The OBOE Tutorial

Joshua S. Madin, Shawn Bowers, Mark P. Schildhauer, and Matt Jones

August 25, 2008

Contents

1 Introduction

The **E**xtensible **OB**servation **O**ntology (reversed to give OBOE) was developed to describe scientific observations and measurements with particular emphasis on capturing observational context. This emphasis is important because automated reasoning tasks, such as searching for data or merging data sets, require "knowledge" of contextual structure in order to determine if data from different sources are compatible (e.g., equivalent or subsumable). For example, if a human or machine process can determine that two data points have the same temporal and spatial context, then the points are potentially comparable, and might be merged to increase the power of a scientific analysis. This tutorial will focus on the use of OBOE for annotating ecological and environmental data sets. A more detailed description of the of OBOE's ontological structure can be found elsewhere (Madin *et al.*, 2007).

1.1 Observation and Measurement

Observations and Measurements form the core of OBOE; represented as black ellipses in Figure 1. An Observation is defined as an assertion that a *thing* (physical, abstract, or conceptual) was observed to belong to a specified class of Entity in a specified ontology. For example, a field researcher might assert that an entity they observed belongs in the class Tree in the ontology portrayed in Figure 2 (i.e., the observed thing is an *instance* of class Tree which is a subclass of Entity).

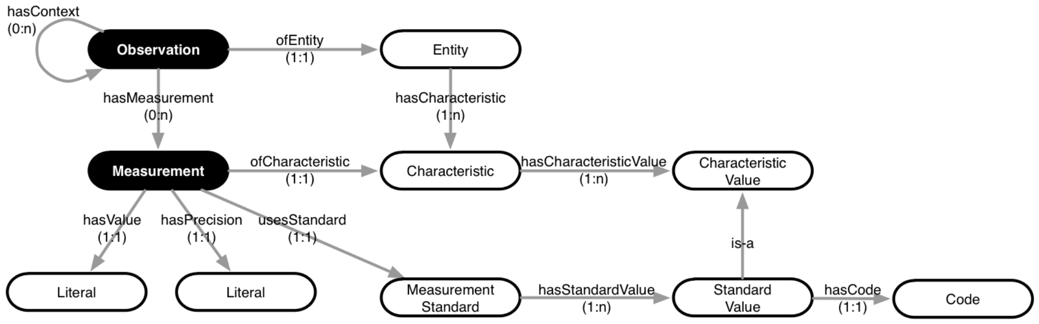


Figure 1: A simplified representation of OBOE's core concepts. Ellipses signify ontology classes and arrows signify properties. Maximum and minimum cardinalities are given for each property and thus indicating the necessary number of connections required for a given data instance.

A Measurement is the subsequent documentation of a characteristic of the observed entity as data (or sometimes metadata). For example, a data set may contain a set of tree heights (Table 1), where the characteristic Height was measured for a number of instances of the class *Quercus lobata*. A Measurement is composed from four parts (Figure 1). The first two parts are the Characteristic that was measured (e.g., weight, color, or name) and the Measurement Standard used to make the measurement (e.g., physical unit or classification scheme), both of which are selected from their respective ontology extensions of OBOE; similar to observed Entity. The other two parts are Value (the actual datum in the data set) and an estimate of precision, both of which are literals (i.e., not represented in an ontology).

Generally, the observed Entity is chosen to be the lowest common denominator for the focal set of data; that is, in Table 1 all the height measurement data apply to independent instances of *Quercus lobata* and nothing else (e.g., not Rock, Bush, nor *Quercus durata*). However, in Table 2 the observed Entity is coarser than for Table 1, and the lowest common denominator is now Tree. In this second example, the classification

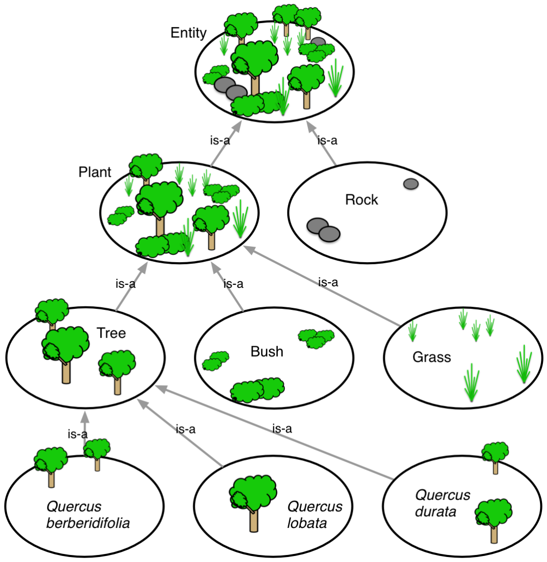


Figure 2: A simple ontology of observable entities illustrating a continuum from coarser to finer level classification. The property "is-a" applies to class instances (not the classes themselves). As an example, this ontology states that an instance of the class Tree is also an instance of the class Plant, but not an instance of the class Rock.

Height (m)
1.2
9.5
4.8
1.5
2.1
...

Table 1: A data set containing tree heights for the oak *Quercus lobata*.

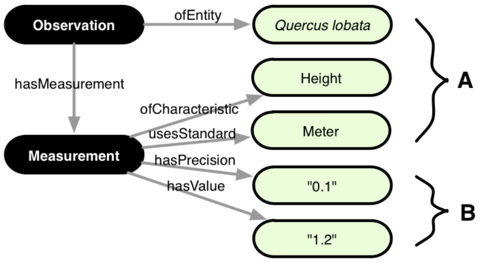


Figure 3: The OBOE representation of the first data point in Table 1. A. Entities, Characteristics, and Standards are represented as instances of ontology classes. B. Measurement Precisions and Values are represented using codes or literals. The other data points in Table 1 are represented identically except for Value, and therefore the hasValue property is usually omitted in OBOE representations.

of instances of Tree into “Species” is now a Measurement, and the Height of each instance is also measured within the same instance of Tree that was observed (Figure 4).

Table 2: A data set containing heights for multiple oak species.

Species	Height (m)
Q. lobata	15.5
Q. berberidifolia	9.0
Q. berberidifolia	8.8
Q. lobata	16.5
Q. durata	5.1
...	...

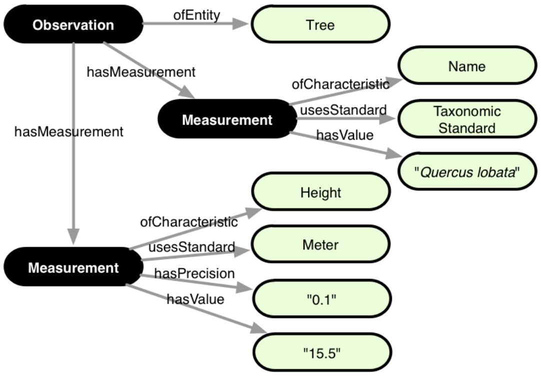


Figure 4: The OBOE representation of the first data point in Table 2.

These examples illustrate several key aspects of OBOE:

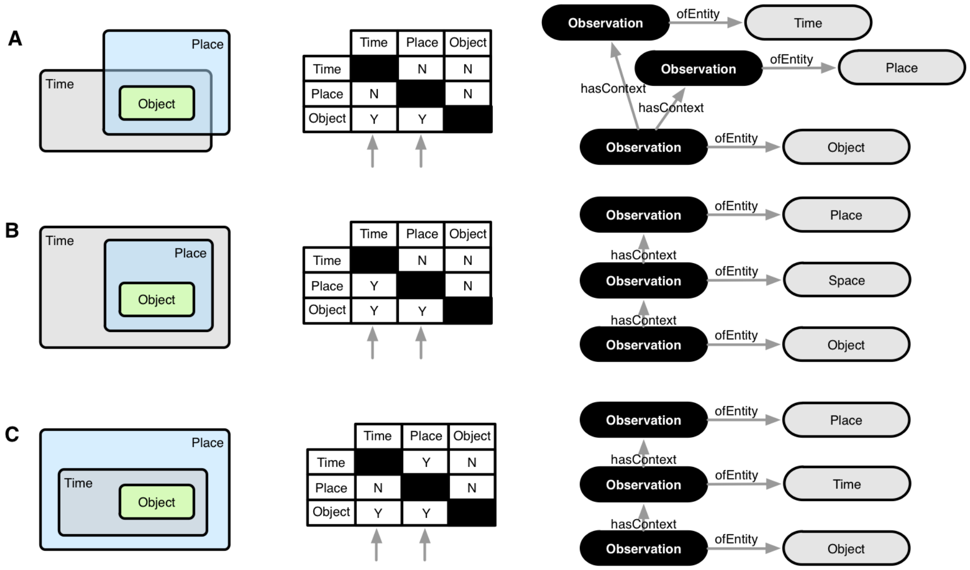
- An Observation implies a *definite* assertion according to a given ontology (or world view). For example, the Heights measured in Table 1 apply to instances of a category of things called *Quercus lobata* according to the world view shown in Figure 2. But world views differ between research groups and change through time (e.g., one group might clump instances of the class *Quercus lobata* with *Quercus durata*). OBOE does not attempt to capture or track mappings among different world views.

- A Measurement is, in essence, the further classification of an entity based its characteristics (e.g., an instance of the class “7.45 Meter High Tree”) but implies an inherent degree of *uncertainty*. An observed entity can be measured (or classified) more than once (e.g., height, width, and name). However, each measurement must apply to the *whole* entity. For example, it is not correct to observe a Lion and measure the length of its tail within the same observation. In this case, the observation of Lion provides *context* for a second observation of Tail (discussed in the next section). Measurements can be made at each level of observation; e.g., the length of the Lion and the length of the Tail (which is associated to the particular instance of Lion via context).

1.2 Context: The Backbone of Observation

Context forms the critical backbone of observational data. That is, Observations and their corresponding Measurements can only be understood within the context they are made. For example, there are few interesting questions that can be asked about a set of tree heights without any contextual information (e.g., the data in Table 1). Whereas, if there is also observational information about location where the trees were observed, then these data are potentially comparable with measurements taken at different locations. Correctly representing the contextual structure of a data set use OBOE takes practice, but if done correctly provides a powerful mechanism for automatically determining the compatibility and mergability of data.

A data set (including metadata) will typically contain information about a object or phenomena observed, as well as the time and place it was observed. However, several contextual structures can potentially describe the relationships among the three entities—e.g., Time, Place, and Object (Figure 5). At this point, it is important to understand that observing what we might be considered the "same" Entity at a different time or place typically results in a different instance of Entity. For example, trees change through time (i.e., they are perdurants). In this sense, observing perdurant Entities at different times can be treated similarly to observing them at different places: they do not represent the same instance of Entity.



	Time	Place	Object
Time	■	N	N
Place	N	■	N
Object	Y	Y	■

	Time	Place	Object
Time	■	N	N
Place	Y	■	N
Object	Y	Y	■

	Time	Place	Object
Time	■	Y	N
Place	N	■	N
Object	Y	Y	■

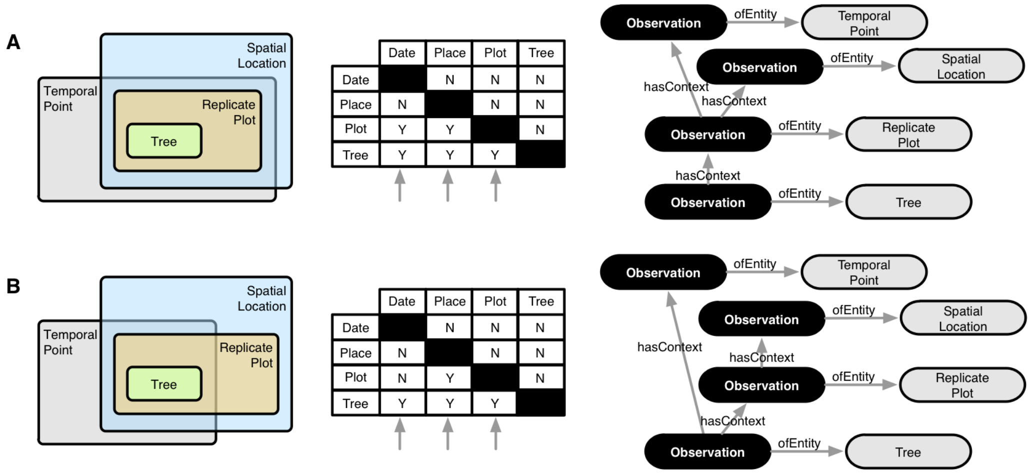
Figure 5: Examples of Entity instance dependence in space and time. Panel **A** illustrates the case when place and time provide independent context for Tree instance observations; that is, each combination of place and time contains a different (unique) instance of Tree. Panel **B** shows the case when space is dependent of time. This might happen if space changes through time, such as when space is sampled randomly at different time steps. Finally, panel **C** shows the case when time is dependent of place. This might happen when time-series are taken at different places or sampling times are selected randomly within a space.

However, this is not always the case, and Space and Time are good examples of this. Figure 5 shows three possible contextual structures for Time, Place, and Object. Figure 5A represents the contextual structure when Time and Place provide independent context for Object. This means that a given instance of Place can be observed within different instances of Time (and visa versa), however a given instance of Object cannot be observed within different instances of Time nor Place. The dependency matrix for Figure 5A illustrates that a different instance of Time does not necessarily imply a different instance of Place (denoted "N" for No). However, a different Time does imply a different Object (denoted "Y" for Yes). Furthermore, a different instance of Place does not imply a different instance of Time, but it does imply a different Object. Finally, a different instance of Object does not necessarily imply a different Time nor Place, because in this example it is possible to observed more than one object at a specific Time and Place combination. Following the grey arrows determines the correct OBOE context representation in OBOE (to the right of the matrix). However, Time and Place are not always independent. Figure 5B shows an example for when Places are different (e.g., randomly chosen) at each for the times observations were made; i.e., Place is dependent on Time. Figure 5C shows an example where Time is now dependent on Place, which might happen when time-series are conducted at different places.

Table 3: A data set containing heights for multiple oak species within plots at Summer Hill on two occassions.

Date	Location	Plot	Species	Height (m)
12/04/2007	Summer Hill	A	*Q. lobata*	2.5
12/04/2007	Summer Hill	A	*Q. lobata*	3.1
12/04/2007	Summer Hill	B	*Q. berberidifolia*	4.1
12/04/2007	Summer Hill	B	*Q. lobata*	8.5
12/04/2007	Summer Hill	C	*Q. durata*	2.2
12/06/2008	Summer Hill	A	*Q. lobata*	7.8
12/06/2008	Summer Hill	A	*Q. lobata*	7.7
12/06/2008	Summer Hill	A	*Q. lobata*	2.1
12/06/2008	Summer Hill	A	*Q. lobata*	2.6
12/06/2008	Summer Hill	A	*Q. lobata*	3.2
12/06/2008	Summer Hill	B	*Q. berberidifolia*	9.1
12/06/2008	Summer Hill	B	*Q. lobata*	9.8
12/06/2008	Summer Hill	B	*Q. berberidifolia*	2.2
12/06/2008	Summer Hill	C	*Q. durata*	7.5

Table 3 contains another hypothetical data set that was collected to monitor the growth of oak trees at a specified location through time. Data collection was accomplished by measuring the names and heights of individual trees within 10 x 10 meter square plots at two different times. In this table, each row (or tuple) is comprised from four Observations: one each of Temporal Point, Spatial Location, Replicate Plot (a second spacial treatment), and Tree (selected from the OBOE Entity extension shown in Figure 7). Figure 6 illustrates two possible ways in which the contextual structure might be represented depending on the research protocol. The key difference between the two representations is whether or not the plots are (A) permanent (i.e., the same instance of Replicate Plot can exist at different instances of Temporal Point) or (B) randomly placed during each census (i.e., an instance of Plot is dependent on a specific instance of Temporal Point).



	Date	Place	Plot	Tree
Date	■	N	N	N
Place	N	■	N	N
Plot	Y	Y	■	N
Tree	Y	Y	Y	■

	Date	Place	Plot	Tree
Date	■	N	N	N
Place	N	■	N	N
Plot	N	Y	■	N
Tree	Y	Y	Y	■

Figure 6: Venn diagrams illustrating two possible contextual structures for the Entities observed in Table 3. Panel **A** shows the contextual structure for when Plots are randomly placed at each consecutive time step; so an instance of Plot is dependent of an instance of Time. Alternatively, panel **B** shows the structure for when Plots are permanently placed and thus “endure” through time; so an instance of Plot is not wholly dependent on an instance of Time.

1.3 Semantic Annotation

Semantic annotation is the mapping between data and OBOE. Outlined below are the steps necessary to annotate data sets using Table 3 as the working example. Computer software is currently being develop to assist the annotation process.

1. Identify the data attributes (in this case, columns) to be capture by the semantic annotation (top row in Figure 8). Meta data that applies to whole-attributes should also be identified. In this example, all replicate plots were 10 meters square, and so there was no need to include a column expressing this in the raw data set (and thus it was relegated to meta data). However, such information is critical for determining important measures like areal density, especially when comparing and potentially merging data sets.

2. The appropriate Entities for each data attribute are then selected from a specified extension ontology (second row in Figure 8; selected from the Entity extension in Figure 7), remembering that sometimes more than one attribute pertains to the same Entity. In this example, Species and Height both pertain to a instance of Tree.

3. Choose the Characteristics from the extension ontology that best describes that the attribute measured, as well as the Measurement Standard used to represent the Characteristic (third and fourth rows of Figure 8; both selected from the respective extensions in Figure 7).

4. Value points to the set of data for which the annotation applies. In this case, the annotation is applied to each data point in the column vectors (bottom row of Figure 8). Precision only applies to continuous variables, and is by default the number of significant decimal places (fifth row of Figure 8).

5. Determine the contextual structure. In this data set, the plots are permanent and therefore the representation in Figure 6B is correct.

6. Complete the OBOE annotation graphically using the information in Figure 8 and the context structure in Figure 6B to give Figure 9).

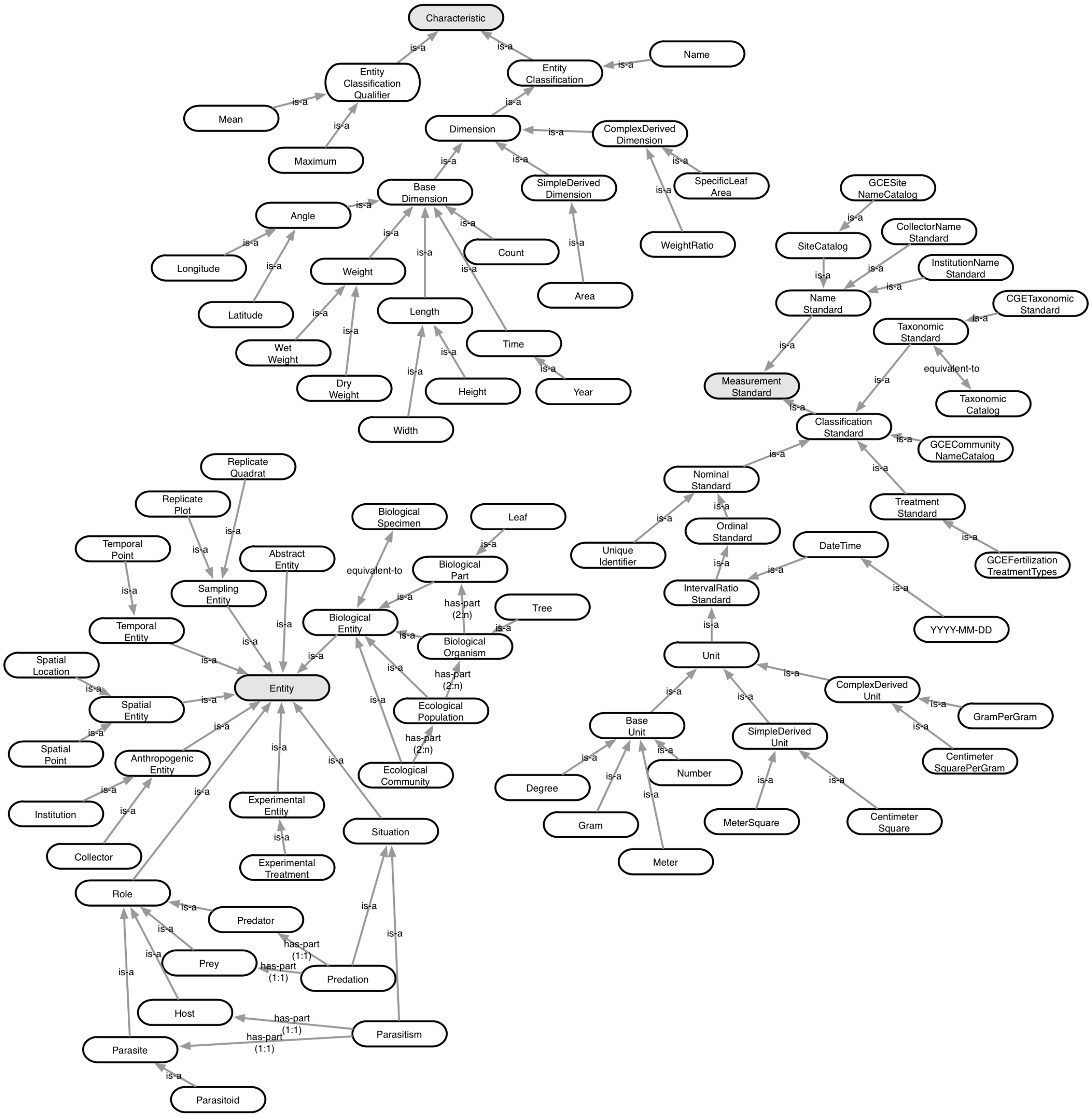
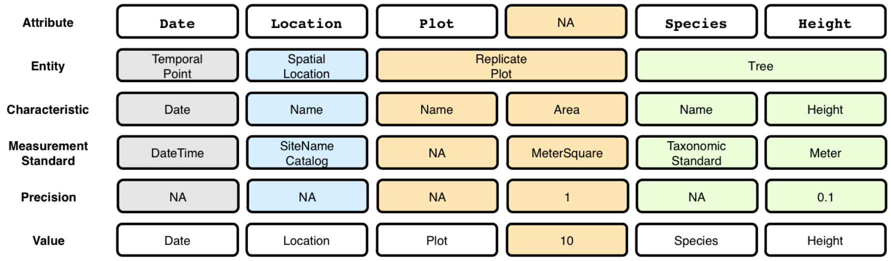


Figure 7: Ontology extensions of core OBOE classes used in this tutorial.



Attribute	Date	Location	Plot	NA	Species	Height
Entity	Temporal Point	Spatial Location	Replicate Plot		Tree	
Characteristic	Date	Name	Name	Area	Name	Height
Measurement Standard	DateTime	SiteName Catalog	NA	MeterSquare	Taxonomic Standard	Meter
Precision	NA	NA	NA	1	NA	0.1
Value	Date	Location	Plot	10	Species	Height

Figure 8: A...

7. Represent the annotation using the Semantic Annotation Language (SAL) that can then be saved in the data sets metadata. The SAL is described elsewhere and will be done automatically when the annotation software is ready.

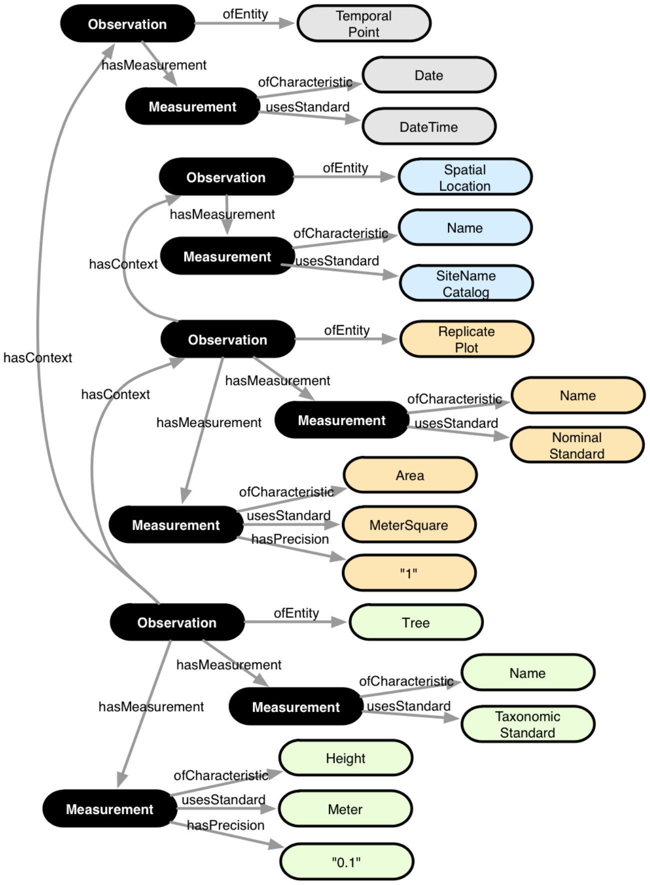


Figure 9: The OBOE representation of the first data point in Table 3.

ANNOTATION 1—

```
<annotation emlPackage="..." dataTable="..." xmlns:oboe="..." xmlns:ext="...">

  <observation label="o1" attributes="Date">
    <entity class="ext:TemporalPoint"/>
    <measurement>
      <characteristic class="oboe:Date"/>
      <standard class="ext:DateTime"/>
      <value attribute="Date"/>
    </measurement>
  </observation>

  <observation label="o2" attributes="Location">
    <entity class="ext:SpatialLocation"/>
    <measurement>
      <characteristic class="oboe:EntityName"/>
      <standard class="ext:SiteNameCatalog"/>
      <value attribute="Location"/>
    </measurement>
  </observation>

  <observation label="o3" attributes="Date␣Location␣ReplicatePlot">
    <entity class="ext:ReplicatePlot"/>
    <measurement>
      <characteristic class="oboe:EntityName"/>
      <standard class="oboe:NominalStandard"/>
      <value attribute="Plot"/>
    </measurement>
    <measurement precision="1">
      <characteristic class="ext:Area"/>
      <standard class="ext:MeterSquare"/>
      <value constant="10"/>
    </measurement>
    <context observation="o1" property="oboe:hasContainmentContext"/>
    <context observation="o2" property="oboe:hasContainmentContext"/>
  </observation>

  <observation label="o4" attributes="*">
    <entity class="ext:Tree"/>
    <measurement>
      <characteristic class="oboe:EntityName"/>
      <standard class="ext:TaxonomicStandard"/>
      <value attribute="Species"/>
    </measurement>
    <measurement precision="0.1">
      <characteristic class="ext:Height"/>
      <standard class="ext:Meter"/>
      <value attribute="Height"/>
    </measurement>
    <context observation="o3" property="oboe:hasContainmentContext"/>
  </observation>

</annotation>
```

2 Examples of Semantic Annotation

In this section, a number of real data sets are annotated using OBOE starting with basic examples and getting progressively harder.

2.1 Height Measurement

2.2 Tree Measurements I

2.3 Tree Measurements II

2.4 Leaf Traits

2.5 GCE Fertilisation Plots

2.6 Specimen Record

2.7 Summary of Organism Heights from Example 3

2.8 Predator-Prey Study

2.9 Parasitoid–Parasite–Host Study