend the DOLCE foundational ontology in support of e-Science with basic science knowledge primitives, such as scientific models, theories, data, and test the resultant SKI ontology (SKIo) by representing several environmental theories. Section 2 describes a typical use-case scenario in the environmental sciences; Section 3 discusses related work; Section 4 discusses our general approach based on computationally inspired renditions of the science knowledge cycle, Section 5 describes SKIo; the results of using SKIo to represent environmental theories and models is outlined in Section 6. Section 7 relays some limitations of SKIo, and Section 8 concludes with a brief summary.

2. An Environmental Modeling Use-Case Scenario

Problem Scenario: Jane is a scientist working on extending a global climate model. She wants to integrate a terrestrial Carbon exchange model into her larger global climate model.

Present Day Solution: Jane begins searching for carbon exchange models using a few keywords in Google as well as in her University library database. She comes up with a huge number of hits and begins the long process of sifting through these, many of which are irrelevant. After much work she has a large collection of papers that seem relevant but cannot be easily differentiated, largely because they use polysemous terminology. For example the key term "model" is used in several senses in the Carbon exchange literature, of which the first two are relevant to her:

1. model = a system of equations to support calculations that generate simulations (Adams, *et al*., 2004)
2. model = a theory which might include equations as well as scientific implications (Baldocchi, *et al*., 2000)
3. model = a simulation software that utilizes a system of equations and implications (Phillips, *et al*., 2006)
4. model = the results of a simulation run, or other modelling process, in which some climatic situation (in a geographical area) is represented (Phillips, *et al*., 2006)

Proposed Solution: Jane logs on to a web-based SKI and begins searching for a relevant "model" by using a number of concepts she is familiar with in Carbon modelling; she expects these concepts to be used as variables in equations. Because the different senses of "model" are well demarcated by the SKI ontology, and because the SKI's contents are annotated by this ontology, she is able to find entities corresponding to senses 1 and 2 that contain the concepts of interest. She soon finds a few candidates, each linked to a set of digital resources. After integrating the newly found model, and running her experiment, she creates a web page documenting this process

and containing a publication draft. She then annotates this resource in the SKI, making it available to other researchers.

Additional Requirements: In addition to aiding Jane as described above, SKIs should also help scientists resolve questions and tasks such as: given problem *p*, has anyone else solved that problem, or a similar problem in another domain? Who is working in the same research field? I want to test a new theory x_2 by replacing existing theory x_1 against a configuration of data *y* as originally written about in journal paper *z*. What other data configurations satisfy theory x_1? What other theories are satisfied by *y*, in which papers, and how do these differ from x_1? How was x_1 derived—what observed data, line of reasoning, and verification procedures were used, in which papers? What other theories is x_1 part of, and what is its role in those theories? What theories have been derived from x_1? What theories could be derived from x_1 that satisfy *y*?

3. Related Work

Although ongoing work on scientific ontologies is vast, and growing, at present a general ontology for science knowledge does not exist. Existing initiatives emphasize the computational representation of the science knowledge cycle or the development of ontologies that span aspects of all sciences, are limited to one science domain, or which incorporate foundational ontologies.

The Science Knowledge Cycle: Several accounts of the scientific knowledge cycle begin to distil the numerous and complex philosophical approaches into representations amenable to computation. The focus is on identification of the key elements in the cycle (Langley, 2000; Shrager & Langley, 1990; Sowa, 2000; 2006; Thagard, 1988), in some cases for schema representation (Hars, 2001; 2003), or for integrating the steps using formal reasoning systems (Ray, 2005), but without adopting ontologies as a formal representation framework.

Ontologies of Science: by science ontologies we mean a conceptualization of general science knowledge primitives that can be applied in many (all) science domains and which are represented, and well defined, in some formal language such First Order Logic or OWL (Antoniou & Van Harmelen, 2003). Existing science ontologies meet this definition only in part because they focus on a fragment of the key knowledge primitives such as scientific experiments (Soldatova & King, 2006) or publications (Benjamins, *et al*., 1998), and omit other key aspects such as theories and models, or they do posses broader contents but without formal representation (Rijgersberg, *et al*., 2008).

Domain Science Ontologies: ontologies are being developed in numerous science domains, such as biosciences, geosciences, environmental and earth sciences (McCray, 2003; Natale, *et al*., 2007; Raskin & Pan, 2005). These ontologies cannot serve as a superstructure for science knowledge because the abstractions are not sufficiently general, They are largely being used to facilitate data interoperability and workflow operation (Ludäscher, *et al.*, 2006; McGuinness, *et al*., 2006), rather than to annotate and test new scientific ideas. Many are organized bottom-up from existing vocabularies (Raskin & Pan, 2005) and not around systematic ontological principles such as those utilized by most foundational ontologies, resulting in diverse ontological assumptions that are not easily recognized nor reconciled.

Foundational Ontologies and Science Knowledge: foundational ontologies provide a superstructure containing the most general abstractions, that can be extended to general science ontologies and domain ontologies, e.g. DOLCE, BFO, GFO, SUMO (Grenon & Smith, 2004; Herre, *et al*., 2006; Masolo, *et al*., 2003; Pease, A. 2006). An ideal arrangement of ontologies would then position a general ontology of science as a layer between foundational ontologies and domain science ontologies (Soldatova & King, 2006), as shown in Figure 1. With the exception of the ontology of experiments (Soldatova & King, 2006), this intermediary layer is at present missing, in that domain science ontologies directly extend from existing foundational ontologies or related foundational theories (Bittner, 2007; Gangemi, *et al*., 2004; Grenon *et al*., 2004).

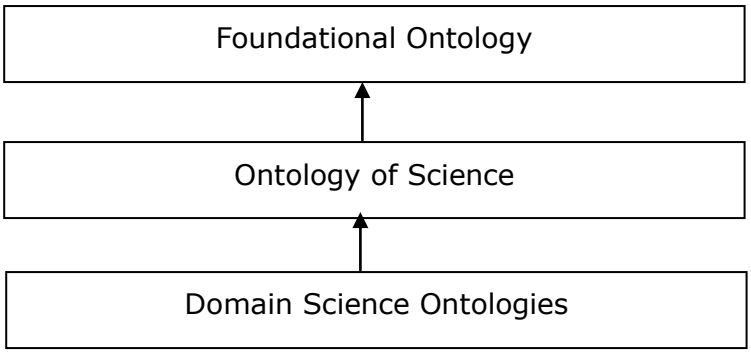


Figure 1: a general science ontology extends a foundational ontology and is extended by domain science ontologies (after Soldatova & King, 2006).

4. Approach

SKIo extends the DOLCE foundational ontology with a modest number of science knowledge primitives synthesized from computational accounts of the science knowledge cycle (Hars, 2003; Sowa, 2006). SKIo is first represented in UML (Brodaric, 2008), and then in OWL-DL. Following OWL terminology conventions, the primitives consist of classes and properties: classes refer to abstractions that can be instantiated in one more individuals, and properties refer to relations between two classes; individuals are single entities that instantiate a class, i.e. instances. Both class and property names will be presented in italics henceforth.

The SKIo extension process involves adding general science classes and properties to the DOLCE hierarchy of classes and properties, using the `subclassOf` and `subpropertyOf` OWL constructs (for subsumption), such that the additions form leaves in these hierarchies, and the original hierarchical structure remains unchanged. This also involves leaving the contents of DOLCE classes largely untouched, i.e. axioms were added to only two existing DOLCE classes. Such modularization enables SKIo to exist as an independent OWL-DL file (with a separate namespace), that can be imported alongside DOLCE as needed.

Several other principles in addition to the modularity principle were to be followed in the design of SKIo (after Gruber 1995):

- ***Semantic Grounding*:** SKIo is to be founded on recognized accounts of the science knowledge cycle that can be formalized for representation purposes.

- ***Semantic Coverage*:** the breadth of the science knowledge cycle is to be encompassed such that SKIo could be extended with both general science and domain extensions.
- *Semantic Precision*: sufficient depth is to be attained to enable annotation of scientific documents through instantiation of SKIo primitives, with a focus on science roles.
- ***Extensibility***: grounding in solid principles of the scientific knowledge cycle is to provide a rigorous and coherent basis for general and domain specific extensions.
- ***Coherence*:** these same principles are to be formalized to enhance definition (and consequently understanding) of SKIo components.

5. The SKI ontology

5.1 The Science Knowledge Cycle

Figure 1 shows some parts of the science knowledge cycle incorporated into SKIo. In this cycle: (1) empirical regularities are induced from observed data, (2) theoretical propositions are abduced from all prior knowledge, (3) predictions about the real world are deduced from empirical patterns or theoretical statements, and (4) predictions are verified through further interaction with the world, which involves activities such as data collection, problem finding, and building models of the world.

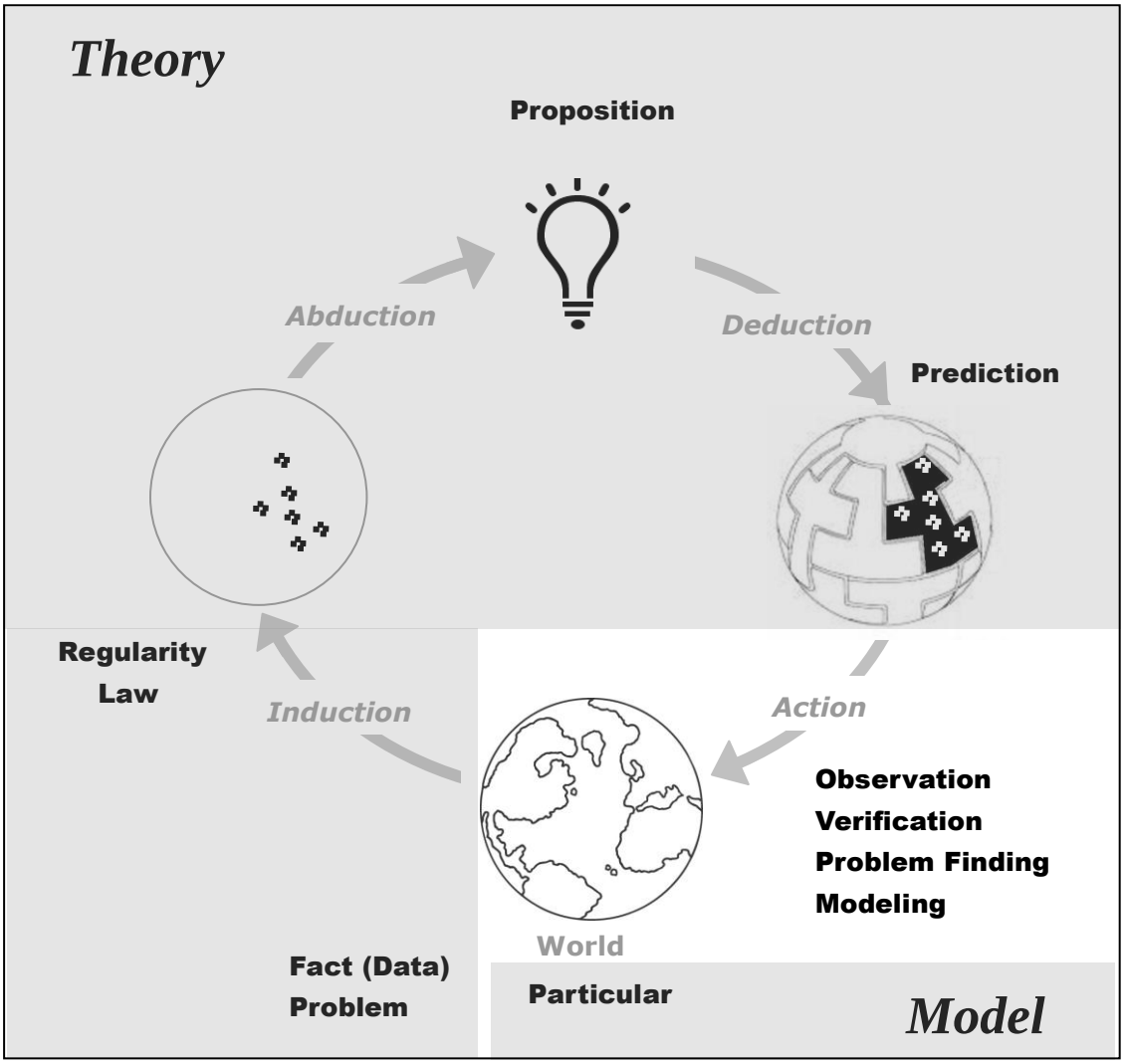


Figure 1: the knowledge cycle in SKI (after Sowa, 2000; 2006).

5.2 SKIing with DOLCE

The DOLCE 2.1 (OWL 397) ontology consists of four core classes which categorize particulars (individual entities in the world): *endurant*, *perdurant*, *quality*, and *abstract*. An *endurant* is an object-like entity that is wholly present at any point in time it exists, but whose characteristics can change over time (rock body, building, country); a *perdurant* is a process-like entity that is not wholly present at any point in time it exists, such as a *process*, *event*, or *state* (San Andreas faulting, San Francisco earthquake, being seismically active); a *quality* is a dependent characteristic inherent in an *endurant*, *perdurant*, or *abstract*, such that an *endurant* inheres physical qualities (geospatial position, size, shape, color), an *perdurant* inheres temporal qualities (duration, age), and an *abstract* or *non-physical-endurant* inheres abstract qualities (the value of the Canadian dollar); an *abstract* is an entity that does not posses physical or temporal qualities, and is often the value of a *quality* or a space containing those values (the number 2, the munsell color space, red).

SKIo extends the *perdurant* and *endurant* DOLCE classes with formalized versions of components of the science knowledge cycle: it extends the *activity* subclass of *perdurant* to include various scientific activities such as those for reasoning, observation, and verification; and it extends some physical and social subclasses of *endurant* (*description*, *situation*, *concept*, *information-object*, *physical-endurant*) to include scientific artifacts. Importantly, in order to foster reproducibility, each SKIo *endurant* is defined according to the scientific activity that produces it.

Descriptions and Situations: DOLCE's descriptions and situations are initially designed to represent intensional socially constructed contexts and the related extensional states-of-affairs interpreted by those contexts, respectively (Gangemi & Mika, 2003). Although descriptions and situations have been used to model specific scientific domain theories and models (Gangemi *et al*., 2004), they are extended by SKIo into general abstractions for science theory, science model, data, and concept definition, as shown in Figure 3, to provide scientific constraints on their meaning. Descriptions are thus viewed as scientific ideas that are syntactically expressed by DOLCE *information-objects*, such as text, tables, figures, reports, papers and web sites; *information-objects* are in turn externally and physically manifested in forms such as hardcopy or computer memories. A science theory is then a scientific idea comprised of one or more coherent descriptions that describe the structure or behavior of some aspect of reality in sufficient generality to satisfy, and be used to predict, a wide number of real-world particulars and the science models that contain them (Hars, 2003). In SKIo, a *ScienceModel* contains particulars which satisfy (are scientifically deducible from) some *ScienceTheory*, a science theory contains descriptions which are satisfied by (can be used to scientifically deduce) the particulars in the science model, and a description contains concepts that classify the particulars in a model. This intent is stated in A1 and implemented in SKIo as a *deducible-by* property on a particular, which takes as its range a *Prediction* indicating the particular is forecast by the prediction.

(A1) *ScienceTheory* (T) $\wedge$ *ScienceModel* (M) $\wedge$ *satisfies* (T, M) $\rightarrow$

$\forall$ (a $\in$ M) $\exists$b (*Deduction* (T, b, a) $\wedge$ *particular* (a) $\wedge$ *particular* (b))

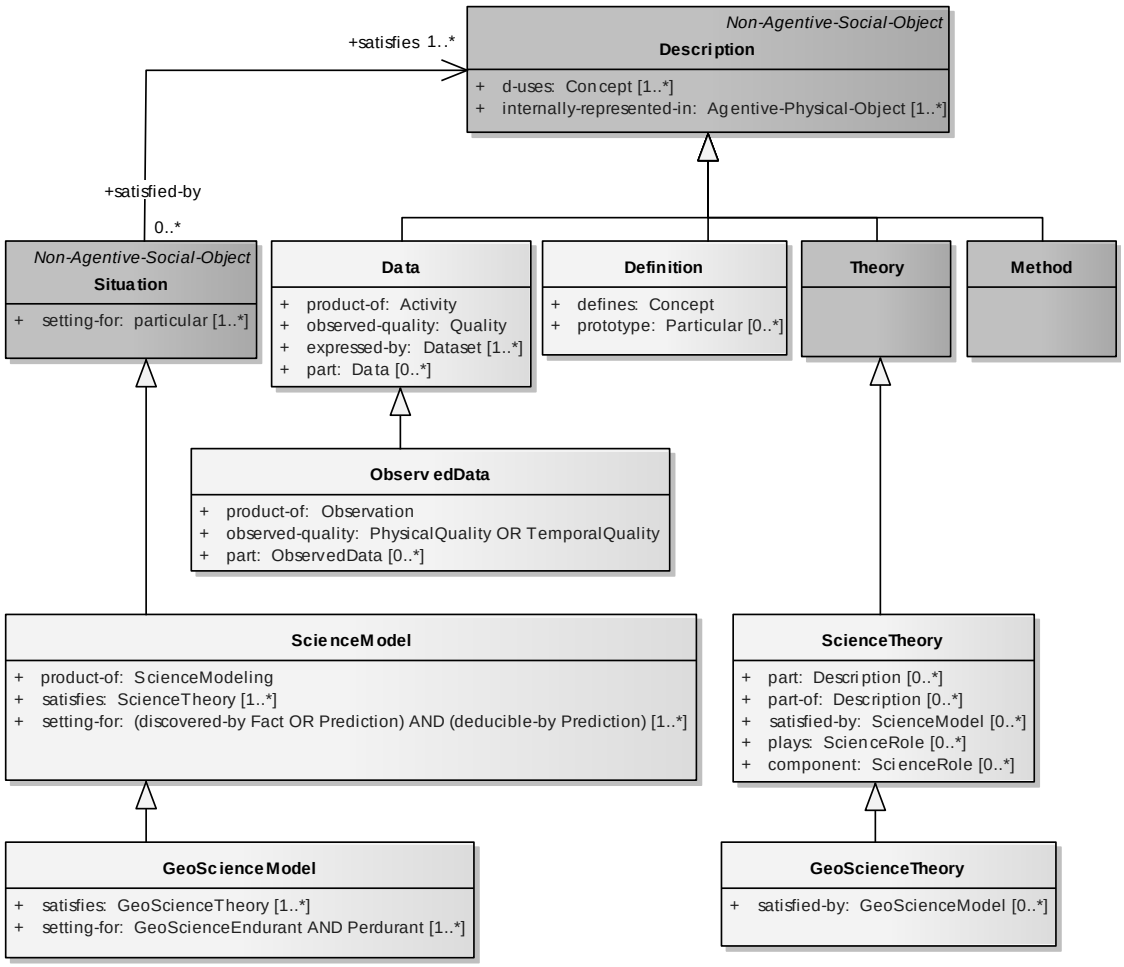


Figure 3: SKIo science model and theory (white) and DOLCE description and situation (grey).

In SKIo, *Data* is also an intensional scientific idea, one that results from the observation or inference of some quality, for some purpose, by some agent in whom the description is internally represented, e.g. in an instrument. Underlying this conceptualization is the assumption that only physical and temporal qualities are observable, and that the inhering physical-endurants and perdurants are then inferred from the observed qualities as part of the science knowledge cycle. *Data* about abstract qualities are also typically inferred in SKIo, while descriptions about the physical or temporal qualities of some *physical-endurant* or *perdurant*, respectively, are obtained via observation (including measurement) and are called *ObservedData*. The syntactic expression of some data is a *DataSet*. SKIo also provides some specializations mainly for convenience: a *GeoscienceModel* contains geoscience endurants and perdurants, and is satisfied by *GeoscienceTheories*; and a *Definition* is a canonical idea included for sake of being explicit—DOLCE descriptions are implicitly definitions.

Concepts: the DOLCE *concept* class is an intensional entity used to classify a particular within a situation to enable it to satisfy a description. DOLCE provides three types of concepts: *role*, *course*, and *parameter*, for classifying endurants, perdurants and quality regions, respectively. Concepts are typically related to descriptions in three ways in SKIo: (1) an atomic description (e.g. a theory part) *d-uses* concepts in its body to describe the idea, (2) atomic descriptions can

play a certain *description-role* concept within aggregate descriptions (e.g. within the theory as a whole), and (3) a description can maintain an index of the *component* concepts that classify situation members or description parts. For example, the theory of special relativity has as a part the idea $e = mc^2$ that: (1) uses *parameter* concepts *energy* (*e*), *mass* (*m*), *constant speed of light* (*c*), (2) plays the *role* concept of *Proposition* within the theory, and (3) the *role* and *parameter* concepts are indexed as components of the theory. In subsequent theories the same idea can play the role of an *Assumption* and a *ScienceProblem* (Kaku, 2004). Figure 4 illustrates the complex relationships between scientific ideas, scientific statements, theories, and roles.

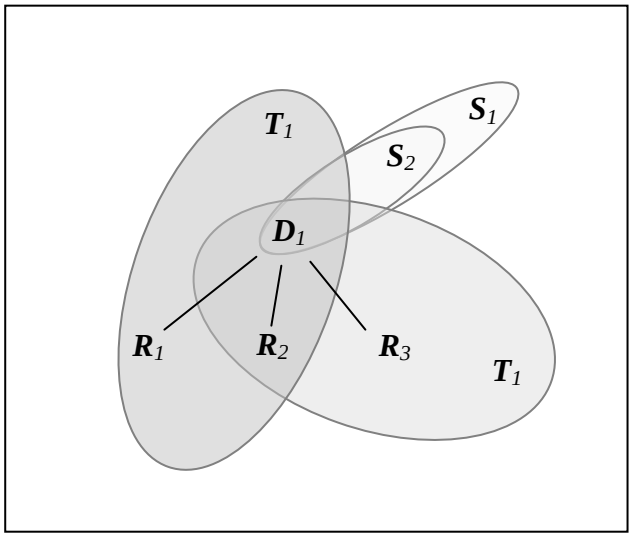


Figure 4: a SKI scenario for relationships between instances of a scientific idea (description $\boldsymbol{D_1}$), two science statements ($\boldsymbol{S_1}$, $\boldsymbol{S_2}$) that express the idea in different papers, two theories ($\boldsymbol{T_1}$, $\boldsymbol{T_2}$) that contain the idea, and the three roles ($\boldsymbol{R_1}$, $\boldsymbol{R_2}$, $\boldsymbol{R_3}$) played by the idea in the theories.

Because the natural language terms for science roles are highly polysemous, SKIo also provides defining conditions in its OWL encoding for six of the seven science roles shown in Figure 5.

- ***Assumption***: is defined by an originating *Assertion*, and is considered to be a primitive such that it is not empirically supported nor inferred.
- ***ScienceProblem***: is defined by an originating *ProblemIdentification* activity operating over a theory part that is disconfirmed by observation or inconsistent with other theory.
- ***Fact***: is the incorporation of some *Data* into some theory. Because facts can only be played by data within some theory, they are always 'theory-laden'. Facts also support specification of the scientific discovery of a particular, insofar as a fact can indicate that some data has lead to identification of the particular, e.g. a rock body.
- ***EmpiricalRegularity***: is defined as an empirical pattern produced by *Induction*, one that is situational but not universal. Situational refers to the case where the regularity is satisfied by only a subset of the possibly valid science models. For example, if the regularity is expressed as a relation amongst concepts, then the relation is present only in some situations in which particulars classified by the concepts are jointly present, and not in all situations where they are jointly present. The regularity might not be present

universally because of insufficient verification or because the pattern is dependent on certain historical conditions that are temporary and change in time. This is implemented in SKIo by requiring the regularity's existence to be dependent on one or more endurants or perdurants, likely some subset of those involved in its original induction.

- ***ScienceLaw***: is defined as an universal empirical pattern produced by *Induction*, i.e. its existence is not dependent on any specific endurants or perdurants. Empirical regularities and science laws can evolve toward each other as a consequence of the logic of induction: a pattern might be scoped as situational but more data might suggest it to be universal, and conversely more data might contradict a science law and demonstrate it to be situational.
- ***Prediction***: is defined as a conjecture about individuals produced via *Deduction*, which can be empirically verified. Because only physical and temporal qualities are observable, it follows only these qualities are predictable. SKIo does not impose this constraint, as it is sometimes convenient as a shorthand to predict individuals as well as qualities.
- ***Proposition***: is a best-guess conjecture produced via *Abduction* from prior theory or data, and which can be situational or universal. Scientific propositions are analogous to logical propositions in that scientific propositions are eventually verifiable (are testable against the world) while logical propositions are eventually resolvable to True or False (are testable against a logical system).

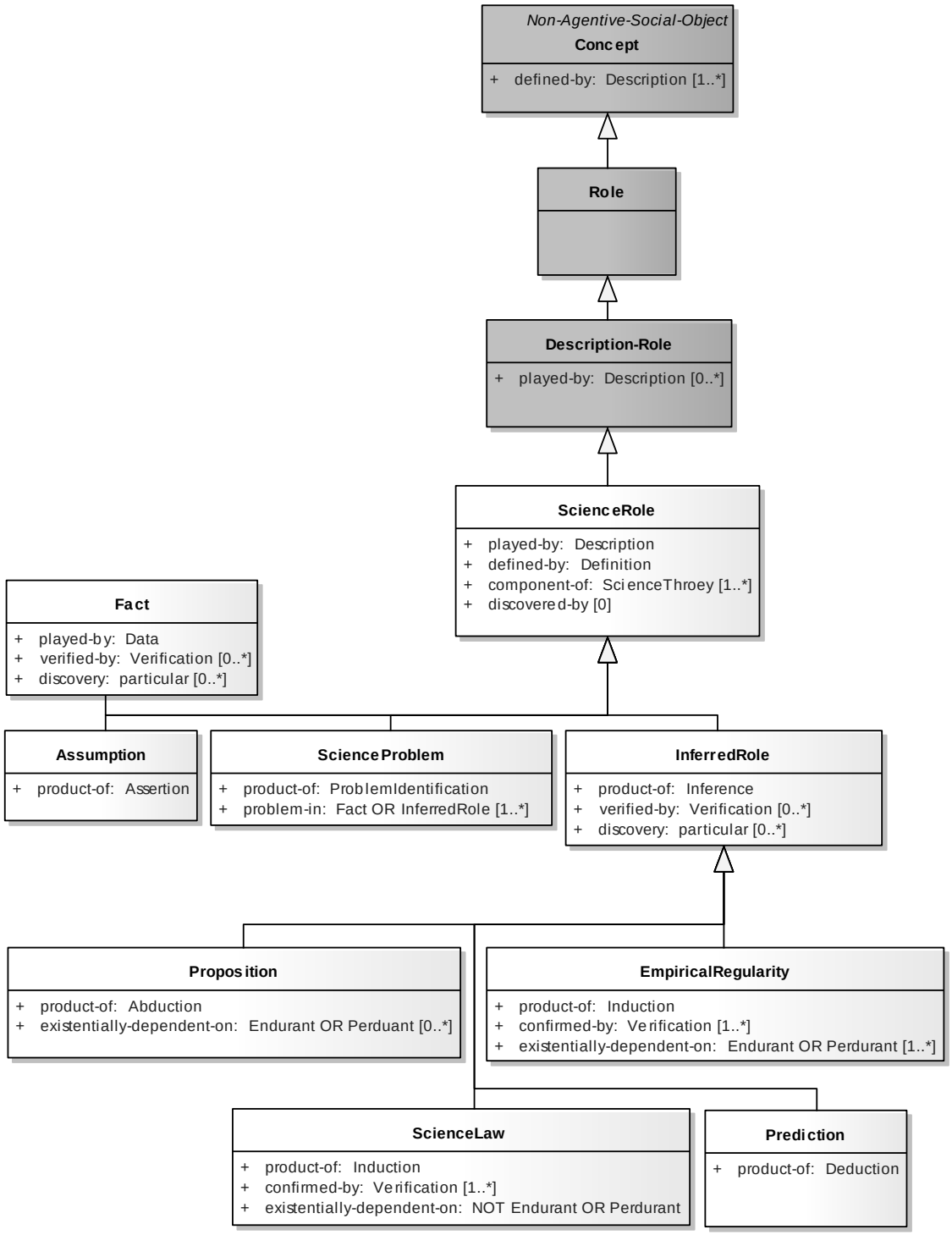


Figure 5: SKIo science roles (white) and DOLCE concepts (grey).

Activities: DOLCE activities are non-atomic perdurants that follow some plan, sequence some tasks, can produce some endurants, and are performed by some agents. In SKI the plan is likely some research project containing tasks performed by scientists. DOLCE activities are extended by SKIo through the addition of properties for the agents' motivation and for the entities upon which the activity is being performed. SKIo includes activities for observation, inference, assertion, verification, problem finding, science modeling, and doing research. These activities are important

because their products are key SKIo elements, such as science models, science roles, and science statements, as shown in Figure 6.

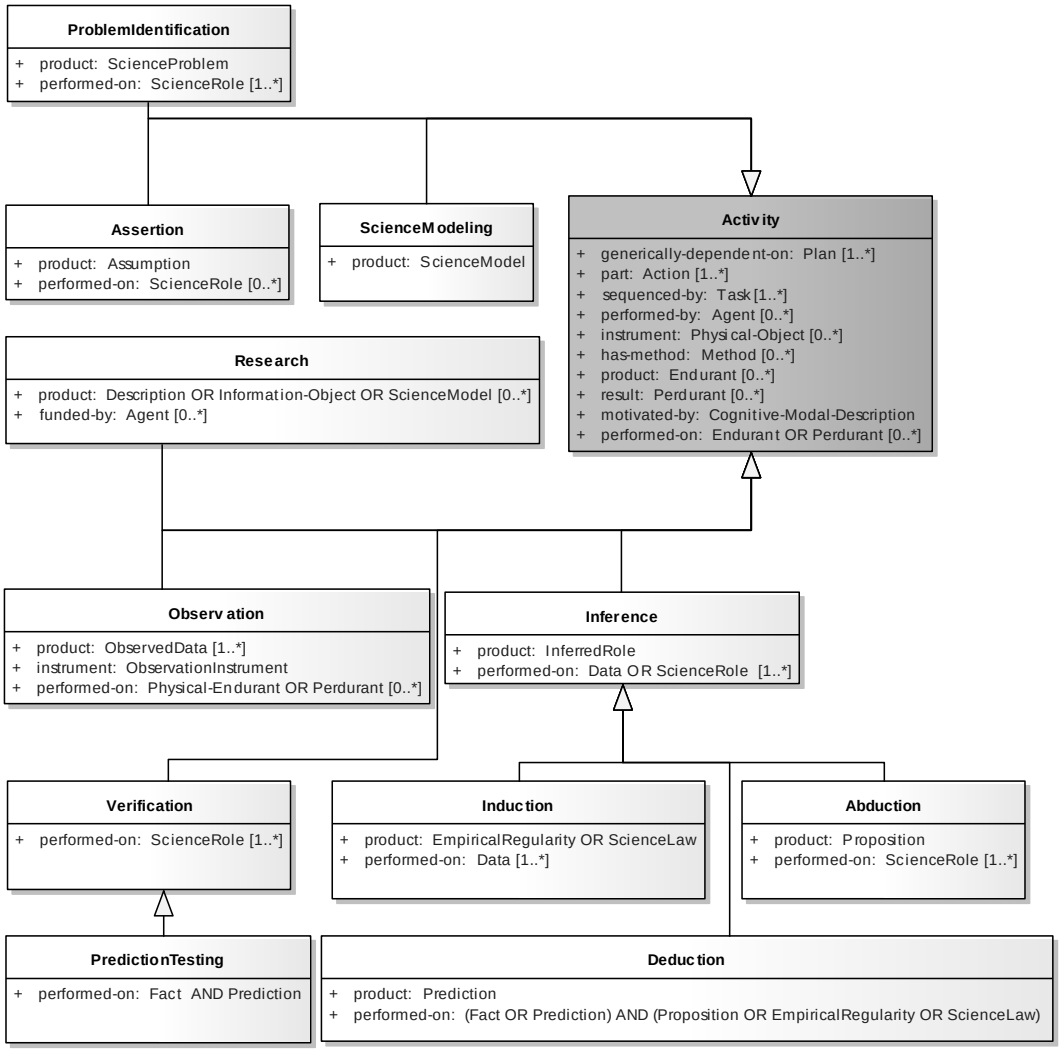


Figure 6: SKIo scientific activities (white), and the DOLCE activity class (grey).

Of particular importance are the *Inference* activities because they bind together much of the knowledge cycle (after Sowa 2000; 2006):

- ***Induction***: involves finding a pattern in data (logical induction), or dis/confirming a pattern via data collection (pragmatic induction). In logical induction: given data { (a_1, b_1), (a_2, b_2), (a_3, b_3) } then infer f (A, B), where f is some relation over concepts A,B, and a_i and b_i are their respective instances. In pragmatic induction, given f (A, B), note in data { (a_1, b_1), (a_2, b_2), (a_3, b_3) } and infer f (A,B) $\models$ TRUE or FALSE, possibly with the aid of intermediary predictions. SKIo *Induction* refers to logical induction, while *Verification* encompasses pragmatic induction and can operate on any science role.
- ***Deduction***: involves generating a *Prediction* about the world using existing theory and data. Logically, given theory T:A→B (A and B are concepts) and instance a_1 (of A), then b_1 (of B) is deduced: $T \wedge a1 \vdash b1$.
- ***Abduction***: involves generating a *Proposition* to enable coherence of discordant scientific

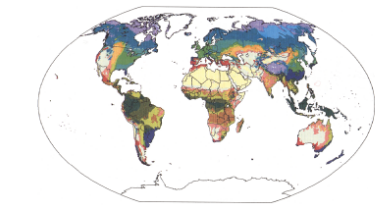
knowledge elements. In SKIo, *Abduction* encompasses both logical and pragmatic abduction. Logical abduction is reverse deduction: in a deduction of the form T: A→B, where T is a theory and A and B are concepts, given data b_1 (instance of B) and some or none of T and a_1 (an instance of A), then guess the missing T and/or a_1. Pragmatic abduction, on the other hand, is more concerned with the mechanism of guessing missing theories and data from prior knowledge (often via analogy ~): given C→D, and A~C ∧ B~D, then T:A→B.

6. Application of SKIo to Environmental Theories

Adams *et al.* (2004) summarizes and compares the mathematical formulation of ten "models" (*ScienceTheories* in SKIo) of terrestrial net primary production (NPP). The theories are expressed primarily as systems of equations containing (1) input variables and (2) fixed values divided into general constants (e.g. atmospheric pressure) and specific parameters (e.g. photosynthesis co-limitation). The theories can be categorized into biogeographic or biochemical, where the former are empirically inferred from data, and the latter are derived from existing theories and express biochemical processes more explicitly. Some of these theories share a few ancillary theories which were obtained from other sources such as books or web pages. SKIo representation of all these elements involves the following instances of SKIo classes:

- Each paper, book, and web site that expresses the theories is represented as an instance of *SciencePublication* containing *ScienceStatements*.
- Each equation, or other theory part such as a table or figure, is expressed as a distinct instance of *ScienceStatement*.
- The intensional description of each theory is represented as a whole *ScienceTheory* instance, and is expressed by the relevant *SciencePublication* instances.
- The intensional description of each equation, or other theory part, is represented as a *ScienceTheory* part instance, and is expressed by the relevant *ScienceStatement* instances.
- The variables and fixed values in each equation are represented as DOLCE *parameter* classes, an instance of each is *d-used* by each equation with a value in a region instance.
- Each part of each empirically inferred theory is represented as an *EmpiricalRegularity* instance, because these are largely induced from empirical data and are local to Earth situations (the fixed values are calibrated to the Earth environment).
- The parts of the theories that were derived from other theories play a *Proposition* science role, because these are theoretically postulated (hence via abduction).
- Some theories share common parts which, however, play different roles amongst the theories. For example, a theory developed by King *et al.* (1995) adopts a *Proposition* from Polglase and Wang (1992) as an *Assumption.*
- Many of the theories, such as the Miami Model (Leith, 1975), are satisfied by well known *ScienceModel* instances which are either based on empirical data (contain particulars generated from observed facts) or predictions (contain particulars created as a result of deduction using the theory and observed input data).
- The activities and data which led to the origin of each theory, or theory part, are not represented in this exercise, mainly because that information was not readily available amongst our sources.

Figure 7a and 7b show how the BIOME3 model, one of the 10 NPP models, can be represented in SKIo.



Predicated vegetation distribution with observed data

BIOME3 Equation

$$NPP(T, C_a, c_i, W, I_{PAR}, A, d) = \frac{365.25}{1000}(P(T, C_a, c_i, I_{PAR}, A, d) - R(T, C_a, c_i, d, A, I_{PAR})$$

Inputs:
P: Photosynthesis rate
R: Respiration rate
T: Temperature
Ca: Atmosphere CO2
ci: Internal leaf CO2 partial pressure (Pa)
W: Soil water or precipitation
IPAR: Photosynthetically active radiation reaching the canopy
A: Leaf area index
d: Day length

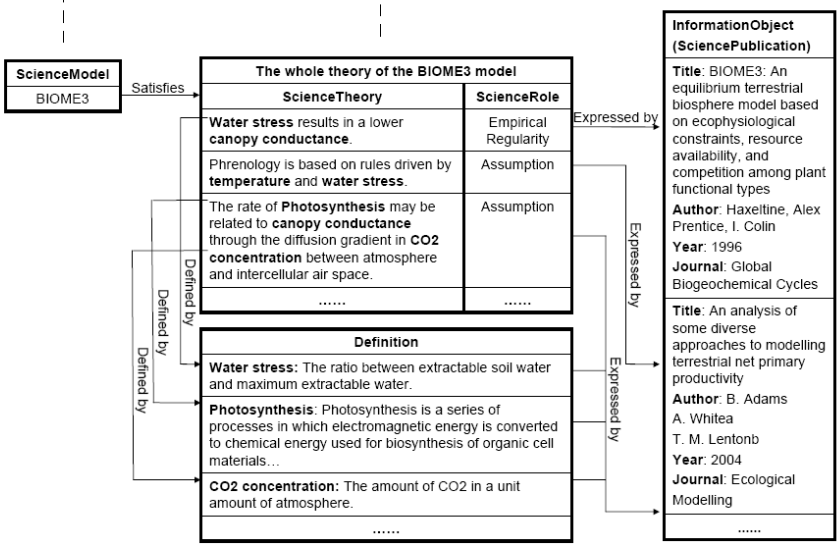


Figure 7a: an example representation of part of a environmental theory and model in SKIo.

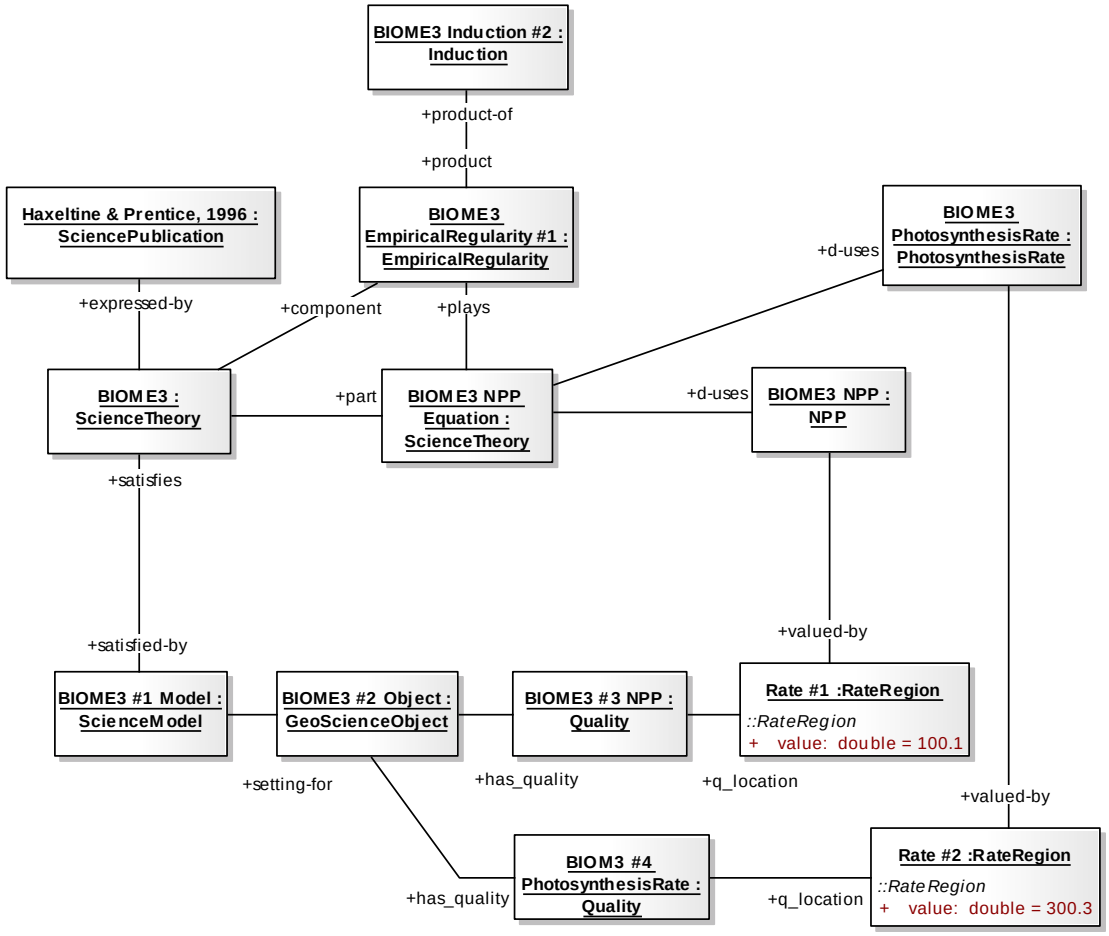


Figure 7b: an example representation of part of an environmental theory and model in SKIo; in the box labels the text before the colon denotes the instance name, and the text after the colon denotes the SKIo class name.

7. Discussion

Achievements: The SKIo ontology meets the representation requirements outlined in the introduction and in the environmental use-case scenario. It clearly and formally distinguishes between general scientific knowledge primitives, such as models and theories, and is shown to adequately represent environmental knowledge as documented in peer reviewed papers. As a result, when SKIo is coupled to a fully operational and semantically-enabled SKI, it should facilitate the search, retrieval and use of the basic science knowledge primitives. For example, knowledge about NPP could then be obtained by searching for the science theories that use the NPP concept. If the user is interested in the correlation between annual average temperature and NPP, then the Miami Model and the theory of King *et al.* (1997) could be retrieved, including a predicted model of world NPP.

Evaluation: SKIo is evaluated according to the principles outlined in Section 4. Formality and modularity are achieved through its extension of DOLCE and the independent OWL encoding, respectively. It is grounded in reasoning accounts of the science knowledge cycle, and includes science knowledge primitives of sufficient coverage and precision to enable at least environmental theories, and we suppose other science theories, to be adequately represented. Ontological

representation of the principles behind the knowledge cycle were shown to help disambiguate the polysemous use of natural language terms, e.g. "model".

Implementation: implementation of SKIo for full support of e-Science is a future activity. Immediate prospects include incorporation of SKIo into existing e-Science infrastructures that use ontologies to enable information discovery, retrieval and workflow operation. Future prospects include use of SKIo-enabled science resource to create and test science knowledge in SKI. This might include using SKIo to support web services for operating over science models as sketched in Figure 8.

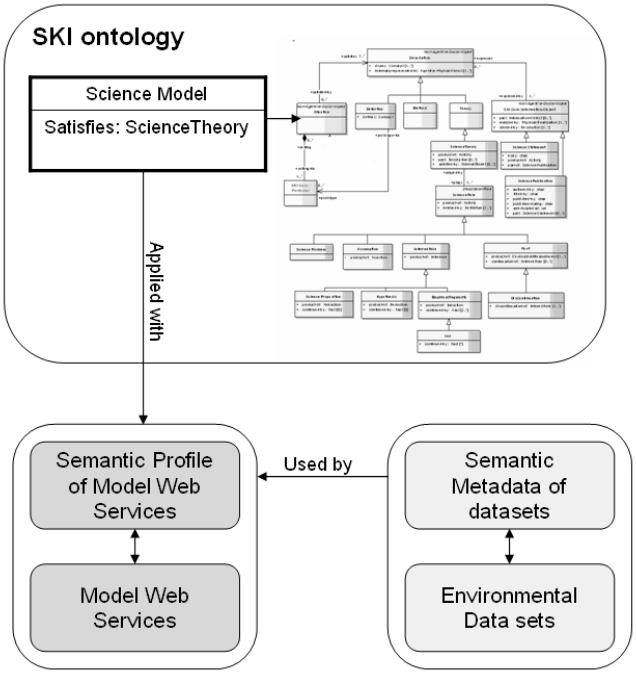


Figure 8: using SKIo in an environmental modeling SKI

Limitations: several limitations were encountered during design and application of SKIo:

- ***completeness***: SKIo is incomplete in several areas, most notably in representing common science methods, instruments, activities, information objects and science metadata.
- ***change***: representing scientific knowledge change, such as theory change, using existing constructs such as perdurants is largely uncharted territory that needs more exploration.
- ***grounding***: DOLCE does not provide adequate guidelines for distinguishing between the ground ontology (non-social classes) and social classes (contextualized entities). For example, should atmospheric pressure be a ground quality or a social parameter, or both? If so, why? Although SKIo provides a mechanism for representing the migration of a scientific idea from conjecture (social artifact?) to accepted dogma (ground artifact?)—i.e. the epistemology—the design choice for extending from a social or ground class needs more rigorous attention.
- ***instantiation***: mechanisms are required to limit instantiation of unwanted but mandatory DOLCE classes triggered by SKIo instantiations.

8. Conclusions

SKIo is an ontology of science that bridges the gap between a general foundational ontology and specific domain science ontologies. It accomplishes this by extending the DOLCE foundational ontology with general science primitives common to all science domains, such as theories, models and data, derived from computational accounts of the science knowledge cycle. Initial testing, involving representation of several environmental theories, confirms the effectiveness of SKIo's overall scope and design, and lends support for the notion that such representations will advance next-generation e-Science by facilitating not only search, retrieval, integration and workflow operations in e-Science infrastructures, but also the discovery and testing of scientific artefacts. Next steps include further testing with a variety of science knowledge from various domains, integrating with existing domain science ontologies, and incorporation into functioning e-Science infrastructures.

Acknowledgements

The authors gratefully acknowledge support of the U.K. e-Science Institute under theme 4 "Spatial Semantics for Automating Geographical Information Processes". Valuable comments that improved the paper were provided by Florian Probst and Mark Gahegan.

References

Adams, B., A. White, and T. M. Lenton. (2004). An analysis of some diverse approaches to modelling terrestrial net primary productivity. *Ecological Modelling* 177 (3-4):353-391.

Antoniou, G., Van Harmelen, F. (2003). Web ontology language: OWL. In: S. Staab, R. Studer (Eds.), Handbook on Ontologies, Springer: Berlin, pp. 67-92.

Baldocchi, D., Finnigan, J., Wilson K., Paw U K.T., Falge E. (2000). On Measuring Net Ecosystem Carbon Exchange Over Tall Vegetation on Complex Terrain. Boundary-Layer Meteorology, 96(1-2), August 2000, pp. 257-291(35).

Benjamins, V. R. Fensel, D., Gómez Pérez, A. (1998). Knowledge Management through Ontologies. In Proceedings, Knowledge Engineering: Principles and Methods, PAKM-98, pp. 5.1-5.12, 1998.

Bittner, T. (2007). From top-level to domain ontologies: Ecosystem classifications as a case study. In M. Duckham, B. Kuipers, L. Kulik, and S. Winter, editors, Spatial Information Theory. Cognitive and Computational Foundations of Geographic Information Science. International Conference, COSIT 2007, 61-77.

Brodaric, B. (2008). Science Knowledge Infrastructure (SKI) Ontology, version 3.0. e-Science Institute Technical Report , available from http://...

Gangemi, A., Catenacci, C., Battaglia, M. (2004). Inflammation Ontology Design Pattern: an Exercise in Building a Core Biomedical Ontology with Descriptions and Situations. In D.M. Pisanelli (ed.) Ontolgies in Medicine, IOS Press, Amsterdam, Berlin, 2004, pp. 64-80.

Gangemi, A., Guarino, N., Masolo C., Oltramari, A. (2003). Sweetening WordNet with DOLCE, AI Magazine, 24(3):13-24.

Gangemi, A., Mika, P. (2003). Understanding the Semantic Web through Descriptions and Situations. In: Meersman R, *et al*. (eds.), International Conference on Ontologies, Databases and Applications of Semantics (ODBASE 2003), Catania, (Italy), November 3-7, 2003.

Gardenfors, P. (2000). Conceptual spaces—the geometry of thought. Cambridge: MIT Press.

Grenon, G., Smith, B., Goldberg, L. (2004). Biodynamic Ontology:Applying BFO in the Biomedical Domain. In: Pisanelli, D.M. (ed.), Ontologies in Medicine, Amsterdam: IOS Press, 2004, 20–38.

Grenon, P., Smith, B. (2004). SNAP and SPAN: Towards Dynamic Spatial Ontology. Spatial Cognition and Computation, 4(1).

Gruber, T.R. (1995). Toward Principles for the Design of Ontologies Used for Knowledge Sharing. International Journal Human-Computer Studies. **43**(5-6): 907-928.

Hars, A. (2001). Designing Scientific Knowledge Infrastructures: The Contribution of Epistemology. Information Systems Frontiers, 3(1):63-73.

Hars, A. (2003). From Publishing to Knowledge Networks: Reinventing Online Knowledge Infrastructures. Spring, Berlin, 211 p.

Herre, H., Heller, B., Burek, P., Hoehndorf, R.Loebe, F., Michalek, H. (2006). General formal ontology (GFO) – part I: Basic principles. Technical Report 8, Onto-Med, University of Leipzig, 2006.

Hey, T., Trefethen, A.E. (2005). Cyberinfrastructure for e-Science. Science 308(5723):817-821.

Kaku, M. (2004). Einstein's Cosmos: How Albert Einstein's Vision Transformed Our Understanding of Space and Time. Norton, New York.

King, A. W., W. R. William, and W. M. Post. (1995). In search of missing carbon sink: a model of terrestrial biospheric response to land-use charge and atmosphere CO2. *Tellus. Series B* 47:501-519.

Langley, P. (2000). The computational support of scientific discovery. Int. Journal of Human-Computer Studies, 53: 393-410.

Leith, H. (1975) Modeling the primary productivity of the world. In *Primary Productivity of the Biosphere*, ed. H. Leith, Whittaker, R.H. New York: Springer-Verlag.

Ludäscher, B., K. Lin, S. Bowers, E. Jaeger-Frank, B. Brodaric, C. Baru, C. (2006). Managing scientific data: From data integration to scientific workflows. In: Sinha, A.K. (Ed.) Geoinformatics: Data to Knowledge, GSA Special Paper 397, 109-130.

Masolo, C., Borgo, S., Gangemi, A., Guarino, N., Oltramari, A. (2003). WonderWeb DeliverableD18, Ontology Library (final). Laboratory For Applied Ontology, http://www.loa-cnr.it/Papers/D18.pdf.

McCray, A. (2003). An upper-level ontology for the biomedical domain. Comparative and Functional Genomics, 4:80–84.

McGuinness, D. L., Fox, P., Cinquini, L., West, P., Garcia, J., Benedict, J. L., Middleton, D. (2006). Enabling Scientific Research using an Interdisciplinary Virtual Observatory: The Virtual Solar-Terrestrial Observatory Example. AI Magazine,

Natale, D.A., Arighi, C., Barker, W.C., Blake, J., Chang, T., Hu, Z., Liu, H., Smith, B., Wu, C.H., Framework for a Protein Ontology. BMC Bioinformatics 2007, 8(Suppl 9):S1.

Pease, A., (2006). Formal representation of concepts: The Suggested Upper Merged Ontology and its use in linguistics. In: Schalley, A. C., Zaefferer, D. (Eds.), Ontolinguistics. How Ontological Status Shapes the Linguistic Coding of Concepts. Vorbereitung Berlin:New York (Mouton de Gruyter).

Phillips, T. J., AchutaRao, K., Bader, D., Covey, C., Doutriaux, C. M., Fiorino, M., Gleckler,

P. J., Sperber, K. R., Taylor, K. E. (2006). Coupled climate model appraisal, a benchmark for future studies. *EOS Trans. AGU* 87(19): 185.

Polglase, P. J., and Y. P. Wang. (1992). Potential CO2-enhanced carbon storage by the terrestrial biosphere. *Australian Journal of Botany* 40 (641-656).

Raskin, R.G., Pan, M.J. (2005). Knowledge representation in the Semantic Web for Earth and Environmental Terminology (SWEET). Computers and Geosciences, 31(9):1119-1125.

Rijgersberg, H., Top, J., Meinders, M. (2008). Use of a Quantitative Research Ontoloy in e-Science. In: McGuinness, D.L., Fox, P., Brodaric, B. (Eds.), Semantic Scientific Knowledge Integration, AAAI Spring Symposium, Technical Report SS-08-05, AAAI Press, Menlo Park, CA, 87-92.

Shrager, J., Langley, P. (1990). Computational Models of Scientific Discovery and Theory Formation. Morgan Kaufman, San Mateo.

Soldatova, L.N., King, R.D. (2006). An ontology of scientific experiments. Journal of the Royal Society Interface, 3(100):795-803.

Sowa, J. F. (2000). Knowledge Representation: Logical, Philosophical, and Computational Foundations. Brooks/Cole: New York.

Sowa, J. F. (2006). The Challenge of Knowledge Soup. In: J. Ramadas, S. Chunawala, (Eds.), Research Trends in Science, Technology and Mathematics Education, Homi Bhabha Centre, Mumbai, 2006, pp. 55-90.

Thagard, P. (1988). Computational philosophy of science. MIT Press, Cambridge, MA.