In press in Ecological Informatics. Please contact for permission to cite.

An ontology for describing and synthesizing ecological observation data

Joshua Madin[1]*, Shawn Bowers[2], Mark Schildhauer[1], Sergeui Krivov[3], Deana Pennington[4], Ferdinando Villa[3]

[1]National Center for Ecological Analysis and Synthesis, University of California, Santa Barbara, California 93101
[2]Genome Center, University of California, Davis, California 95616
[3]Gund Institute for Ecological Economics, University of Vermont, Burlington, Vermont 05405
[4]University of New Mexico, Albuquerque, New Mexico 87131
*Corresponding author: madin@nceas.ucsb.edu

Abstract

Research in ecology increasingly relies on the integration of traditionally small, focused studies, to produce larger datasets that allow for more powerful, synthetic analyses. The results of these synthetic analyses are critical in guiding decisions about how to sustainably manage our natural environment, so it is important for researchers to effectively discover relevant data, and appropriately integrate these within their analysis. This is a major challenge however, as ecological data encompass an extremely broad range of data types, structures, and semantic concepts. Moreover, ecological data is widely distributed, with few well-established repositories or standard protocols for their archiving and retrieval. These factors presently make the discovery and integration of ecological data sets a highly labor-intensive task. Metadata standards such as EML and Darwin Core are important steps for improving our ability to discover and access ecological data, but are limited to describing only a few, relatively specific aspects of data content (e.g., data owner and contact information, variable "names", keyword descriptions, etc.). A more flexible and powerful way to capture the semantic subtleties of complex ecological data, its structure and contents, and the inter-relationships among data variables is needed.

We present a formal ontology for capturing the semantics of generic scientific observation and measurement. The ontology provides a convenient basis for adding detailed semantic annotations to scientific data, which crystallize the inherent "meaning" of observational data. The ontology can be used to characterize the context of an observation (e.g., space and time), and clarify inter-observational relationships such as dependency hierarchies (e.g., nested experimental observations) and meaningful dimensions within the data (e.g., axes for cross-classified categorical summarization). It also enables the robust description of measurement units (e.g., grams of carbon per liter of seawater), and can facilitate automatic unit conversions (e.g., pounds to kilograms). The ontology can be easily extended with specialized domain vocabularies, making it both broadly applicable and highly customizable. In particular, explicit "extension points" allow new types of observable entities (e.g., tree, rock, population), characteristics (e.g., height, color, diversity), and unit definitions to be easily added. Finally, we describe the utility of the ontology for enriching the capabilities of data discovery and integration processes.

1 Introduction

Ecology is an inherently multidisciplinary science that explores how physical and biological factors and their inter-relationships establish the structure and function of living systems. Accordingly, the range of data that can inform ecological analyses is incredibly broad, often involving perspectives from many fields in the earth sciences (e.g., geography, oceanography and hydrology) and life sciences (e.g., genetics and physiology). The need to access these diverse data sources becomes especially acute when undertaking *synthetic* analyses to address broad ecological questions, such as the impacts of deforestation on the global balance of greenhouse gases, or the link between biodiversity losses and the productivity of our world's fisheries. Ecological insights are critical to understanding many complex real world issues that have vital implications for the quality and sustainability of life on this planet.

Approaches to ecological synthesis today leverage the rapidly growing amounts of data available through the Internet. However, current methods for finding and interpreting potentially relevant data are extremely primitive and inefficient, which severely impedes progress in accomplishing synthetic ecological science. The lack of advanced technical tools for data exploration and interpretation has been long recognized, and is fundamentally due to semantic limitations of current relational database systems (Chen, 1976; Batini et al., 1992; Hammer and McLeod, 1999). Proposed enhancements are still largely lacking in practical, non-proprietary implementations.

Effective data discovery is particularly problematic in ecology, where traditionally small, focused studies employed largely *ad hoc* data management solutions, often consisting of flat files or spreadsheets with minimal formal structure and little to no metadata documentation. This situation was viable when researchers worked only with their own data, and data management was considered merely a provisional framework for accomplishing some specific analyses, after which, one moved on to other research questions and data analyses. Researchers maintained many of the details of their data in their memory, with maximum cognizance of the relevant subtleties and issues in the data ideally occurring simultaneously with the period when they were actively being analyzed (Michener et al., 1997).

Recently there has been a growing recognition of the need to both preserve ecological data after their "intended" usage was completed, as well as to extend data collection events through time to discern long-term trends in ecological processes through intensive site-based studies. This recognition raised concerns about the lack of protocols and services for preserving ecological data (Gross and Pake, 1995), and clarified the need for reducing the possibilities of "data entropy" (Michener et al., 1997). Efforts grew to develop metadata standards that can systematically structure the types of information that researchers should document about their data, making these far more effective for informing future studies (Jones et al., 2006).

The Ecological Metadata Language (EML) was developed through a community-based effort involving researchers and information managers from several institutions charged with accomplishing ecological synthesis and long-term research (Jones et al., 2001). While intended primarily for the purpose of preserving critical metadata about ecological data sets, it is essentially a generic standard for describing tabular data, in addition to a number of other data formats (Fegraus et al., 2005). While EML is a growing standard for data documentation in the ecological field, practical experience using this standard has revealed that metadata alone has some serious shortcomings in terms of the capabilities it can provide scientists in data discovery and interpretation. These shortcomings are particularly severe in ecology due to the incredible heterogeneity of topics studied, and the relative lack of standardized protocols and methods when measuring variables of ecological interest.

Metadata languages have also been developed for describing natural history specimen data, such as Darwin Core (Darwin Core, 2004). Both the Darwin Core and EML metadata standards primarily focus on describing data format (i.e., describing data structure) along with high-level contextual information (often by adopting Dublin-core style attributes such as who created a data set and when [DCMI, 2006]). Furthermore, these standards generally lack support for capturing even the basic "semantics" of data—i.e., information that broadens the capability for understanding or interpreting the content and relevance of the data from a disciplinary perspective, and even in terms of fundamental physical quantities. For example, while EML allows one to declare that a data set contains an attribute labeled "biom" (e.g., referring to a biomass measure), it is not possible using EML to determine if it is compatible with another attribute labeled "weight" or "kg". What is needed is a way to capture such concepts and relationships in formal models that can then be used to draw logical conclusions (e.g., consistency, equivalence) without human intervention. The creation of such a framework must also address the need for a simple mechanism to assist scientists in mapping their observations onto such models.

This paper describes an approach for enhancing the capability of ecological scientists to more powerfully discover, interpret, and reuse data in support of synthetic research. While provision of access to "others" data raises some interesting issues and challenges with regards to the sociology of data-sharing and intellectual property rights, the focus here is solely on addressing several of the most pressing technological impediments to accomplishing scientific synthesis: *data discovery* and (legitimate) *integration*. Discovery is the process of locating relevant and available data related to a specified topic of interest. This process is currently hampered by the lack of well-described data to begin with, and compounded by the inability to clearly explicate and explore basic semantic notions within and across data sets (e.g., that biomass is a weight and that "dry weight" is a biomass and a weight). Integration is the process of merging compatible data once these are discovered. Here, we present a formal *ontological* framework for capturing the essential semantic information of observational data sets to better facilitate the discovery and integration of ecological information, thus aiding ecologists in synthesizing knowledge for answering larger ecological questions.

Ontologies are representations of the knowledge within a domain of interest, defined via the terminology (concepts) used within the domain and the properties and relationships among domain objects (Baader et al., 2003). Ontologies capture the semantic definitions within domains, and represent one enabling mechanism for providing more comprehensive data discovery and integration (Jones et al., 2006). As a simple example, if instances of *biomass* are defined as instances of *weight* in a particular domain ontology, then data about *biomass* will be discovered when searching for data about *weight* (i.e., based on an ontology reasoning system [Baader et al., 2003]). Moreover, these data are compatible through being semantically classified as *weights*, and can potentially be merged. A number of formal (logic-based) languages exist to capture ontologies, including description logics (Baader et al., 2003), semantic networks (Sowa, 1999), and the more recent RDF and OWL web-based standards (McGuinness and van Harmel, 2004). While ontologies are being used successfully in a number of biological and medical informatics projects (The Gene Ontology Consortium, 2000; Rosse and Mejino Jr., 2003; Bard and Rhee, 2004), widespread support for ontology-based approaches and ontology development has yet to be adopted within the field of ecology.

Our work on ontologies is within the context of the Science Environment for Ecological Knowledge (SEEK; http://seek.ecoinformatics.org) project, which aims to develop technology to discover, access, integrate, and analyze distributed ecological information (e.g., using scientific workflows [Ludäscher, 2006; Kepler Project, 2006]). The project's approach is to extend EML to support the semantic annotation of ecological data sets, such that EML data-set descriptions can use terms drawn from OWL-DL ontologies. A benefit of using OWL-DL (which itself is based on description logic) is that it supports a "natural" representation for formalizing terms. In particular, named classes (such as *Biomass* or *Plot*) define sets in OWL-DL, where each member of a set is considered an "instance" of the corresponding class (e.g., the class *Biomass* might denote the set of all biomasses; *Plot* the set of all physical plots). Class definitions are typically intentional, i.e., classes are defined based on their name or possibly other constraints, without enumerating their instances. The "intent" of a class is usually further elaborated by relating it to other classes. The "is-a" relationship defines class specializations as subsets. For example, the expression "*Biomass* is-a *Weight*" implies all biomass instances are also valid weight instances. OWL-DL also supports user-defined properties, e.g., to capture part-of relationships between classes, as well as cardinality and other (set-based) constraints (Baader et al. 2003).

While formal languages such as OWL-DL provide a means to capture ontologies, the quality of the realized ontology will determine its utility for assisting in data discovery and integration. Additionally, as the number of ontologies and their included terms increases, organizing these into a coherent framework becomes increasingly complex, as recognized within the biological community, e.g., see [Bard and Rhee, 2004]. In this paper, we describe the SEEK Extensible Observation Ontology (OBOE), which aims at providing a core ontology framework for semantically annotating observational data sets. Our framework defines a formal ontology based on the concepts of *Observation*, *Measurement*, *Unit*, *Characteristic*, and (Ecological) *Entity* (Figure 1), providing a structured yet generic approach for semantic data annotation and for developing (and combining) domain-specific ecological ontologies. Our approach differs from other ontology-based descriptions of ecological information (e.g., Keet, 2005; Smith and Varzi, 1999a; Smith and Varzi, 1999b; Smith, 2001; Gruber and Olsen, 1994; Brilhante, 2003) in that we focus specifically on providing: (i) a robust framework for describing generic scientific observations; (ii) a structured approach for easily building and sharing domain-specific ontology extensions; and (iii) data discovery and integration services, via semantic annotations to the ontology, across varied ecological observation data (and not just for a specific, specialized domain). In Section 2, we describe our framework using a number of real-world examples, and illustrate how it can be extended with domain-specific ontologies. The ontology framework presented here has evolved through various earlier efforts within SEEK to develop formal ecological ontologies (Berkley et al., 2005; Bowers et al., 2005; Bowers and Ludäscher, 2006; Villa, 2007), and is based on a number of working meetings with members of the SEEK project as well as participants from the broader ecological community. In Section 3, we outline applications of our ontology for data discovery and integration. Finally, Section 4 concludes with a summary of our contributions and future work.

2 The Observation Ontology

The goal of the *Extensible Observation Ontology* (OBOE) is to serve as a formal and generic conceptual framework for describing the semantics of observational data sets (i.e., data sets consisting of observations and measurements). OBOE also prescribes a *structured* approach for organizing domain-specific ontologies through the use of "extension points." OBOE extension points allow ontology classes, properties, and constraints to be easily defined for a particular domain-specific terminology, and existing domain extensions to be interrelated. Thus, OBOE can serve as a *framework* for defining new domain ontologies as well as interoperating and relating existing ones. Figure 1 graphically depicts the basic core structure of OBOE, which consists of five classes labeled *Observation*, *Entity*, *Measurement*, *Characteristic*, and *Measurement Standard*, and six properties labeled *hasContext*, *ofEntity*, *hasMeasurement*, *hasValue*, *hasPrecision*, *usesStandard*, and *ofCharacteristic*. (Note that we capitalize OBOE classes to distinguish them from more general concepts, e.g., 'Observation' denotes an OBOE class whereas 'observation' denotes the more general concept.)

Most details of observational data are not recorded. Instead, the physical representation of data is often optimized for data collection; for use within a specific tool, e.g., R (R Development Core Team 2005) or SAS® Software; or for a particular analysis, e.g., to perform a calculation requiring a site-by-species matrix. As a consequence, contextual information concerning data is typically *implicit*, where context is (possibly) encoded by attribute labels, implied by the proximity of attributes (i.e., neighboring data), stored in metadata as natural-language descriptions, or altogether missing. Consider the first data table in Figure 2. The column of data labeled "Ht" can be assumed to represent height, giving information about the *characteristic* of some entity that was measured. However, no explicit information is given about the *entity* itself. The neighboring column labeled "Sp" suggests that each height measurement was for a species given in the column. Further, if all such species values in the column correspond to types of *reef coral*, one could surmise the kind of *entities* to which the height measurements pertain. The goal of OBOE is to provide a generic model for making such information explicit, which can then enable automated approaches for merging data (e.g., in this case with other coral data) and data discovery (e.g., for researchers looking for data on coral heights).

The rest of this section details the core OBOE classes and properties shown in Figure 1, and demonstrates the use of OBOE for capturing observation data and for extending OBOE with domain ontologies. Although not discussed here, OBOE is encoded using the OWL-DL ontology language (McGuinness and van Harmelen, 2004), which gives a description-logic (Baader et al., 2003) formalization of the various parts of the OBOE ontology described below.

2.1 Observations

In OBOE, an observation is a statement that an entity of a particular type was observed. As shown in Figure 1, all *Observations* are composed of exactly one *Entity* (expressed by the cardinality restriction "1:1" on the *ofEntity* property). The Entity class in OBOE represents all concrete and conceptual objects that are "observable." While this notion of entity is extremely generic, it serves as a placeholder (i.e., extension point) for more specific types of objects. The bottom portion of Figure 3 gives a simple example of an Entity-class extension model, which is used in the examples of this paper. As shown, Entity classes are extended via *is-a* relations. For example, every *Ecological Entity* is also an Entity, every *Population Entity* is an Ecological Entity (and hence an Entity), and so on. Hierarchies, like the Entity class hierarchy of Figure 3, can be additionally constrained in OBOE using OWL-DL language constructs, specifying that classes are disjoint, equivalent, or related to multiple other classes combined through set union and intersection operations (e.g., stating that one class is equivalent to the union of two or more other classes). Although not shown, a number of Entity classes in Figure 3 are defined as non-overlapping (disjoint). For instance, the *Male* and *Female Entity* classes are represented as disjoint sets of objects implying that no Male Entities are Female Entities, and vice versa. While the domain extensions portrayed in Figure 3 are narrow and specialized for the purposes of this paper, we do not anticipate a single, universally accepted OBOE Entity classification. Instead, we aim to support multiple domain-specific extensions through OBOE, in which scientific groups and communities can flexibly build, share, and extend their own specialized entity (and other extension) models.

The Entity classes associated with observations in the first example data set of Figure 2 include *Reefs*, *Replicate Transects*, *Animals*, and *Populations* of coral crab. In addition, some of these observations serve as context for other observations. For instance, each observation of a coral crab population occurs within the context of an animal (i.e., a coral colony). The *hasContext* property shown in Figure 1 is used to capture these kinds of contextual relationships. In particular, an observation can serve as context for zero or more other observations, and can itself have multiple contexts (e.g., a replicate transect may occur within a broader spatial context as well as a particular temporal context). Context in OBOE is defined independently from the observed entity, allowing the notion of observation scale to be efficiently formalized as well as systems implementing OBOE to perform automatic re-contextualization when merging observations (e.g., see Villa, 2007). Examples using context for merging observations are shown in Section 3.3.

The *hasContext* property asserts a "dependency" relationship between corresponding entities at the time of the observation, and thus, is defined to be transitive. For instance, each observation serving as context for replicate transect observations in the first example data set of Figure 1 is also context for corresponding coral-colony observations. The *hasContext* property can also be extended to represent more specific types of contextualization, e.g., many of the mereological (i.e., part-of or containment) relations defined in Wand et al. (1999) would be suitable extensions. The transitive nature of *hasContext* can simplify the semantic annotation process in that by specifying only direct observation dependencies, it is possible to automatically infer all other indirect dependencies.

In OBOE, the type given to an observed entity is considered an *essential* quality (Guarino and Welty, 2002). An entity's essential qualities help to define it, and always hold (are invariant) regardless of the entity's context. For example, observing an entity of type *Animal Entity* implies that regardless of context, the particular object being observed is an "animal". In this case, if the object was not an animal, it would be a different object altogether. Alternatively, one may assert that for a given context, a particular object is observed to be of a certain type, possibly to some degree of confidence. In this case, the type may not be an essential quality of the entity, since in a different context the quality may not hold. For instance, an object's observed height may change in different contexts. Assuming height is not an essential quality of the object, it would not be correct to assign it a type such as *Tall Animal Entity*. Qualities that are not essential are attributed to entities in OBOE through measurements, which we describe in the following subsection.

2.2 Measurements

In OBOE, Observations can be composed of *Measurements*, which represent measurable *Characteristics* (i.e., qualities) of the entity being observed. As mentioned above, measurements in OBOE are assertions about an entity, and are not necessarily essential to the entity. Although not shown in Figure 1, a Measurement is always associated to an Observation (i.e., no Measurement can exist without an associated Observation). Moreover, a particular Measurement can be associated with *at most* one Observation (i.e., two Observations may not share the same Measurement). Measurements assign values, via a *Measurement Standard*, to the characteristic of the associated entity. For certain types of measurements (e.g., physical quantities), a *Precision* is also given. Measurements in OBOE can be used to denote a wide range of instantiations of entity characteristics, including name; a categorical assignment such as color, i.e., nominal and ordinal measurements (Stevens, 1946); or the existence of some entity, e.g., measured as "presence". Thus, the concept of measurement in OBOE is more generic than physical measurements alone. Figure 4 gives a number of example observation-measurement instantiations. Note that we use the UML convention "id : Observation", or simply ": Observation" when the id is unknown, to denote an instance of the Observation class (Jacobson et al., 1998). Figure 4a states that a coral was observed such that the height of the coral was measured as 0.46 meters with a precision of 0.01. Figure 4b states that the diversity (a characteristic) of a community was measured as 1.24 according to the Shannon-diversity index. Similar to observations, OBOE does not elaborate the methods used to measure characteristics (e.g., whether a characteristic was measured via measuring tape, a reference to a species description, etc.), but could be extended to do so.

In OBOE, *Characteristics* represent the types of measurable traits of entities, and denote another OBOE extension point. Example Characteristic class extensions are shown at the top of Figure 3, again defined via the *is-a* relation. *Dimension* is a core OBOE class that represents the distinguished set of physical dimensions that are used to define unit system (described further below). Concrete Characteristic classes in our example extension ontology denote "stand-alone" trait types, e.g., which can be used directly in semantic annotations. Abstract Characteristic classes, however, must be combined—through intersection—with concrete characteristics to be used in annotation. For instance, the class *Maximum Depth* used in Figure 4c is defined as the intersection of the *Maximum* and *Depth* Characteristic classes.

Characteristics and Entities are assumed to be disjoint (similar to Parsons and Wand [2000] and Bunge [1977]), even though the same term can often be used to refer to a characteristic or entity in natural language. For instance, 'area' could be the subject (i.e., entity) of an observation, and characteristics of that area entity measured (e.g., width). Conversely, 'area' could be a characteristic measured about an entity (e.g., the physical size of a study plot). Unique names are required in OBOE to distinguish these terms, e.g., *Area Entity* and *Area Characteristic*.

2.3 Measurement Standards

A measured value (or "data point" in an observational data set) cannot be interpreted without reference to a defined measurement standard. Moreover, data integration relies on the ability to determine if two values are compatible, and if conversion to a common standard is possible. In OBOE, *Measurement Standards* are all the units, scales, categories, catalogs, and lists that are utilized when measuring a characteristic. Figure 3 illustrates core OBOE classes for representing measurement standards, as well as a selection of classes used by examples in this paper (i.e., as a Measurement-Standard class extension). OBOE enforces a constraint (*forCharacteristic*) between the measurement standards and the characteristics that they represent. For example, the physical unit meter can only be used to represent characteristics that belong to dimension length, such as height (see Figure 3).

OBOE defines four subclasses of Measurement Standard: *Unit*, *Index*, *DateTime*, and *Classification Domain*. The first of these, *Unit*, is subdivided into three disjoint classes. *Base Unit* contains the fundamental physical units, including SI units (e.g., meter, kilogram, second), and all manifestations of these units (e.g., millimeter, gram, hour). *Base Unit* also contains units for angle (e.g., degree), as well as number of items (individual). *Simple Derived Unit* contains all the physical units that are raised to a power other than 1, e.g., *Meter Square* (see Figure 5a). The final unit class, *Complex Derived Unit*, is composed of two or more *Simple Derived Units* and/or *Base Units*, an example of which is given in Figure 5b. Here, the *Individual per Meter Square* class is composed of the *Individual* Base-Unit class and the *Per Meter Square* Simple Derived-Unit class (i.e., *Meter Square* raised to the power -1). The OBOE representation for units as well as the corresponding representation for domains (given in Figure 3) is adopted from the approach used by the EML unit dictionary (Jones et al., 2001; Michener 1997), which is based on the STMML language (Murray-Rust and Rzepa, 2002) and the NIST Reference on Constants, Units, and Uncertainty (http://physics.nist.gov/cuu/Units/introduction.html).

Measurement standards can be related to the type of entity being measured via the *forEntity* property. In the case of complex derived units (e.g., *areal density*, Figure 5b)—composed from two or more independent unit types—each independent unit type can additionally be related to a corresponding entity type. In this way, OBOE provides a mechanism to describe so-called *semantic units*, e.g., grams of carbon per cubic meter of seawater, or individuals of rabbit per individuals of fox. For instance, in Table II of Figure 2, the data value 6 represents the density of coral crabs in a coral, which can be expressed using a semantic unit (corresponding to the *Individual per Meter Square* class) defined as the composition of two unit components: the unit component *Individual* linked to the *Crab Entity* class, and the component *Meter Square* (raised to the power -1) linked to the *Coral Entity* class (shown as dashed ellipses in Figure 5b). Semantic units are commonly used in ecological data sets, and using OBOE it is possible to formally define the meaning of these units, making them available for discovery and integration processes (see Section 3).

The second subclass of Measurement Standard, *Index*, is a container for all the indices, scales, and surrogates of non-dimensional measures such as pH, the Richter Scale, and the various representations of biological diversity and evenness. *Indices* are often calculated using physical units, but have lost physical dimensionality due to a functional transformation (e.g., logarithmic or exponential transformation). For example, by log-transforming a measurement of height, dimensionality is lost, and the measurement instead becomes an index for height. Note that we do not consider a "dimensionless" unit of measurement in OBOE, since units are essential information in data integration even when the components of a complex unit cancel. For example, a ratio of two heights measured in meters is represented using a complex derived unit consisting of the base unit *meter* and the simple derived unit that represents *meter* raised to the power of -1.

The third subclass, *DateTime*, is a container for the different ways of expressing a point in time (e.g., 12:30 am, 12-23-06, or 234 mya [million years ago]), as opposed to time intervals, which are represented using physical units (e.g., second, year, etc.). Many standards for representing the notion of date-time exist, and can be used to extend OBOE's *DateTime* class. Extensions for the *DateTime* class can include standards for representing Gregorian and Julian calendars, geological time scales, and the correlations among them. Measurements that use the first three subclasses of Measurement Standard—Unit, Index, and DateTime—also have *precision*, which indicates an estimate of the proximity of the measurement value to the real world value. Precision differs from accuracy, which is a methodological consideration, and thus, like methods for observation and measurement, is not elaborated here.

The final subclass, *Classification Domain*, contains measurement standards representing holistic characteristics of entities, such as name, taxonomic classification, color, or rank. A classification domain contains *Classification Entities*, which are entities used as standards for comparison when making a measurement (be it a conscious or subconscious comparison). For example, the *Sex Classification-Domain* class for the *Sex* Characteristic class has two elements: *Female Entity* and *Male Entity* (Figure 5d). Figure 4c illustrates the use of this classification domain, stating that an animal was observed, the sex characteristic was measured, resulting in the value "F", which is the representation used in the data set for a *Female Entity* (from the classification domain). Therefore, this example asserts that the entity, in the particular observation context, is the intersection of animal (given as an essential quality) and female (asserted via measurement).

OBOE requires a measurement standard for all measurements. However, in certain situations a measurement standard does not exist, e.g., when an *ad hoc* naming scheme is used in a data set. The default OBOE measurement standard is Classification Domain, which assumes a corresponding (exemplar) Classification Entity whose name is the value of the measurement. For instance, Figure 4d states that a particular reef was observed and that its name (in the current context) is "bird", however, a reference to a known naming standard is not given. In this case, a default standard is used containing a single Classification Entity also named "bird". Here, the reef entity used in the default classification domain is not the same entity that is the subject of observation, and instead, is considered an exemplar entity for the particular name (Stevens, 1946). These entities are different due to the name assertion, i.e., because we are only asserting that the observed reef is named "bird", and that this name may not be an essential quality of the reef.

This approach for naming the observed reef is necessary when the extension ontology does not contain the pertinent entity for classification. However, if the *Bird Reef Entity* class does exist in the extension ontology, then there are two additional possibilities. First, if being a "Bird Reef" is an essential quality of the entity, we can directly state the observed entity type as Bird Reef Entity (Figure 4e). Alternatively, if this is not an essential quality, we can assert via a measurement that it is "most like" the Bird Reef Entity in the Lizard Island Reef domain (Figure 4f). Finally, Figure 4g-i also illustrates various representations of taxonomic name using OBOE. Figure 4g considers *Acopora hyacinthus* the subject of observation (and therefore an essential quality), which is not correct when describing organisms due to the fact that taxonomy varies with opinion, but might be correct if describing data about a taxonomic concept. Figure 4h considers the case where a relevant taxonomic domain is absent. Figure 4i demonstrates the use of prescribing additional attributes to a classification entity, in which data pertaining to *Acropora hyacinthus* is described as being "most like" the taxonomic concept denoted by the *Acropora hyacinthus Entity* according to "Dana 1846" in the *Coral Taxonomic Domain* (Figure 5d); where *taxonomic concept* denotes the representation standard defined by Kennedy et al. (2005).

3 Applications of the Observation Ontology

This section gives an overview of ways that OBOE can be used to facilitate the discovery and integration of ecological data. To expose the semantics of data using OBOE, data must first be annotated with relevant terms and relationships from the ontology. Semantic annotation is the process by which data are mapped to the ontology, and is accomplished by asserting the membership of data in ontology classes (e.g., Animal Entity, Height, Meter) and any additional relationships among classes (e.g., hasContext). An annotation language has been developed (e.g., see [Berkley et al., 2005; Bowers and Ludäscher, 2006]) that formalizes this mapping, and allows the annotation to be applied flexibly to multiple rows or cells depending on the data structure (e.g., table, cross-tabulation, etc.). Annotations expressed in this language are serialized according to an XML Schema, e.g., allowing them to be embedded within existing EML documents (or alternatively, as stand-alone documents that reference the corresponding EML). A graphical user interface for selecting different domain ontologies and annotating ecological data sets is being developed, which draw elements from EML to reduce the effort on the behalf of the user, and will eventually be integrated into a data set markup wizard.

Figure 6 is a graphical representation of how the first row of the first data table (Table I in Figure 2) might be annotated according to OBOE (Figure 1) and the example domain extensions given in this paper (Figures 3). The annotation in Figure 6 would be applied to every row of the data in Table 1. The left hand region of the figure shows the contextual hierarchy of observed entities. The observation of *Temporal Point* comes from the EML accompanying the raw data set (dashed box), and applies to the whole data set. Therefore, this temporal observation provides context for the observation of the *Reef*, and the *Reef* in turn provides context for the *Replicate Transect* (corresponding with "Site" and "Trans." in Table I, respectively). As mentioned above, because context is transitive, *Temporal Point* also provides context for the replicate transect, although the observation of the transect still depends on that of the *Reef*. Further, the replicate transect provides context for the observation of coral colony, and the coral provides context for the observation of a crab population.

At each level of observation hierarchy, measurements were taken. Note that *measurements* tend to represent single columns in data tables (or else are derived from metadata), but *observations* tend to span one or more columns. For example, when the replicate transect was observed, only its name was measured to indicate that it differed from other replicate transects at a given reef. However, additional knowledge about the replicate transect was recorded, possibly in field notes, e.g., that the transect was 30 meters long. On the other hand, four measurements (representing four separate columns of data) were taken for a given coral entity found on the replicate transect: (1) the coral colony's taxonomic name according to a given taxonomic domain, (2) the distance along to transect where the colony was found, (3) a measure of colony area, and (4) colony height. Meanwhile, within the coral colony, the population of coral crabs was observed, and the number of individuals measured. Because all the crab populations were of the same crab species, taxonomic name was absent from the raw data, but was recorded as metadata. However, there was no taxonomic domain available for coral crabs in the ontology, so only the name was given with no reference to a prescribed taxonomic domain.

Figure 7 is a graphical representation of how the first row of the first data table (Table II in Figure 2). Here, much of the semantic information is extracted from the meta-data (dashed box), i.e., when and where the observations were conducted, including geographic coordinates. Within this space-time context, a coral colony was observed, and its taxonomic name and height measured. In turn, the coral colony provides context for an observation of a population of coral crabs, where the number of individuals per unit area of the colony was measured, however data about how this calculation was made is missing and therefore implicit. Figure 5b illustrates how the crab population density can be represented using a semantic unit (where sub-units are for different entities), which can aid the data integration process outlined below.

Using the semantic annotations shown in Figures 5 and 6, the rest of this section illustrates three useful applications leveraging the formal structure of OBOE: data discovery, summarization, and integration.

3.1 Data discovery

A major application facilitated by OBOE is the capability to discover data sets based on the concepts they represent (i.e., their semantics), rather than just the labels and keywords that are used in traditional searches (Berkley, 2001). As mentioned previously, a data attribute labeled "Ht" in the first row of Table 1 is ambiguous, even to human interpretation. However, annotating this attribute with the Height (a Characteristic) from OBOE clarifies its meaning and relationship with other ontology terms (i.e., via *is-a* relations, *part-of* relations, and other description-logic constraints). An enhanced keyword search for data about "height", e.g., can leverage OBOE definitions to discover the various data sets annotated with ontology terms related to Height, such as those assigned more specific terms like Body Height (Figure 3). Search can also exploit relationships defined in OBOE Entity extension models, including the use of *part-of* relationships between classes. A search for "coral", e.g., could include entities that are *part-of* a coral colony, such as branch, tissue, skeleton, polyp, and so on. Note here that transitivity (e.g., of coral parts), and other description-logic constraints (set intersection, etc.) can also be exploited to further enhance search.

More complex forms of inference can also be used, leveraging the logical structure of OBOE. For example, measurement dimensionality can be exploited to enhance keyword search, where a search for “density” data sets will not only return those annotated with Areal Density, but also those data sets that contain the Count and Area dimensions via appropriately contextualized observations (see Figure 10). That is, this search would discover not only data having Areal Density attributes (Figure 6), but also data sets having an attribute for Area and another attribute, functionally dependent via context (Figure 5), for Count. In general, we are exploring the use of OBOE in this way for providing enhanced data-discovery query results, in which keyword queries given by scientists are: (1) expanded into their corresponding ontology classes, similar to traditional approaches based on formal terminologies (Voorhees, 1994; Moldovan and Mihalcea, 2000; Jarvelin et al., 2001); and (2) these ontology classes are compared (via *is-a*, *part-of*, and so on) to the explicit (e.g., Height) and implicit (e.g., where Count and Area imply Density) classes expressed in semantic annotations.

3.2 Data summarization

Upon finding a potentially relevant data set, an important aspect of the discovery process is to rapidly understand the content of the data set, e.g., to determine whether the data is relevant for a particular analysis. An often-used approach for understanding data content is to aggregate (i.e., *summarize*) attributes at various combinations of observation and measurement. The OBOE framework can be used to suggest appropriate data summarizations, and in so doing, also determine when a particular summarization is “sensible”. This notion of determining sensible summarizations exploits the basic structure of OBOE. For example, measurements can only be “sensibly” aggregated by other measurements that are of the same entity or measurements of entities providing observation context (i.e., “higher” contextual entities). Therefore, it makes “sense” to summarize animal height by reef name, because the observation of reef provides context for the observation of animal (Figure 8b). However, it does not make “sense” to summarize reef water temperature by animal taxon name, which gives an arbitrary average temperature of the reefs where a taxon was measured, and which is dependent on the disparity in the number of animals measured at reefs.

Replication is a mechanism to ensure unbiased parameter estimates when unable to measure an entire population, and therefore aggregating over replicate transects is also not “sensible”. Therefore, OBOE can be used to help avoid summarizing attributes by replicate samples in general (i.e., including any random, haphazard, uniform replication procedures). For instance, summarizing counts of crabs in coral colonies based on coral taxon name, transect, and reef is not “sensible”, because there are no distinguishing features between the arbitrarily named transects (Figure 8c). Moreover, summarizing by transect alone is also not “sensible” (Figure 8d), unless one wanted to test for systematic biases in observation (e.g., observer error increased as a function of the number of transects). Whereas, summarizations that include replicate observations make “sense”, as shown in Figures 8b and e.

In general, the logical structure and constraints of OBOE can be used to test the usefulness of various statistical operations and modeling procedures. For example, when summarizing a nominal or categorical variable, the aggregate count is applicable, whereas continuous aggregations (e.g., sum, average, maximum, standard deviation, etc.) are not. Moreover, types of modeling approaches can be recommended based on measurement types. For example, parametric linear models require continuous measures, whereas the inclusion of a categorical variable (i.e., a classification domain measurement standard) requires non-parametric model fitting approaches. By making the semantics of observation and measurement explicit in a formal structure such as OBOE, automated inferencing procedures (i.e., "machine reasoning") can be applied to enhance the various decision making processes used in exploring and modeling observational data sets.

3.3 Data integration

Another major application that OBOE facilitates is the capability to determine if two data sets can be either fully or partially merged once they are discovered (and vetted, e.g., through summarization). To do so, a number of steps must be taken, starting at the lowest level semantic resolution (i.e., at the level of single measurements and observations), and building up to the data set level (e.g., aligning observation context). As a simple example, assume we are interested in merging data about heights following the discovery of the two data tables in Figure 2. Figure 9 illustrates the reasoning process involved in merging the first height observations in each data table, which is based on the semantic annotations using OBOE shown in Figures 5 and 6. The first step is to determine if, and at what semantic resolution, the data are compatible. At the observation level, instances of the Animal and Coral entities are compatible at the semantic the resolution of Animal. Observation **b** loses semantic resolution when standardized for the merge (i.e., the Animal entity class is the "lowest common denominator"). Similarly for measurement, instances of the Height and Width characteristics are compatible at the resolution of Length (i.e., Length is the "common ancestor" of the Height and Width characteristics), where both instances in this case lose semantic resolution. Finally, the measurement standards used for the two measurements have the same dimension, and therefore they are compatible under OBOE's model. Assuming merging lengths of coral in this way is desirable (e.g., for a particular analysis), Figure 9 illustrates the resulting merged data, where both observations are now of type Animal, both measurements are of characteristic Length, both measurement standards are in Meters, and the values are recalculated at the coarsest level of precision.

As a more sophisticated example, assume the result of a discovery search for data about "animal" "density" on "reefs" returns the two example data sets in Figure 2. The data in Table I were discovered for the reasons discussed above. To integrate the data pertaining to areal densities, the OBOE structure can be used to determine if the data are compatible and then calculate the merge. The numerator in a semantic unit typically refers to the focal entity (count of individuals), whereas the denominator is contextual (area of coral). Therefore, semantics units are in essence a nested contextual dependency (required when more detailed information about context is missing). Therefore, if a focal entity is compatible with the numerator of the semantic unit, and the contextual entity is compatible with the denominator (e.g., Figure 10a and 10b), the merge result shown at the bottom of Figure 10 can be automatically computed. In the case of the areal densities of crab populations, Figure 10a and 10b illustrate the semantic equivalencies between the two data tables, and calculates the merge, where "**a**" loses resolution when being coersed into the semantic unit form. However, knowledge about the coral area and crab population in data Table I can still be retained (Table III), although there are no corresponding measurements from data Table II.

4 Summary

This paper presents the OBOE ontology framework for capturing the process of ecological field observation and measurement, facilitating logic-based reasoning (via description logic and OWL-DL) to be utilized to automate important data-management applications for data synthesis. The ontology formalizes an interpretation of observation, which focuses on observing a concrete or conceptual *entity*, and measuring one or more of the entity's *characteristics* by comparison with a *measurement standard*. In general, each data point in an observational data set is an instance of observation, and can provide contextual information for other instances of observation. Semantic annotations define a standard representation for mapping observational data to the ontology, and can be exploited in data discovery and integration applications. Annotation via OBOE and associated domain extensions makes explicit the basic definition of data and their relationships with other data, allowing annotated data sets to be easily contrasted. This approach can facilitate more powerful data discovery and integration approaches, and can provide guidance for, and automate, data aggregation and summary.

Annotation to OBOE enables compatibility testing among data attributes, both at the level of the attribute (i.e., is the entity, characteristic, and measurement standard compatible?) and the data-set level (i.e., are entity nesting structures in two independent data sets compatible?). If compatible, the ontology contains the necessary details (i.e., constraints) to conduct the appropriate conversions so that data can be merged. Finally, OBOE provides a structured approach for creating domain-specific ontologlies, allowing new ontologies to extend core OBOE classes. In this way, OBOE can also be used as a "glue structure" to incorporate and inter-relate existing domain ontologies, allowing otherwise *ad hoc* ontologies to be structured and placed within a broader, cross-discipline scientific context.

Our ongoing and future work includes the development of an easy-to-use graphical user interface for data annotation based on OBOE. This tool will leverage existing EML metadata definitions (e.g., for basic data structure information and measurement units), will leverage OBOE ontology constraints to help direct and fill-in annotations when possible, and will transparently store semantic annotations using the XML serialization syntax mentioned in Section 3. We are also using OBOE as the foundation ontology in the SEEK Semantic Mediation System, which will take semantic annotations as input, and provide discovery, summarization, and integration services for use within, e.g., the Kepler scientific workflow system (Ludäscher, 2006; Berkley, 2005) and the EML-based Morpho application (Higgins, 2002).

References

Baader F., Calvanese D., McGuinness D., Nardi, D. and Patel-Schneider P., 2003. *The Description Logic Handbook: Theory, Implementation, and Applications*. Cambridge University Press.

Bard, J. R. L. and Rhee, S. Y., 2004. Ontologies in biology: Design, applications and future challenges. *Nat. Rev. Genet.* 5, 213-222.

Batini C, Ceri S, Navathe SB. 1992. Conceptual Database Design: An Entity-Relationship Approach. Redwood City, CA: Benjamin Cummings

Berkley, C., Jones, M. B., Bojilova, J., and Higgins, D., 2001. Metacat: a Schema-Independent XML Database System. *Proc. of the 13th Intl. Conf. on Scientific and Statistical Database Management*. IEEE Computer Society.

Berkley, C., Bowers, S., Jones, M. B., Ludaescher, B., Schildhauer, M. and Tao, J., 2005. Incorporating Semantics in Scientific Workflow Authoring. Proceedings of the 17th International Conference on Scientific and Statistical Database Management. *IEEE Computer Society*.

Bowers, S. and Ludäscher, B., 2006. A calculus for propagating semantic annotations through scientific workflow queries. Proc. of the Wkshp. on Query Langauges and Query Processing. *Lect. Notes Comput.,* 4254, 712–723.

Bowers, S., Thau, D., Williams, R. and Ludäscher, B., 2005. Data procurement for enabling scientific workflows: On exploring inter-ant parasitism. Proc. of the Intl. Wkshp. on SemanticWeb and Databases, *Lect. Notes Comput.,* 3372.

Brilhante, V. B., 2003. Ontology and reuse in model synthesis. PhD thesis. Univ. Edinburgh

Bunge, M., 1977. *Treatise on Basic Philosophy: vol. 1, Ontology 1: Furniture of the Word*. Reidel, Boston.

Chen, P.P., 1976. The Entity-Relationship Model: Toward a Unified View of Data. *ACM TODS* 1, 9-36.

Davies J., Fensel, D., van Harmelen, F., 2003. *Towards the Semantic Web: Ontology-Driven Knowledge Management*. John Wiley & Sons, London.

Darwin Core, 2004. Darwin Core Schema (version 1.3), A draft standard of the Taxonomic Database Working Group (TDWG), http://wiki.tdwg.org/DarwinCore

DCMI, 2006. DCMI Metadata Terms. http://www.dublincore.org/documents/dcmi-terms

Ecological Metadata Language (EML) Specification. http://knb.ecoinformatics.org/software/eml/

The Gene Ontology Consortium, 2000. Gene Ontology: tool for the unification of biology. *Nat. Rev. Genet*. 25, 25-29.

Hammer J, McLeod D. 1999. Resolution of representational diversity in multidatabase systems. In Management of Heterogeneous and Autonomous Database Systems, ed. A Elmagarmid, M Rusinkiewicz, A Sheth, 4:91–117. San Francisco: Morgan Kaufmann. 413 pp.

Higgins, D., Berkley. C., and Jones, M. B., 2002. Managing Heterogeneous Ecological Data using Morpho. *Proc. of the 14th Intl. Conf. on Scientific and Statistical Database Management*.

Fegraus, E.H., Andelman, S., Jones, M.B. and Schildhauer, M., 2005. Maximizing the Value of Ecological Data with Structured Metadata: An Introduction to Ecological Metadata Language (EML) and Principles for Metadata Creation. *Bull. Ecological Soc. Am.* 86, 158-168.

Gray, J., Chaudhuri, S., Bosworth, A., Layman, A. , Reichart, D., Venkatrao, M., Pellow, F. and Pirahesh, H., 1997. Data Cube: A Relational Aggregation Operator Generalizing Group-By, Cross-Tab, and Sub-Totals. *Data Min. Knowl. Disc.* 1, 29-53.

Gross, K. L. and Pake, C.E., eds. (1995) The Future of Long-term Ecological Data. A Report to the Ecological Society of America. Vol. I: Text of the Report. 123 pp. Vol. II: Directories to Sources of Long-term Ecological Data. 114 pp.

Gruber, T. R., and Olsen, G.R., (1994). An Ontology for Engineering Mathematics. Proceedings of the Intl. Conf. on Principles of Knowledge Representation and Reasoning (KR), 258-269.

Guarino, N. and Welty, C. A., 2002. Evaluating ontological decisions with OntoClean. *Commun. ACM* 45, 61-65.

Jacobson, I., Booch, G. and Rumbaugh, J., 1998. *The Unified Software Development Process*. Addison Wesley Longman.

Jarvelin, K., Kekalainen, J., and Niemi, T., 2001. ExpansionTool: Concept-based query expansion and construction. *Information Retrieval*, 4(3-4):231-255.

Jones, M., Berkley, C., Bojilova, J. and Schildhauer, M., 2001. Managing Scientific Metadata. *IEEE Internet Comput.* 5, 59-68.

Jones, M. B., Schildhauer M., Reichman O. J., and Bowers S., 2006. The New Bioinformatics: Integrating ecological data from the gene to the biosphere. *Annu. Rev. Ecol. Evol. Syst.* 37, 519-544.

Keet, C. M., 2005. Factors Affecting Ontology Development in Ecology. *Lect. Notes Comput.*, 3615, 46-62.

Kennedy, J.B., Kukla, J. and Paterson, T., 2005. Scientific Names Are Ambiguous as Identifiers for Biological Taxa: Their Context and Definition Are Required for Accurate Data Integration. *Lect. Notes Comput. Sc.* 3615, 80-95.

Kepler Project, 2006. Kepler: An Extensible Scientific Workflow System. http://kepler-project.org.

Ludäscher, B., Altintas, I., Berkley, C., Higgins, H., Jaeger, E., Jones, M., Lee, E. A., Tao, J., and Zhao, Y., 2006. Scientific Workflow Management and the Kepler System. *Concurrency and Computation: Practice & Experience*, 18(10):1039–1065.

McGuinness, D.L. and van Harmelen, F. eds., 2004. OWL Web Ontology Language Overview, W3C Recommendation 10 Feb 2004, http://www.w3.org/TR/2004/REC-owl-features-20040210/

Michener, W. K., 2000. Research Design: Translating Ideas to Data. In: W. K. Michener and J. W. Brunt (Eds.), *Ecological Data: Design, Management and Processing* (pp. 1-24). Blackwell Science, Oxford.

Michener, W. K., Brunt, J. W., Helly, J. J., Kirchner, T. B. and Stafford, S. G., 1997. Non-geospatial metadata for the ecological sciences. *Ecol. Appl*. 7, 330-342.

Moldovan D.I., and Mihalcea, R., 2000. Using WordNet and lexical operators to improve internet searchers. *IEEE Internet Computing*, 4(1):34-43.

Murray-Rust, P. and Rzepa, H. S., 2002. STMML. A markup language for scientific, technical and medical publishing. *Data Science Journal* 1, 1-65.

The Observation Ontology, OBOE. http://ecoinformatics.org/ontologies/observation-0.1.0.

Parsons, J. and Wand, Y., 2000. Emancipating instances from the tyranny of classes in information modeling. *ACM T. Database Syst.* 25, 228-268.

Pickett, S. T. A., Kolasa, J. and Jones, C. G., 1994. *Ecological Understanding: The Nature of Theory and the Theory of Nature*. Academic Press, San Diego.

Rosse C., and Mejino J. L. V., Jr., 2003. A reference ontology for bioinformatics: the Foundational Model of Anatomy. *J Biomed Inform.*, 36, 478-500.

R Development Core Team, 2005. R: A language and environment for statistical computing. R Foundation for Statistical Computing,Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Smith, B., 1996. Mereotopology: A Theory of Parts and Boundaries. *Data Knowl. Eng.* 20, 287–303.

Smith, B., 2001. Objects and their environments: From Aristotle to ecological ontology. In: Frank, A., Raper, J., and Cheylan, J. P. (eds.), *The Life and Motion of Socio-Economic Units* (pp. 79-97). London: Taylor and Francis.

Smith, B., and Varzi, A., 1999. The Niche, *Nous,* 33(2):198–222.

Smith, B., and Varzi, A., 1999. The Formal Structure of Ecological Contexts. In P. Bouquet, P. Brezillon, L. Serafini, M. Beneceretti, F. Castellani (eds.), *Modeling and Using Context*, *Lect. Notes Artif. Int.,* 1688, 339–350.

Sowa, J. F., 1999. *Knowledge Representation: Logical, Philosophical, and Computational Foundations*. PWS Publishing Co., Boston.

Stevens, S. S., 1946. On the theory of scales of measurement. *Science* 103, 677-680.

Villa, F., in press. A semantic framework and software design to enable the transparent integration, reorganization and discovery of natural systems knowledge. *J. Intell. Inf. Syst.* XX, XX-XX.

Voorhees, E.M., 1994. Query expansion using lexical-semantic relations. *Proc. of the Annual Intl. ACM-SIGIR Conf. on Research and Development in Information Retrieval* (SIGIR).

Wand, Y., Storey, V.C. and Weber, R., 1999. An Ontological Analysis of the Relationship Construct in Conceptual Modeling. *ACM T. Database Syst.* 24, 494-528.

Figure captions

Figure 1. The core classes (ellipses) and properties (arrows) of the Extensible Observation Ontology (OBOE). Each Observation is of some Entity, and can provide context for the Observation of another Entity. A Characteristic of an Entity can be represented through a Measurement. Measurements relate Characteristics to a Measurement Standard via a Value and, if applicable, a Precision. Observations may have multiple Measurements. Entity (**a**), Characteristic (**b**), and Measurement Standard (**c**) (i.e., the shaded classes) provide extension points for domain-specific ontologies (see Figure 3). Numbers in parentheses denoted min:max cardinality for properties.

Figure 2. Example data sets where: **I** contains data about coral crabs living within different coral species, including the given location and replicate transect, the species name, distance along the transect, colony area for coral colonies, and the number of crabs found in each colony; **II** contains data (from another study) about the density of coral crabs in different coral colonies, including the species name and crab density (other attributes, such as date and location, are described in the metadata, e.g., field notes); and **III** contains the results of merging **I** and **II**.

Figure 3. Detailed representation of OBOE core classes and examples of ontology extensions for Entity, Characteristic, and Measurement Standard. Shaded ellipses represent OBOE core classes, and open ellipses represent domain extensions for the examples given in this paper.

Figure 4. Examples of OBOE observational-data instantiations: (**a**) represents the observation of a coral where the coral height is measured in meters; (**b**) represents the observation of a community (an ecological concept) where diversity of taxa is measured using the Shannon Diversity index; (**c**) represents the observation of an animal classified as female via reference to classification standard; (**d**) represents the observation of a reef entity classified as a "bird", without reference to a specific classification standard; (**e**) represents the maximum depth of a particular reef; (**f**) represents the name of a particular reef according to a naming standard for the lizard island reef; (**g**) represents a trivial observation of an *Acropora hyacinthus* entity; (**h**) represents an assertion of an entities taxa, without an associated naming standard; and **i** represents a similar observation as (**h**), but where the naming standard is given.

Figure 5. Example OBOE Measurement Standard extensions: (**a**) a simple derived unit; (**b**) a complex derived unit composed of two other units, as well as corresponding 'semantic units' (dashed ellipse), i.e., units linked to particular types of entities; (**c**) an ordered classification domain (i.e., an ordinal scale); and (**d**) three (nominal) classification domains, one for male/female values, geographic names, and taxonomic names with "according to" relations.

Figure 6. OBOE representation of the first row of data in Table **I** in Figure 2. The colon placed before a class name denotes the instantiation of the class, i.e., ": Distance" is a member of the ontology class "Distance." Solid ellipses represent the observation and

measurement structure of the data set, including metadata (dashed box). Open ellipses represent terms selected from ontology extensions for each of Entity, Characteristics, and Measurement Standard. White squares represent the physical values from the data table or the metadata, and *Precision* where applicable.

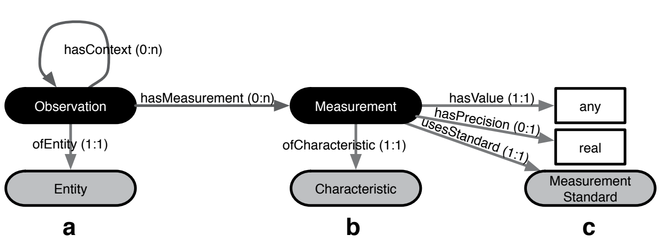
Figure 7. Formal OBOE representation of the first row of data in Table **II** in Figure 2. Dashed box represents metadata. **b** corresponds with example **b** in Figure 5, illustrating the more detailed representation using the semantic unit. See text for details.

Figure 8. Graphical examples of data summarization leveraging OBOE's formal structure. **(a)** The basic observation and measurement structure for a subset of Table I. **(b)** Summarizing animal height by reef site ("sensible"). **(c)** Summarizing number of crab by animal species, replicate transect, and reef site (*not* "sensible"). **(d)** Summarizing average number of coral crabs by replicate transect name (*not* "sensible"). **(e)** Summarizing average number of coral crabs replicate transect ("sensible"). See text for details.

Figure 9. Merging measurements of animal height and coral length (from Tables I and II, respectively) based on semantic annotation with terms from OBOE (Figures 5 and 6). See text for details.

Figure 10. Merging measurements of crab population count and coral area (from Table I) and crab population areal density (from Tables II) based on semantic annotation with terms from OBOE (Figures 5 and 6). See text for details.

Figure 1



hasContext (0:n)
Observation
hasMeasurement (0:n)
Measurement
hasValue (1:1)
any
hasPrecision (0:1)
real
usesStandard (1:1)
ofEntity (1:1)
ofCharacteristic (1:1)
Entity
Characteristic
Measurement Standard
a
b
c

Figure 2

I

Site	Temp.	Trans.	Sp.	Dist.	Ht.	Area	Crabs
bird	27	1	hya	0.55	0.46	1260	2
bird	27	1	hya	5.90	0.44	2830	4
bird	27	1	hya	19.35	0.13	180	0
bird	27	2	hya	3.55	0.32	5030	6
bird	27	2	hya	18.20	0.58	17670	18
bird	27	2	hya	29.75	0.15	310	1
south	25	1	hya	4.15	0.08	20	0
south	25	1	hya	15.00	0.34	4420	7
south	25	1	hya	20.05	0.45	6360	7
south	25	2	hya	12.30	0.40	1260	2
...	...	...	...	...	...	...	...

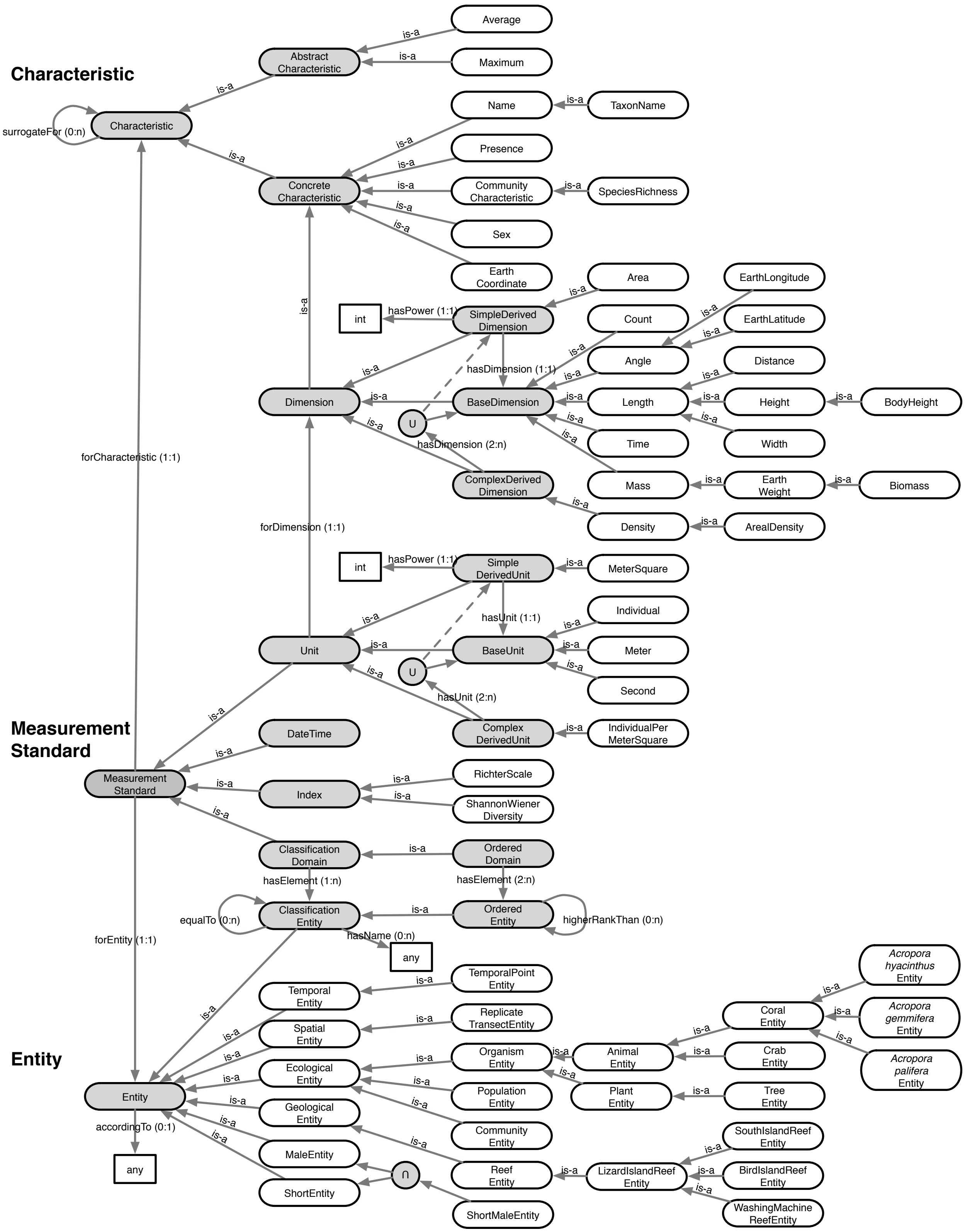
II

Species	Wth.	Dens.
A. hyacinthus	30	6
A. hyacinthus	40	19
A. hyacinthus	20	4
A. hyacinthus	20	4
A. gemmifera	10	3
A. gemmifera	20	1
A. gemmifera	20	2
A. gemmifera	10	3
A. gemmifera	10	1
A. palifera	10	0
...	...	...

III

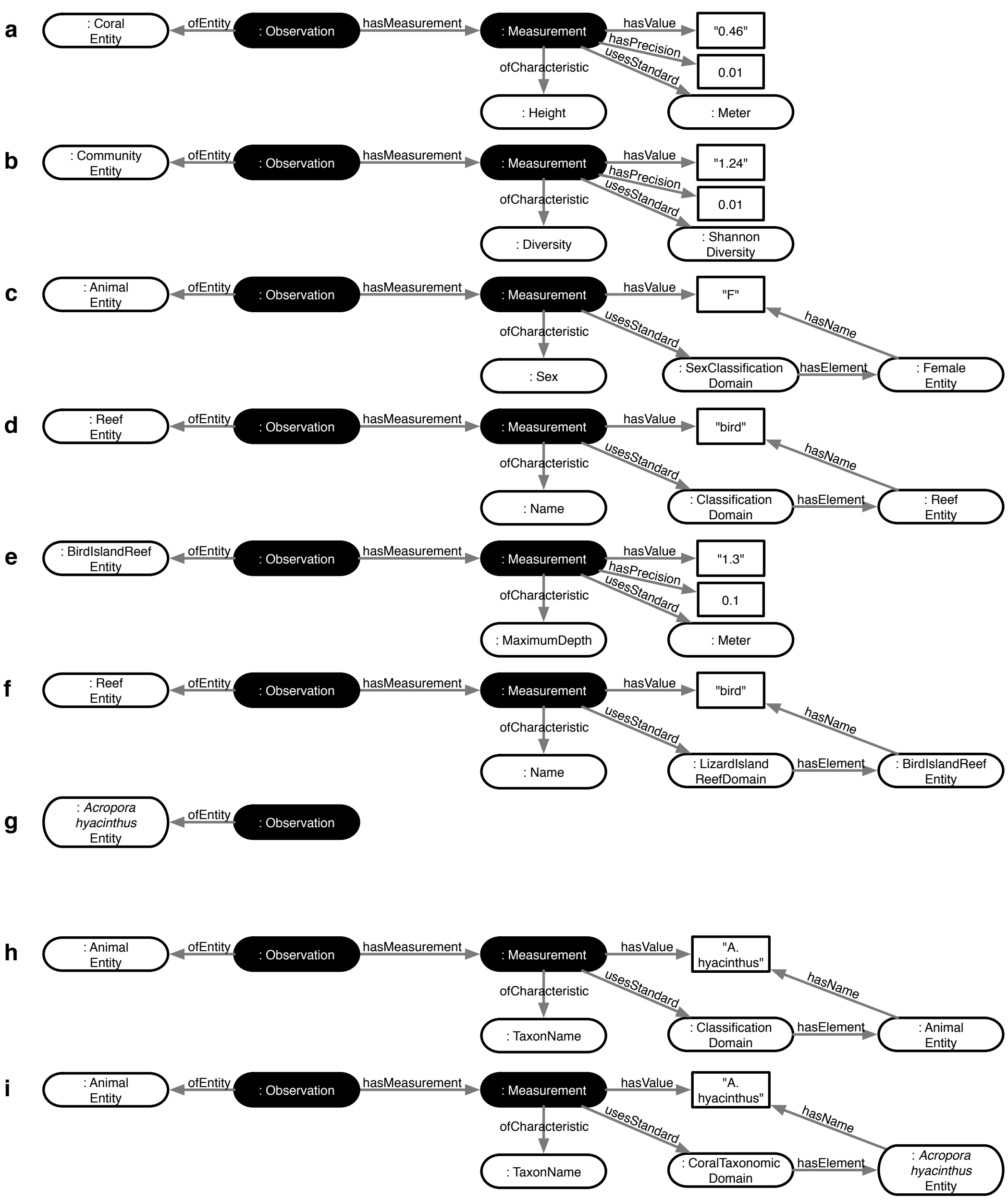
Date	Site	Temperature	Transect	Species	Distance	Length	Area	Crabs	Density
12-08-2006	Bird Island	27	1	A. hyacinthus	0.55	0.5	1260	2	16
12-08-2006	Bird Island	27	1	A. hyacinthus	5.90	0.4	2830	4	14
12-08-2006	Bird Island	27	1	A. hyacinthus	19.35	0.1	180	1	0
12-08-2006	Bird Island	27	2	A. hyacinthus	3.55	0.3	5030	6	12
12-08-2006	Bird Island	27	2	A. hyacinthus	18.20	0.6	17670	13	10
12-08-2006	Bird Island	27	2	A. hyacinthus	29.75	0.2	310	4	32
12-08-2006	South Island	25	1	A. hyacinthus	4.15	0.1	20	3	0
12-08-2006	South Island	25	1	A. hyacinthus	15.00	0.3	4420	7	16
12-08-2006	South Island	25	1	A. hyacinthus	20.05	0.5	6360	7	11
12-08-2006	South Island	25	2	A. hyacinthus	12.30	0.4	1260	6	16
...	...	...	...	...	...	...	...	...	...
10-18-2006	South Island	24		A. hyacinthus		0.3			6
10-18-2006	South Island	24		A. hyacinthus		0.4			19
10-18-2006	South Island	24		A. hyacinthus		0.2			4
10-18-2006	South Island	24		A. hyacinthus		0.2			4
10-18-2006	South Island	24		A. gemmifera		0.1			3
10-18-2006	South Island	24		A. gemmifera		0.2			1
10-18-2006	South Island	24		A. gemmifera		0.2			2
10-18-2006	South Island	24		A. gemmifera		0.1			3
10-18-2006	South Island	24		A. gemmifera		0.1			1
10-18-2006	South Island	24		A. palifera		0.1			0
...	...	...		...		...			...

Figure 3



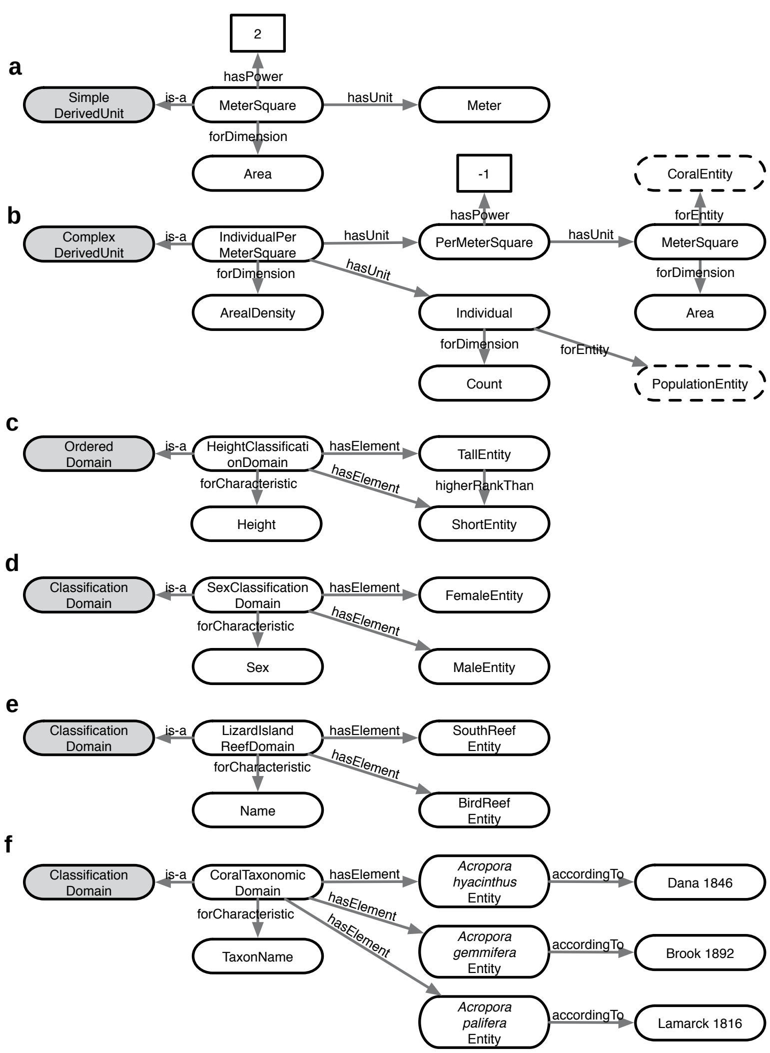
Characteristic
Measurement Standard
Entity
Average
Maximum
Abstract Characteristic
Characteristic
surrogateFor (0:n)
Name
TaxonName
Presence
Community Characteristic
SpeciesRichness
Concrete Characteristic
Sex
Earth Coordinate
Area
EarthLongitude
int
hasPower (1:1)
SimpleDerived Dimension
Count
EarthLatitude
Angle
Distance
hasDimension (1:1)
Dimension
BaseDimension
Length
Height
BodyHeight
U
Time
Width
hasDimension (2:n)
forCharacteristic (1:1)
ComplexDerived Dimension
Mass
Earth Weight
Biomass
Density
ArealDensity
forDimension (1:1)
Simple DerivedUnit
MeterSquare
hasUnit (1:1)
Individual
Unit
BaseUnit
Meter
Second
hasUnit (2:n)
Complex DerivedUnit
IndividualPer MeterSquare
DateTime
Measurement Standard
Index
RichterScale
ShannonWiener Diversity
Classification Domain
Ordered Domain
hasElement (1:n)
hasElement (2:n)
equalTo (0:n)
Classification Entity
Ordered Entity
higherRankThan (0:n)
hasName (0:n)
any
forEntity (1:1)
Acropora hyacinthus Entity
TemporalPoint Entity
Temporal Entity
Coral Entity
Replicate TransectEntity
Acropora gemmifera Entity
Spatial Entity
Organism Entity
Animal Entity
Crab Entity
Acropora palifera Entity
Ecological Entity
Population Entity
Plant Entity
Tree Entity
Geological Entity
Community Entity
accordingTo (0:1)
SouthIslandReef Entity
MaleEntity
Reef Entity
LizardIslandReef Entity
BirdIslandReef Entity
∩
ShortEntity
ShortMaleEntity
WashingMachine ReefEntity
is-a

Figure 4



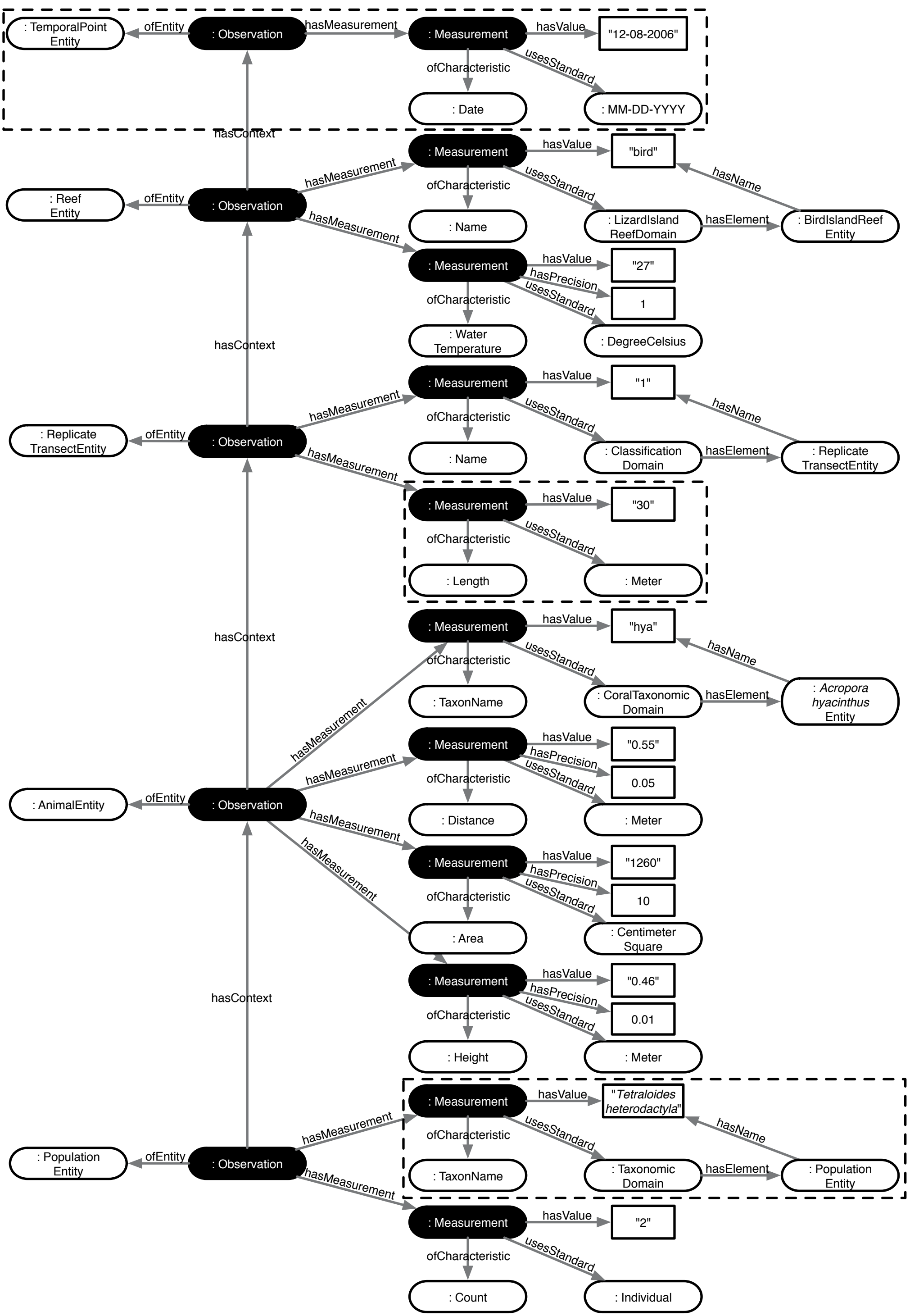
a
: Coral Entity
ofEntity
: Observation
hasMeasurement
: Measurement
hasValue
"0.46"
hasPrecision
0.01
usesStandard
: Meter
ofCharacteristic
: Height
b
: Community Entity
: Diversity
"1.24"
: Shannon Diversity
c
: Animal Entity
"F"
hasName
: Sex
: SexClassification Domain
hasElement
: Female Entity
d
: Reef Entity
"bird"
: Name
: Classification Domain
e
: BirdIslandReef Entity
"1.3"
0.1
: MaximumDepth
f
: LizardIsland ReefDomain
g
: *Acropora hyacinthus* Entity
h
"A. hyacinthus"
: TaxonName
i
: CoralTaxonomic Domain

Figure 5



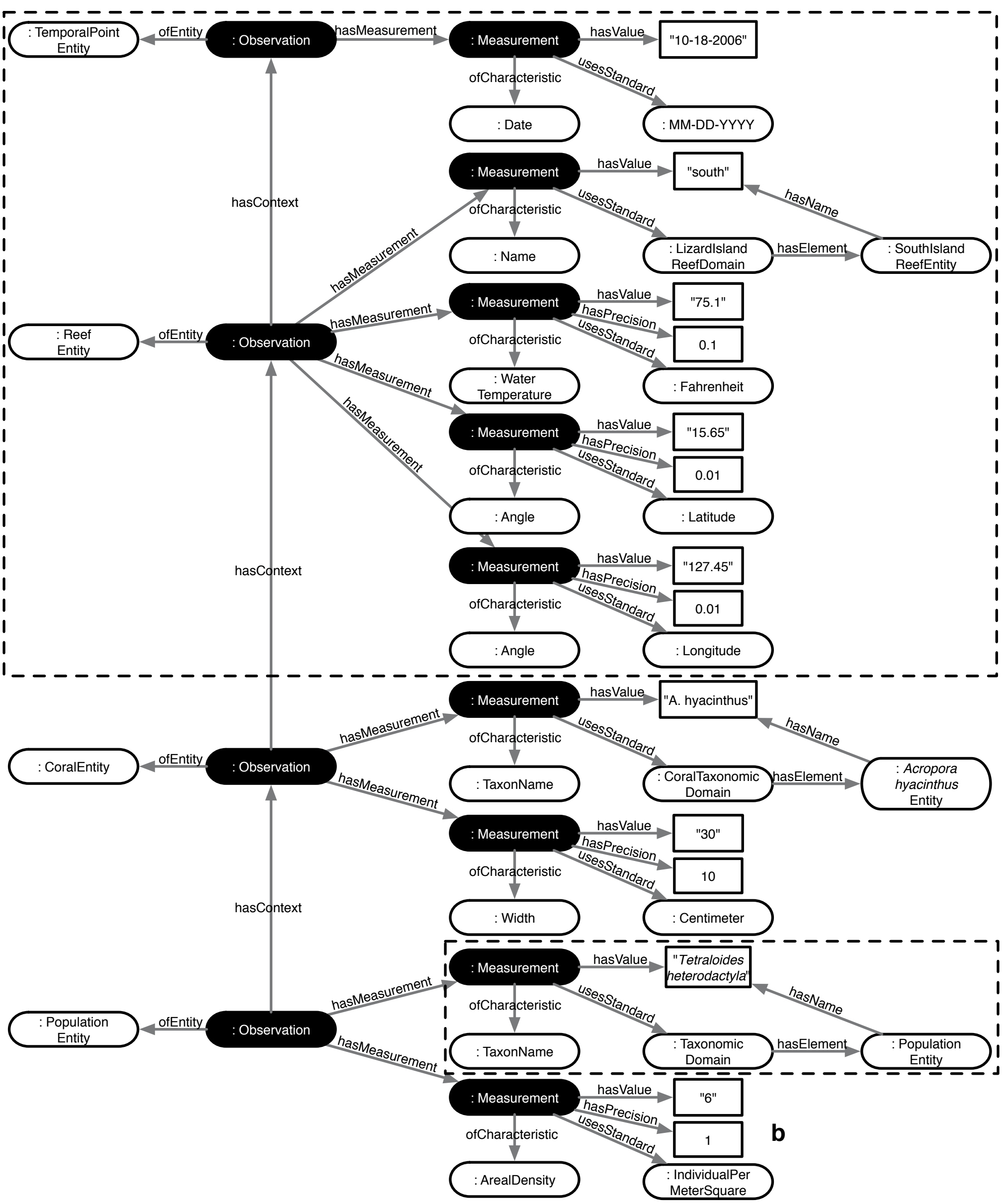
a
2
hasPower
Simple DerivedUnit
is-a
MeterSquare
hasUnit
Meter
forDimension
Area
b
-1
CoralEntity
hasPower
forEntity
Complex DerivedUnit
is-a
IndividualPer MeterSquare
hasUnit
PerMeterSquare
hasUnit
MeterSquare
forDimension
hasUnit
forDimension
ArealDensity
Individual
Area
forDimension
forEntity
Count
PopulationEntity
c
Ordered Domain
is-a
HeightClassificationDomain
hasElement
TallEntity
forCharacteristic
hasElement
higherRankThan
Height
ShortEntity
d
Classification Domain
is-a
SexClassification Domain
hasElement
FemaleEntity
forCharacteristic
hasElement
Sex
MaleEntity
e
Classification Domain
is-a
LizardIsland ReefDomain
hasElement
SouthReef Entity
forCharacteristic
hasElement
Name
BirdReef Entity
f
Classification Domain
is-a
CoralTaxonomic Domain
hasElement
Acropora hyacinthus Entity
accordingTo
Dana 1846
forCharacteristic
hasElement
hasElement
TaxonName
Acropora gemmifera Entity
accordingTo
Brook 1892
Acropora palifera Entity
accordingTo
Lamarck 1816

Figure 6



: TemporalPoint Entity
ofEntity
: Observation
hasMeasurement
: Measurement
hasValue
"12-08-2006"
ofCharacteristic
usesStandard
: Date
: MM-DD-YYYY
hasContext
: Measurement
hasValue
"bird"
hasName
: Reef Entity
ofEntity
: Observation
hasMeasurement
ofCharacteristic
usesStandard
: Name
: LizardIsland ReefDomain
hasElement
: BirdIslandReef Entity
hasMeasurement
: Measurement
hasValue
"27"
hasPrecision
1
usesStandard
ofCharacteristic
: Water Temperature
: DegreeCelsius
hasContext
: Measurement
hasValue
"1"
hasName
: Replicate TransectEntity
ofEntity
: Observation
hasMeasurement
ofCharacteristic
usesStandard
: Name
: Classification Domain
hasElement
: Replicate TransectEntity
hasMeasurement
: Measurement
hasValue
"30"
ofCharacteristic
usesStandard
: Length
: Meter
hasContext
: Measurement
hasValue
"hya"
hasName
ofCharacteristic
usesStandard
: TaxonName
: CoralTaxonomic Domain
hasElement
: *Acropora hyacinthus* Entity
hasMeasurement
: Measurement
hasValue
"0.55"
hasPrecision
0.05
usesStandard
ofCharacteristic
: AnimalEntity
ofEntity
: Observation
hasMeasurement
: Distance
: Meter
hasMeasurement
: Measurement
hasValue
"1260"
hasPrecision
10
usesStandard
ofCharacteristic
hasMeasurement
: Area
: Centimeter Square
: Measurement
hasValue
"0.46"
hasPrecision
0.01
usesStandard
hasContext
ofCharacteristic
: Height
: Meter
: Measurement
hasValue
"*Tetraloides heterodactyla*"
hasName
hasMeasurement
ofCharacteristic
usesStandard
: Population Entity
ofEntity
: Observation
: TaxonName
: Taxonomic Domain
hasElement
: Population Entity
hasMeasurement
: Measurement
hasValue
"2"
ofCharacteristic
usesStandard
: Count
: Individual

Figure 7

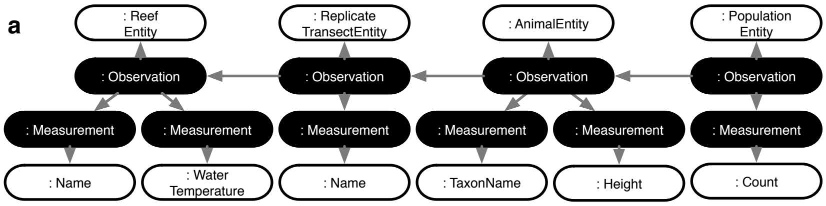


: TemporalPoint Entity
ofEntity
: Observation
hasMeasurement
: Measurement
hasValue
"10-18-2006"
ofCharacteristic
usesStandard
: Date
: MM-DD-YYYY
hasContext
: Measurement
hasValue
"south"
hasName
ofCharacteristic
usesStandard
: Name
: LizardIsland ReefDomain
hasElement
: SouthIsland ReefEntity
hasMeasurement
: Measurement
hasValue
"75.1"
hasPrecision
0.1
usesStandard
: Reef Entity
ofEntity
: Observation
hasMeasurement
ofCharacteristic
: Water Temperature
: Fahrenheit
hasMeasurement
hasMeasurement
: Measurement
hasValue
"15.65"
hasPrecision
0.01
usesStandard
ofCharacteristic
: Angle
: Latitude
: Measurement
hasValue
"127.45"
hasPrecision
0.01
usesStandard
hasContext
ofCharacteristic
: Angle
: Longitude
: Measurement
hasValue
"A. hyacinthus"
hasName
hasMeasurement
ofCharacteristic
usesStandard
: CoralEntity
ofEntity
: Observation
: TaxonName
: CoralTaxonomic Domain
hasElement
: *Acropora hyacinthus* Entity
hasMeasurement
: Measurement
hasValue
"30"
hasPrecision
10
usesStandard
ofCharacteristic
hasContext
: Width
: Centimeter
: Measurement
hasValue
"*Tetraloides heterodactyla*"
hasName
hasMeasurement
ofCharacteristic
usesStandard
: Population Entity
ofEntity
: Observation
: TaxonName
: Taxonomic Domain
hasElement
: Population Entity
hasMeasurement
: Measurement
hasValue
"6"
hasPrecision
1
usesStandard
ofCharacteristic
: ArealDensity
: IndividualPer MeterSquare

b

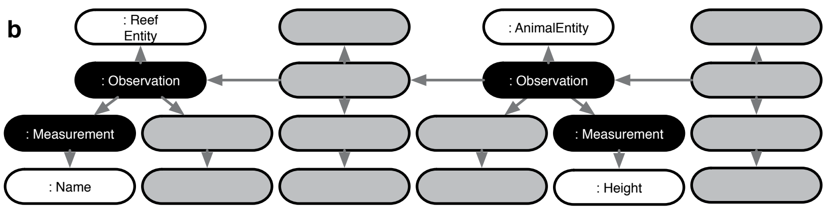
Figure 8

a



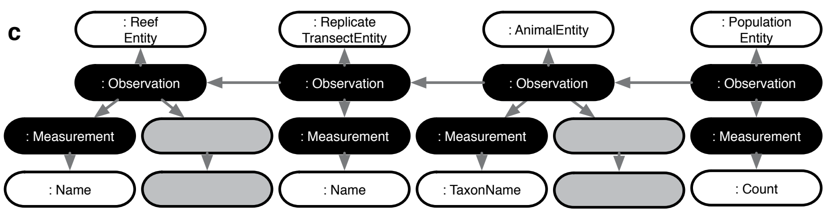
Site	Temp.	Trans.	Sp.	Ht.	Crabs
bird	27	1	hya	0.46	2
bird	27	1	hya	0.44	4
bird	27	1	hya	0.13	0
bird	27	2	hya	0.32	6
bird	27	2	hya	0.58	18
bird	27	2	hya	0.15	1
south	25	1	hya	0.08	0
south	25	1	hya	0.34	7
south	25	1	hya	0.45	7
south	25	2	hya	0.40	2
...	...	...	...	...	...

b



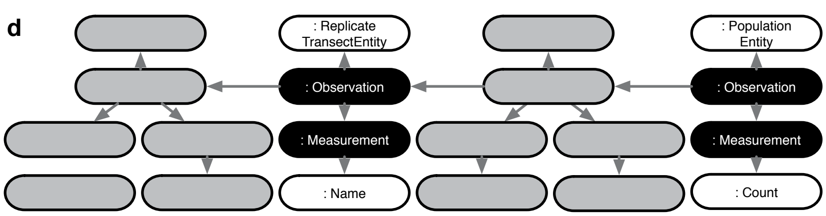
Site	Average Ht.
bird	0.35
south	0.32
...	...

c



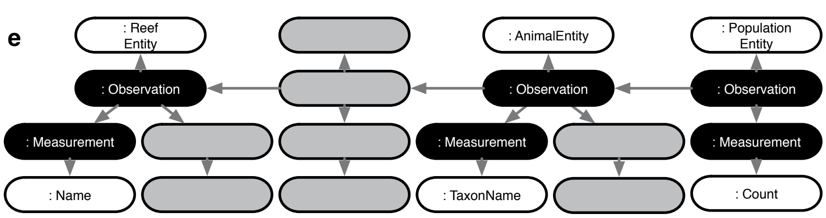
Site	Trans.	Sp.	Average Crabs
bird	1	hya	2
bird	2	hya	8.3
south	1	hya	4.7
south	2	hya	2
...	...	...	...

d



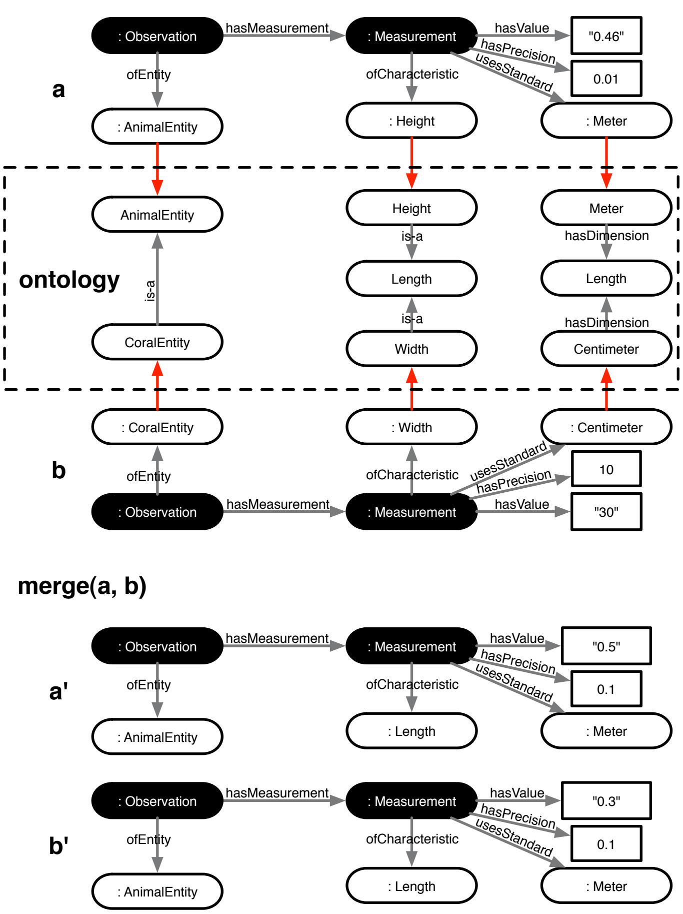
Trans.	Average Crabs
1	3.3
2	6.8
...	...

e



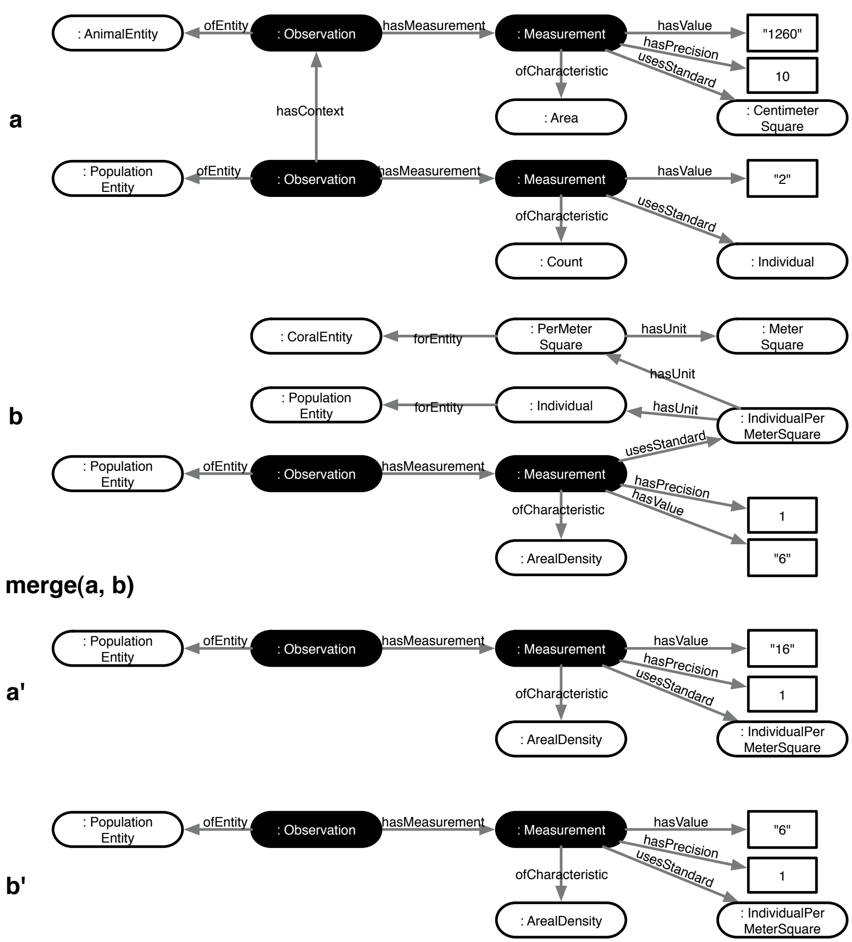
Site	Sp.	Average Crabs
bird	hya	5.2
south	hya	4
...	...	...

Figure 9



: Observation
hasMeasurement
: Measurement
hasValue
"0.46"
hasPrecision
0.01
usesStandard
a
ofEntity
ofCharacteristic
: AnimalEntity
: Height
: Meter
AnimalEntity
Height
Meter
is-a
hasDimension
ontology
is-a
Length
Length
is-a
hasDimension
CoralEntity
Width
Centimeter
: CoralEntity
: Width
: Centimeter
b
ofEntity
ofCharacteristic
usesStandard
10
hasPrecision
: Observation
hasMeasurement
: Measurement
hasValue
"30"
merge(a, b)
: Observation
hasMeasurement
: Measurement
hasValue
"0.5"
hasPrecision
0.1
usesStandard
a'
ofEntity
ofCharacteristic
: AnimalEntity
: Length
: Meter
: Observation
hasMeasurement
: Measurement
hasValue
"0.3"
hasPrecision
0.1
usesStandard
b'
ofEntity
ofCharacteristic
: AnimalEntity
: Length
: Meter

Figure 10



a
: AnimalEntity
ofEntity
: Observation
hasMeasurement
: Measurement
hasValue
"1260"
hasPrecision
10
usesStandard
: Centimeter Square
ofCharacteristic
: Area
hasContext
: Population Entity
ofEntity
: Observation
hasMeasurement
: Measurement
hasValue
"2"
ofCharacteristic
usesStandard
: Count
: Individual
b
: CoralEntity
forEntity
: PerMeter Square
hasUnit
: Meter Square
hasUnit
: Population Entity
forEntity
: Individual
hasUnit
: IndividualPer MeterSquare
usesStandard
: Population Entity
ofEntity
: Observation
hasMeasurement
: Measurement
hasPrecision
hasValue
1
"6"
ofCharacteristic
: ArealDensity
merge(a, b)
a'
: Population Entity
ofEntity
: Observation
hasMeasurement
: Measurement
hasValue
"16"
hasPrecision
1
usesStandard
ofCharacteristic
: ArealDensity
: IndividualPer MeterSquare
b'
: Population Entity
ofEntity
: Observation
hasMeasurement
: Measurement
hasValue
"6"
hasPrecision
1
usesStandard
ofCharacteristic
: ArealDensity
: IndividualPer MeterSquare