A Conceptual Modeling Framework for Expressing Observational Data Semantics*

Shawn Bowers[1], Joshua S. Madin[2], and Mark P. Schildhauer[3]

[1]Genome Center, University of California, Davis
[2]Dept. of Biological Sciences, Macquarie University, Australia
[3]National Center for Ecological Analysis and Synthesis, UC Santa Barbara
sbowers@ucdavis.edu, {madin,schild}@nceas.ucsb.edu

Abstract. Observational data (i.e., data that records observations and measurements) plays a key role in many scientific disciplines. Observational data, however, are typically structured and described in *ad hoc* ways, making its discovery and integration difficult. The wide range of data collected, the variety of ways the data are used, and the needs of existing analysis applications make it impractical to define "one-size-fits-all" schemas for most observational data sets. Instead, new approaches are needed to flexibly describe observational data for effective discovery and integration. In this paper, we present a generic conceptual-modeling framework for capturing the semantics of observational data. The framework extends standard conceptual modeling approaches with new constructs for describing observations and measurements. Key to the framework is the ability to describe observation context, including complex, nested context relationships. We describe our proposed modeling framework, focusing on context and its use in expressing observational data semantics.

1 Introduction

Scientific knowledge is typically derived from relatively simple, underlying measurements directly linked to real-world phenomena. Such measurements are often recorded and stored in *observational data sets*, which are then analyzed by researchers using a variety of tools and methodologies. Many fields increasingly use observational data from multiple disciplines (genetics, biology, geology, hydrology, sociology, etc.) to tackle broader and more complex scientific questions. Within ecology, e.g., cross-disciplinary data is necessary to investigate complex environmental issues at broad geographic and temporal scales. Carrying out such studies requires the integration and synthesis of observational data from multiple research efforts [1,2]. These investigations, however, are hindered by the heterogeneity of observational data, which impedes the ability of researchers to discover, interpret, and integrate relevant data collected by others.

The heterogeneity of observational data is due to a number of factors: (1) most observational data are collected by individuals, institutions, or scientific communities through independent (i.e., uncoordinated) research projects; (2) the structure of observational

* This work supported in part by NSF grants #0533368, #0553768, #0612326, #0225676, #0630033, and #0612326.

Q. Li et al. (Eds.): ER 2008, LNCS 5231, pp. 41–54, 2008.

data is often chosen based on collection methods (e.g., to make data easier to record "in the field") or the format requirements of analysis tools, as opposed to standard schemas; and (3) the terms and concepts used to label data are not standardized, both within and across scientific disciplines and research groups [3]. This need for a more uniform mechanism to describe observational data has led to a number of proposals for observational data models [4,5,6] and ontologies [7,8,9,10]. While many of these approaches provide domain-specific vocabularies for describing data, or data models for storing certain types of observational data, generic and extensible approaches for modeling observational data semantics are still needed.

We present an initial step towards such a generic conceptual modeling framework for observational data. Our framework extends traditional conceptual modeling languages with constructs for explicitly modeling observations, measurements, and observation context. Our approach aims to address challenges associated with the following general characteristics of observational data:

- Observational data are primarily stored as tables within text files, where each data set corresponds to a single table within one file. This situation stems from data being generated for use in common analytical programs, e.g., spreadsheet tools.
- Observational data sets are represented in first normal form (1NF), but are not otherwise normalized, with no integrity constraints given.
- Observational data are not initially created from explicit conceptual models (e.g., ER or UML diagrams).
- Observational data do not represent a set of facts, or "ontological" truths about the world; instead, they represent (possibly conflicting) measurements of phenomena within some broader context.
- Observational data do not use standardized terms for attribute names and coded values (e.g., species or location names). The terms used, however, may be informally described within plain-text metadata descriptions.

We envision conceptual models being created within our framework to describe *existing* observational data, primarily for the purpose of enabling discovery and integration of data sets. That is, while it may be possible to employ a more traditional "top down" modeling approach using our framework, we are primarily focused on the case of supplementing existing data with formal, semantic content descriptions. We call this process *semantic annotation*, whereby annotations referencing an external conceptual model serve to clarify and constrain the interpretation of the original data set.

As others have noted (e.g., [11,12,13]), it is often difficult to represent observations and their context in traditional conceptual-modeling languages. For example, Fig. 1 shows three hypothetical observational data sets together with their corresponding ER diagrams. Fig. 1(a) shows diameter measurements of trunks of different tree species taken in different years. Fig. 1(b) depicts similar yearly measurements of tree trunk diameters, but where trees are located within plots, plots have average daily temperatures, and plots are located within sites. Fig. 1(c) also consists of tree diameter measurements, where trees are located along a transect (a fixed path within an area) and within a particular type of soil, soil acidity is measured, and each transect has a particular type of treatment applied (either a high or low watering regime). These relatively simple examples demonstrate the need for semantic descriptions to clarify similarities

(a)

tree	*spp*	*yr*	*dbh*
1	piru	2007	35.8
1	piru	2008	36.2
2	abba	2008	33.2
...	...	...	...

(b)

site	*plot*	*tmp*	*yr*	*piru*	*abba*
1	A	53.2	2007	50.1	40.1
1	A	53.2	2008	50.5	40.7
1	B	50.2	2007	41.2	35.2
...	...	...	...	...	...

(c)

trns	*trmt*	*soil*	*pH*	*tree*	*spp*	*dbh*
1	hi	s	3.9	t1	piru	14.1
1	hi	s	5.1	t2	piru	16.2
2	lo	h	6.0	t3	abba	12.5
...	...	...	...	...	...	...

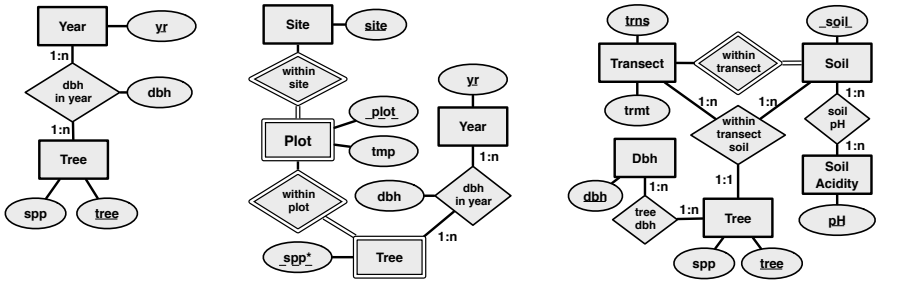

Fig. 1. Three simple observational data sets and example ER representations: (a) diameter-at-breast-height (dbh) measurements per year for tree species; (b) dbh per year for tree species located in plots within sites; and (c) dbh and soil pH (acidity) measurements along transects. The attribute spp* in (b) generalizes the two attributes labeled with species names, *piru* and *abba*. Cardinality restrictions x:y denote the min and max participation of the entity in the corresponding role of the relationship.

and differences among data sets. For example, it is not obvious from the attributes and data values, nor the ER diagrams, whether these three data sets contain similar types of measurements.

While the conceptual models of Fig. 1 help to describe these data sets, they also highlight challenges in expressing observational data semantics that are crucial to the scientific interpretation and potential usage of these data for an integrated analysis:

Implicit context. In each example data set, the same tree entity has different diameter (dbh) values. These discrepancies are explained by the context in which the diameter measurements occur. In general, context describes the meaningful "surroundings" of an observation, such as the other entities observed, their measured values, and their relationship to the observed entity. However, context is only implicitly modeled in Fig. 1: it is unclear which relationships denote context (e.g., "dbh in year", "within plot") and which denote measurements (e.g., "tree dbh", "soil pH"). Similarly, context is only partially specified: it is not explicitly stated that transects and soils are context for trees, or that trees also serve as context for soils. Without an understanding of the contextual relationships within a data set, it becomes difficult to interpret and analyze data. In Fig. 1, e.g., it is not trivial to determine whether it is meaningful to summarize temperatures across years (computing a yearly average), or how to compute average tree diameter by soil type. This in turn has ramifications for data integration, which often requires the aggregation of observations to combine data [10].

Coupled structure and semantics. Although similar, the conceptual models in each of the examples reflect potentially important differences. These differences are primarily due to variation in methodologies used to collect data, and are expressed through

relationships, cardinality constraints, weak-entity constraints, promotion of attributes to entities, etc. While the same general types of entities and relationships exist across the three examples, the difficulty of capturing the methodological constraints (such as context) within models impairs the ability to: (1) define domain-specific concepts and relationships (e.g., within a shared ontology) that can be used to semantically annotate *multiple* datasets; and (2) easily compare the semantics of different data sets for discovery and integration.

Complex constraints. Similarly, constructing conceptual models of observational data using traditional modeling languages often requires the combination of complex constraints in conjunction with "advanced" modeling features (e.g., n-ary relationships, cardinality restrictions). Because of the complexity of observational data, constructing appropriate conceptual models is often tricky, and thus a time-consuming and error-prone task. Similarly, these approaches often require knowledge of esoteric concepts that would not be intuitive to most scientific researchers who ultimately need to understand and use the data.

The rest of this paper is organized as follows. In Section 2 we describe our proposed framework for modeling observational data. Our approach addresses a number of the challenges highlighted above: (1) we introduce explicit constructs for modeling observations and their context, thereby allowing domain-specific concepts and relationships to be decoupled from the constraints imposed by data-collection methods; (2) because of this separation of concerns, the complex constraints needed to represent observational data are reduced; and (3) the framework provides a natural approach for data annotation and summarization. In Section 3 we describe related work, and discuss future directions in Section 4.

2 Modeling Observational Data

The basic constructs of our modeling framework are depicted in Fig. 2. We introduce new constructs (left) for representing measurement standards, measurements, observations, and context. These constructs are layered upon "traditional" ER constructs (right), namely, entities, relationships, attributes (called "characteristics" in Fig. 2), and values.

Measurement standards represent the various criteria used for comparing measured values. Examples of measurement standards include units (e.g., meter, gram, square centimeter), nominal and ordinal codes (e.g., location or color names, gender codes), scales (e.g., pH, Richter scale, drought severity index), and date-time standards. Values are combined with measurement standards to form **Standard Values**. Although not described here, measurement standards are often classified by a standard typology that differentiates nominal, ordinal, interval, and ratio measurements [14].

Measurements consist of a characteristic (i.e., attribute) and a standard value.[1] In our framework, each value within an observational data set represents a measurement. For example, the first value in the *dbh* column of Fig. 1(a) denotes a measurement consisting of a characteristic of type 'diameter-at-breast-height', the unit 'centimeter', and the value '35.8'. Values representing categorical and identifying information are

[1] Measurements may also have additional information, such as *precision* and *accuracy*.

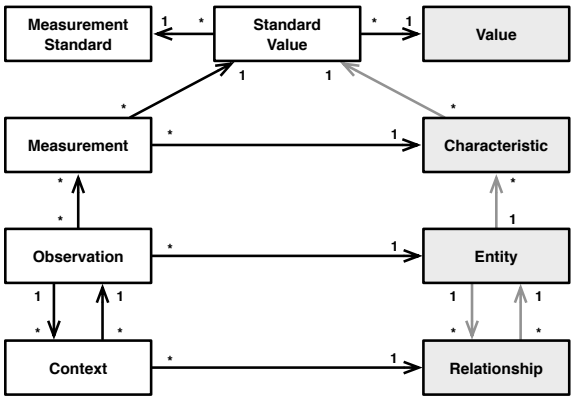

Fig. 2. A metamodel for describing observational data

also considered measurements. For example, in Fig. 1(c), the *trns* column denotes measurements having characteristics of type 'name' according to a local transect naming scheme, and similarly, the *trmt* column represents measurements having characteristics of type 'water-level' and a measurement standard that defines the values 'hi' and 'lo'.

Observations consist of an entity (denoting the entity observed) and a set of measurements. Each measurement associated with an observation applies to the observed entity. That is, an observation asserts through a measurement that a particular value was observed for one of the characteristics of the entity (implicitly shown by the gray arrows on the right of Fig. 2). In addition, an observation can be related to zero or more observations through context. A **Context** consists of a relationship and an observation, and states that an observation was made within the scope of another observation. A contextual relationship between observations asserts that the relationship was observed between the corresponding entities.

As shown in Fig. 2, binary directed relationships are used within the framework for modeling observational data. Binary, as opposed to more general n-ary relationships, are employed for two primary reasons: (1) they allow for ontology languages based on description logics (e.g., OWL-DL[2]) to be easily used within our framework to define domain-specific vocabularies for data annotation (e.g., where entities are expressed via OWL-DL classes, relationships through object properties, and characteristics through datatype properties); and (2) they generally result in models that are simpler and easier to define (although less restrictive). We also use the term 'characteristic' instead of 'attribute' to distinguish the semantic property being described from the particular column label used within a data set. In particular, the process of annotation involves associating the attributes of a given data set with specific characteristics defined in domain-specific vocabularies (described further below).

Fig. 3 shows three observational models for describing the data sets of Fig. 1. Instead of defining entity, relationship, and attribute types (as in Fig. 1), the diagrams of Fig. 3 define the observation, measurement, and context types for the data sets.

[2] `http://www.w3.org/TR/owl-ref/`

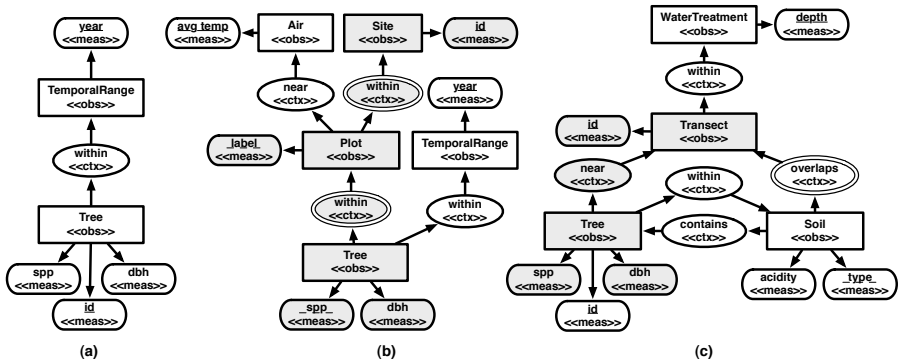

Fig. 3. Example observational conceptual models for the data sets of Fig. 1. Rectangles denote observations labeled with the corresponding entity, rounded boxes represent measurements labeled with the corresponding characteristic, and ovals represent context labeled with the corresponding relationship. To simplify the diagrams, measurement standards are not shown. Closely related concepts of (b) and (c) are highlighted.

These types *reference* the appropriate entity, relationship, and characteristic types defined, e.g., within one or more shared ontologies. As shown, measurements and context relationships can be used to denote distinct entities (via keys, weak entities, and identifying relationships), where the same entity may be involved in multiple observations.

The examples of Fig. 3 demonstrate many of the advantages of our framework for describing observational data. Because observational structures (observations, measurements, context) and semantic structures (entities, relationships, characteristics) are decoupled, the latter can be used uniformly across observational models (e.g., dbh or the 'within' relationship). Similarly, creating models for observational data can be driven by the definitions within standard ontologies, thereby simplifying the annotation process, and lowering the potential for terminological ambiguity. As an example, characteristic and relationship types can be defined within an external ontology to be used only with specific types of entities, thus suggesting entity types for characteristics, and vice versa.

Another advantage is that context is explicitly represented and distinguished from measurements. In contrast to Fig. 1, context relationships are directed, allowing one to easily determine the context hierarchies (or "paths") induced by observations. Similar to summarizability in multidimensional databases [15], context hierarchies can help determine the meaningful summarizations available within a data set. Context relationships also encourage the full disclosure of *what* was observed, which is critical metadata that is often left implicit in observational data. This is demonstrated in Fig. 3(c), where an explicit observation type is used to denote water entities used as experimental treatments, in which the corresponding depth measurement denotes the height of the water level. Similarly, in Fig. 3(b), an explicit air observation type is used to signify that air temperature was measured (as opposed to water or body temperature, e.g.).

Below we further describe the framework of Fig. 2. We first give a formal definition of the modeling constructs, focusing on instance-level descriptions. We then describe types, which are used to construct observational models (e.g., as in Fig. 3). We also show how existing data sets can be annotated with conceptual models defined in our framework, and finally discuss issues related to summarization.

2.1 Observation Instances

An instance of a model is constructed from the following base and derived sets. *Val* is the set of measurement values (integers, doubles, strings, etc.). *Std* is the set of measurement standards (units, scales, etc.). *Ent* is the set of entity objects. *ObsId* is the set of observation identifiers. *Rel* is the set of identifiers denoting binary, directed relationships between entities. And *Char* is the set of identifiers denoting entity characteristics that relate specific entities to standard values. The derived structures are built from these base sets as follows.

$$\begin{aligned} \mathit{StdVal} &\subseteq \mathit{Val} \times \mathit{Std} \\ \mathit{EntRel} &\subseteq \mathit{Ent} \times \mathit{Rel} \times \mathit{Ent} \\ \mathit{EntChar} &\subseteq \mathit{Ent} \times \mathit{Char} \times \mathit{StdVal} \\ \mathit{Obs} &\subseteq \mathit{ObsId} \times \mathit{Ent} \times \mathcal{P}(\mathit{Meas}) \times \mathcal{P}(\mathit{Ctx}) \\ \mathit{Meas} &\subseteq \mathit{Char} \times \mathit{StdVal} \\ \mathit{Ctx} &\subseteq \mathit{Rel} \times \mathit{ObsId} \end{aligned}$$

A standard value consists of a measurement standard and a value, e.g., $\mathit{StdVal}(5, \mathrm{cm})$ denotes the quantity "5 centimeters" (where cm represents the unit centimeter). The elements of *Rel* and *Char* act as "handles" to specific relationship and characteristic occurrences such that *EntRel* and *EntChar* specify the relationships and characteristics, respectively. If $\mathit{EntRel}(e_1, r, e_2)$, we say e_1 is r-related to e_2, and that r goes from e_1 to e_2. Similarly, if $\mathit{EntChar}(e, c, v)$, we say that e has the standard value v for characteristic c. Entities may have at most one value for a characteristic. Each observation has an explicit identifier and consists of an entity, a set of measurements, and a set of contexts. For convenience, we often write $o = \mathit{Obs}(e, M, C)$ to denote an observation $\mathit{Obs}(o, e, M, C)$. A measurement consists of a characteristic and a standard value. And a context consists of a relationship and a reference to an observation.

Example 1 (Observation instance). A portion of the instance of the observational model of Fig. 3(a) corresponding to the first row of the data set in Fig. 1(a) is given below, where c_1 to c_4 are characteristics of type `Year`, `Dbh` (diameter at breast height), `Spp` (taxonomic name), and `Id`, respectively, and r_1 is a relationship of type `Within`.

$$\begin{aligned} o_1 &= \mathit{Obs}(e_1, \{m_1\}, \emptyset) \\ m_1 &= \mathit{Meas}(\mathrm{c}_1, \mathit{StdVal}(2007, \text{datetime})) \\ o_2 &= \mathit{Obs}(e_2, \{m_2, m_3, m_4\}, \{\mathit{Ctx}(\mathrm{r}_1, o_1)\}) \\ m_2 &= \mathit{Meas}(\mathrm{c}_2, \mathit{StdVal}(35.8, \mathrm{cm})) \\ m_3 &= \mathit{Meas}(\mathrm{c}_3, \mathit{StdVal}(\text{Picea rubens}, \text{ITIS})) \\ m_4 &= \mathit{Meas}(\mathrm{c}_4, \mathit{StdVal}(1, \text{LocalTreeIds})) \end{aligned}$$

Here, e_1 and e_2 denote entities of type `TemporalRange` and `Tree`, respectively; ITIS represents a taxonomic name standard[3]; and LocalTreeIds represents a catalog of tree ids local to the study.

As mentioned above, observations represent assertions about entities. In particular, measurements imply that within a given context, an entity was observed to have the corresponding measured characteristic values. Similarly, observations inherit the assertions of their contextual observations. The assertions of an observation are obtained by "entering" the observation, given by the operation *enter*(o)[4], for an observation o. Let

$$context : ObsId \rightarrow \mathcal{P}(ObsId)$$

be a function that takes an observation and returns its corresponding contextual observations. For an observation $o = Obs(e, M, C)$, we define

$$context(o) = \{o' \mid \exists r\ Ctx(r, o') \in C\},$$

where $context^+$ denotes the transitive closure of *context*. For $context^+(o) = O$, we define $enter(o) = E_m \cup E_r \cup E_c$ such that:

$$\begin{aligned} E_m &= \{EntChar(e, c, v) \mid Meas(c, v) \in M\} \\ E_r &= \{EntRel(e, r, e') \mid \exists o' M' C'\ (Ctx(r, o') \in C) \wedge (o' = Obs(e', M', C'))\} \\ E_c &= \bigcup_{o' \in O} enter(o') \end{aligned}$$

For example, Fig. 4 shows the result of entering two different observations corresponding to the first two rows of data in Fig. 1(b). Entering a tree observation (i.e., for the *piru* attribute; denoted by o5 in the figure) results in assertions "up" the context hierarchy of Fig. 3(b), and includes the corresponding temporal, plot, air, and site observations. Entering a plot observation (denoted by o3 in the figure), however, results only in corresponding plot, air, and site assertions.

By providing a semantics for observation context, the *enter* operation can also help verify the consistency of conceptual models and their instances. In particular, for an observation to be consistent, the result of entering the observation must be consistent. The latter is determined by the constraints implied by the corresponding semantic constructs (entities, relationships, characteristics). For example, because entities have at most one value for a characteristic of a given type (such as dbh in Fig. 3), the result of entering an observation must also satisfy this constraint. Adding a new observation o8 to Fig. 4 with observation context o5 and o7 would violate this constraint, e.g., since the union of *enter*(o5) and *enter*(o7) is inconsistent. Similarly, cardinality constraints on relationships must be satisfied after entering an observation.

2.2 Observation Types and Models

As mentioned above, decoupling observational and semantic structures allows semantic types (i.e., the entity, relationship, and characteristic types) to be defined independently

[3] Integrated Taxonomic Information System, `http://www.itis.gov`

[4] The notion of entering an observation is similar in spirit to "lifting" operators in [16].

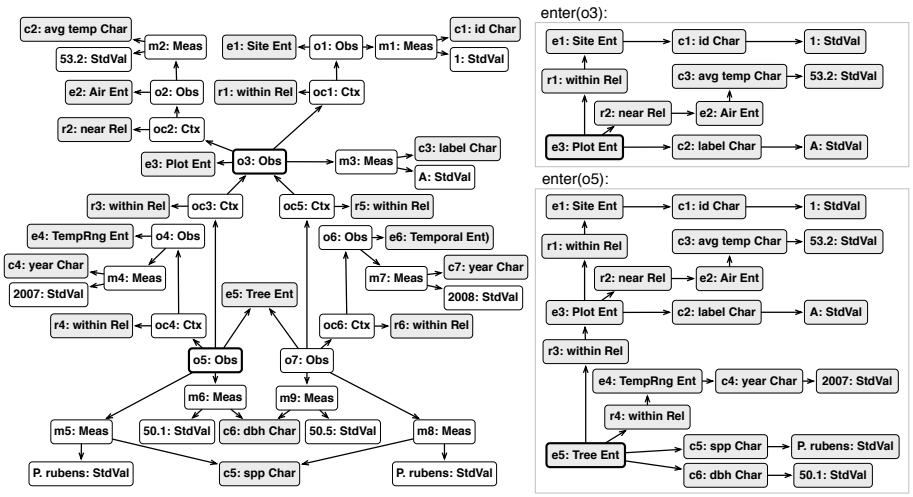

Fig. 4. Example instances of the first two rows of Fig. 1(b), for the *Picea rubens* (*piru*) species (left), and the corresponding result of entering observations o3 and o5 (right). For simplicity, measurement standards are not shown for standard values.

of observations and measurements. Thus, the same semantic types can be used to describe multiple observational data sets. In prior work [3,10], we have used OWL-DL for defining entity, relationship, and characteristic types. However, the framework is not limited to a specific approach (such as OWL), and instead can support the use of a number of different conceptual modeling languages. Here we consider a simple typing language for the purpose of defining observational models such as those of Fig. 3.

We define base types `Std`, `Val`, `StdVal`, `Meas`, `Obs`, `ObsId`, `Ctx`, `Char`, `Ent`, and `Rel` for constructing observational models. We require subtypes τ' of types τ, written $\tau' \sqsubseteq \tau$, to imply subset relations. That is, $\tau' \sqsubseteq \tau$ iff $[\![\tau']\!] \subseteq [\![\tau]\!]$, where $[\![\tau]\!]$ denotes the set of instances of a type τ. If $\tau' \sqsubseteq \tau$ we say that τ' *is-a* τ. Similarly, if x is an instance of a type τ, we write $x : \tau$ such that $x : \tau$ iff $x \in [\![\tau]\!]$. Each base type denotes its corresponding instance-level set, e.g., $[\![\texttt{Val}]\!] \equiv \mathit{Val}$, $[\![\texttt{Std}]\!] \equiv \mathit{Std}$, etc. With slight abuse of notation, we define the standard-value, observation, measurement, and context base types as follows.

$$\begin{aligned}
\texttt{StdVal} &\sqsubseteq \texttt{Val} \times \texttt{Std} \\
\texttt{Obs} &\sqsubseteq \texttt{ObsId} \times \texttt{Ent} \times \mathcal{P}(\texttt{Meas}) \times \mathcal{P}(\texttt{Ctx}) \\
\texttt{Meas} &\sqsubseteq \texttt{Char} \times \texttt{StdVal} \\
\texttt{Ctx} &\sqsubseteq \texttt{Rel} \times \texttt{ObsId}
\end{aligned}$$

These types correspond to the instance-level sets defined above as follows.

$$\begin{aligned}
[\![\texttt{StdVal}]\!] &\subseteq [\![\texttt{Val}]\!] \times [\![\texttt{Std}]\!] \\
[\![\texttt{Obs}]\!] &\subseteq [\![\texttt{ObsId}]\!] \times [\![\texttt{Ent}]\!] \times \mathcal{P}([\![\texttt{Meas}]\!]) \times \mathcal{P}([\![\texttt{Ctx}]\!]) \\
[\![\texttt{Meas}]\!] &\subseteq [\![\texttt{Char}]\!] \times [\![\texttt{StdVal}]\!] \\
[\![\texttt{Ctx}]\!] &\subseteq [\![\texttt{Rel}]\!] \times [\![\texttt{ObsId}]\!]
\end{aligned}$$

Similar to observation instances, for convenience we write $\mathtt{O} = \mathtt{Obs}(\mathtt{E}, \{\mathtt{M}_1, ...\}, \{\mathtt{C}_1, ...\})$ to denote the type $\mathtt{Obs}(\mathtt{O}, \mathtt{E}, \{\mathtt{M}_1, ...\}, \{\mathtt{C}_1, ...\})$. In general, a type definition $\mathtt{Obs}(\mathtt{O}, \mathtt{E}, \{\mathtt{M}_1, \mathtt{M}_2, ...\}, \{\mathtt{C}_1, \mathtt{C}_2, ...\})$ implies a type $\tau \sqsubseteq \mathtt{Obs}$ such that:

$$\tau \sqsubseteq \mathtt{O} \times \mathtt{E} \times \mathcal{P}(\mathtt{M}_1 \cup \mathtt{M}_2 \cup \dots) \times \mathcal{P}(\mathtt{C}_1 \cup \mathtt{C}_2 \cup \dots).$$

This definition similarly states that:

$$[\![\tau]\!] \subseteq [\![\mathtt{Obs}]\!] \cap ([\![\mathtt{O}]\!] \times [\![\mathtt{E}]\!] \times \mathcal{P}([\![\mathtt{M}_1]\!] \cup [\![\mathtt{M}_2]\!] \cup \dots) \times \mathcal{P}([\![\mathtt{C}_1]\!] \cup [\![\mathtt{C}_2]\!] \cup \dots)).$$

Thus, using these definitions it is straightforward to test whether an instance x is of an observational type τ, or whether $\tau \sqsubseteq \tau'$ for two observational types.

Example 2 (Observation types). Let $\mathtt{DateTime}, \mathtt{ITIS}, \mathtt{Cm}, \mathtt{LocalTreeIds} \sqsubseteq \mathtt{Std}$; $\mathtt{Int}, \mathtt{String}, \mathtt{Double} \sqsubseteq \mathtt{Val}$; $\mathtt{Year}, \mathtt{Spp}, \mathtt{Id}, \mathtt{Dbh} \sqsubseteq \mathtt{Char}$; $\mathtt{TemporalRange}, \mathtt{Tree} \sqsubseteq \mathtt{Ent}$; and $\mathtt{Within} \sqsubseteq \mathtt{Rel}$. The observation types of the conceptual model of Fig. 3(a) can be expressed as follows.

$$\begin{aligned}
\mathtt{DateTimeVal} &= \mathtt{StdVal}(\mathtt{Int}, \mathtt{DateTime}) \\
\mathtt{YearMeas} &= \mathtt{Meas}(\mathtt{Year}, \mathtt{DateTimeVal}) \\
\mathtt{TemporalRangeObs} &= \mathtt{Obs}(\mathtt{TemporalRange}, \{\mathtt{YearMeas}\}, \bot) \\
\mathtt{SppMeas} &= \mathtt{Meas}(\mathtt{Spp}, \mathtt{StdVal}(\mathtt{String}, \mathtt{ITIS})) \\
\mathtt{IdMeas} &= \mathtt{Meas}(\mathtt{Id}, \mathtt{StdVal}(\mathtt{Int}, \mathtt{LocalTreeIds})) \\
\mathtt{DbhMeas} &= \mathtt{Meas}(\mathtt{Dbh}, \mathtt{StdVal}(\mathtt{Double}, \mathtt{Cm})) \\
\mathtt{TreeObs} &= \mathtt{Obs}(\mathtt{Tree}, \{\mathtt{SppMeas}, \mathtt{IdMeas}, \mathtt{DbhMeas}\}, \{\mathtt{WithinCtx}\}) \\
\mathtt{WithinCtx} &= \mathtt{Ctx}(\mathtt{Within}, \mathtt{TemporalRangeObs})
\end{aligned}$$

An observational **model** $M = (O, K, W)$ consists of a set of observation types O, a set of key constraints $K \subseteq O \times \mathcal{P}(C)$, and a set of weak-entity constraints $W \subseteq O \times \mathcal{P}(C) \times \mathcal{P}(R \times O)$, where C and R denote the set of characteristic and relationship types, respectively. For every $(\mathtt{O}, \{\mathtt{C}_1, \dots, \mathtt{C}_n\}) \in K$, we require observation type $\mathtt{O}$ to have a measurement type $\mathtt{M}_i$ with characteristic type $\mathtt{C}_i$, for $1 \leq i \leq n$. We similarly constrain $(\mathtt{O}, \{\mathtt{C}_1, \dots, \mathtt{C}_n\}, \{(\mathtt{R}_1, \mathtt{O}_1), \dots, (\mathtt{R}_m, \mathtt{O}_m)\}) \in W$, adding the additional constraint that observation type $\mathtt{O}$ also consist of a context type having relationship type $\mathtt{R}_j$ and observation type $\mathtt{O}_j$ for $1 \leq j \leq m$

Example 3 (Observational model). Assume we have the following type definitions for the model of Fig. 3(b):

$$\begin{aligned}
\mathtt{TreeObs} &= \mathtt{Obs}(\mathtt{Tree}, \{\mathtt{SppMeas}, \mathtt{DbhMeas}\}, \{\mathtt{PlotCtx}, \mathtt{YearCtx}\}) \\
\mathtt{PlotObs} &= \mathtt{Obs}(\mathtt{Plot}, \{\mathtt{LabelMeas}\}, \{\mathtt{AirCtx}, \mathtt{SiteCtx}\}) \\
\mathtt{AirObs} &= \mathtt{Obs}(\mathtt{Air}, \{\mathtt{AvgTmpMeas}\}, \bot) \\
\mathtt{SiteObs} &= \mathtt{Obs}(\mathtt{Site}, \{\mathtt{IdMeas}\}, \bot) \\
\mathtt{YearObs} &= \mathtt{Obs}(\mathtt{TemporalRange}, \{\mathtt{YearMeas}\}, \bot) \\
\mathtt{PlotCtx} &= \mathtt{Ctx}(\mathtt{Within}, \mathtt{PlotObs}) \\
\mathtt{AirCtx} &= \mathtt{Ctx}(\mathtt{Near}, \mathtt{AirObs}) \\
\mathtt{SiteCtx} &= \mathtt{Ctx}(\mathtt{Within}, \mathtt{SiteObs}) \\
\mathtt{YearCtx} &= \mathtt{Ctx}(\mathtt{Within}, \mathtt{YearObs}).
\end{aligned}$$

The model $M = (O, K, W)$ shown in Fig. 3(b) is defined as:

$$O = \{\texttt{TreeObs}, \texttt{PlotObs}, \texttt{AirObs}, \texttt{SiteObs}, \texttt{YearObs}\}$$
$$K = \{(\texttt{SiteObs}, \{\texttt{Id}\}), (\texttt{YearObs}, \{\texttt{Year}\}), (\texttt{AirObs}, \{\texttt{AvgTmp}\}\}$$
$$W = \{(\texttt{TreeObs}, \{\texttt{Spp}\}, \{(\texttt{Within}, \texttt{PlotObs})\})$$
$$(\texttt{PlotObs}, \{\texttt{Label}\}, \{(\texttt{Within}, \texttt{SiteObs})\})\},$$

An **instance** $I \subseteq Obs$ of a model $M = (O, K, W)$, denoted $I : M$, consists of a set of observations that are instances of types in O. If $I : M$, then I must satisfy the key and weak-entity constraints of M. These constraints are the same as those of standard ER models, but apply indirectly through observations and context. For example, if $o_1, o_2 \in I$ are of type $\texttt{O}$ in M such that $(\texttt{O}, \{\texttt{C}_1, \ldots, \texttt{C}_n\}) \in K$ and both o_1 and o_2 have the same characteristic instances (implying the same characteristic values, see Fig. 4) for $\texttt{C}_1$ to $\texttt{C}_n$, then both instances must reference the same entity instance. Additionally, if $I : M$, we require $enter(o)$ to be consistent for each $o \in I$.

2.3 Annotation

Given a data set D and a conceptual model M, we annotate D with M by relating attributes of D to measurements in M. The result of this process is an *annotation* $A = (D, M, \Sigma)$ consisting of a set of mappings $\Sigma \subseteq V \times O \times C$, where V is a set of attribute names, O is a set of observation types, and C is a set of characteristic types. If $(\texttt{V}, \texttt{O}, \texttt{C}) \in \Sigma$, we require that $\texttt{V}$ be an attribute of D, $\texttt{O}$ be an observation type in M, and $\texttt{C}$ be a characteristic type for some measurement type of $\texttt{O}$.

Example 4 (Annotation). The annotation $A = (D, M, \Sigma)$ for data set D of Fig. 1(a) and model M of Fig. 3(a) consists of the mappings:

$$\Sigma = \{(tree, \texttt{TreeObs}, \texttt{Id}), (spp, \texttt{TreeObs}, \texttt{Spp}),$$
$$(yr, \texttt{TemporalRangeObs}, \texttt{Year}), (dbh, \texttt{TreeObs}, \texttt{Dbh})\}.$$

Additional rules are often needed to define Σ, e.g., for converting data-set values to allowable values of a measurement standard. Thus, a set Σ is often accompanied by more complex expressions. Observational models may also contain measurements not directly linked to data-set attributes. In Fig. 1(b), e.g., we may know that all plots of the study have a $10m^2$ area, which typically would not be stored in a data set since the corresponding column would contain the same value in every row.

Annotations provide a mechanism to determine the semantics of attributes in a data set. For instance, from an annotation we can determine for each attribute: (1) the corresponding measurement in the conceptual model; (2) the observation in which the measurement was made; (3) the characteristic, measurement standard, and entity associated with the attribute; (4) other attributes of the data set associated with the same observation; and (5) the observations, measurements, and attributes serving as context for the attribute. It is also possible to construct schema mappings (i.e., views [17]) from annotations that map instances of data-set schemas to instances of observational models. Such mappings can be used to generate instances of the model, query data sets via the model, or integrate data sets across models.

2.4 Data Summarization

Meaningful summaries of data are often constrained by the direction of context relationships. In particular, similar to "roll-up" operations in multidimensional and statistical databases [15], summarization is often performed over contextualized observations and measurements, where measurements of observations that are "lower" in a context hierarchy are summarized by observations that are "higher" in the context hierarchy. For instance, computing an average water-treatment depth by transect in Fig. 3(c) is not meaningful, since each transect has exactly one depth. However, computing average tree-trunk diameter by transect in Fig. 3(c) is a generally meaningful summarization.

The types and constraints defined within observational models can be used to enable summarization testing [18], i.e., to automatically determine and compute meaningful summarizations. For instance, the measurement standard (e.g., nominal, ordinal, interval, ratio) determines the kinds of summaries that can be applied to an observation [14]. Cardinality constraints on observations imposed by context relationships also can suggest summaries. For instance, in Fig. 3(b) each plot within a site contains a *single* average air temperature, which can be used to compute average air temperatures by site (via the plots within the site). Key and weak-entity constraints also enable summarization by determining when two observations reference the same entity. For example, averaging soil acidity by tree species in Fig. 3(c) is made possible by first averaging acidity for each tree entity, and then aggregating over the set of entities of each species.

Finally, annotations allow summarizations over data-set attributes to be analyzed and computed according to the constraints of the corresponding observational model. For example, given a desired summarization expressed over data-set attributes, the corresponding measurement types within the model can be obtained, and then used to check whether the summarization is meaningful, and if so, to determine how it should be carried out (i.e., based on context relationships, key, and weak-entity constraints).

3 Related Work

A number of data models [4,5,6] and ontologies [7,8,9,10] have been proposed to support observational data (see [3] for a general survey). Our approach differs by providing formal and generic constructs for describing observations, measurements, and contexts that are compatible with well-established conceptual-modeling languages (ER, UML, description logics) and suitable for data annotation. Our approach also supports generic context relationships that are either missing or provided only through specific properties in existing approaches (e.g., recording when or where a measurement was taken). In [10], we describe an OWL ontology developed within the SEEK project[5] that includes concepts similar to those presented here. We extend these ideas in this paper by: (1) identifying and formalizing the constructs of Fig. 2; (2) providing a general definition and formalization of context that can reference arbitrary relationships; (3) defining observational models that include key and weak-entity constraints; and (4)

[5] Science Environment for Ecological Knowledge: `http://seek.ecoinformatics.org`

describing a formal approach for annotating data sets with observational models. In [10], we define concept hierarchies and properties for measurement standards including units and unit conversions, which can also be used in the framework presented here.

Approaches for representing context have been widely studied in logic [16,19] and conceptual modeling [20,12,11]. For instance, in [11], an ER model is extended by adding "weak attributes" to support context and data quality; in [12], ORM extensions are proposed to support context-aware applications; and in [20] context is modeled via sets of objects that can be related via classification, generalization, and attribution. In contrast, our approach distinctly separates observations of entities from entities and indirectly assigns context to entities via observations, thus providing additional flexibility for describing observational data and associated context relationships. Similar to context, a number of ER extensions have also been proposed to explicitly support temporal aspects of data [13]. In [21,22,23], approaches for reverse-engineering databases into ER models are proposed, where [21] defines an approach to generate ER models from denormalized relational sources (as in Fig. 1). Many annotation approaches have been proposed, ranging from column-level tagging [24] to query-based mappings [17]. Our approach differs by employing simple annotations to observational models (as opposed to arbitrary ontologies) from which more complex mappings can be constructed. Finally, summarization is well-established in multidimensional and statistical database systems with techniques developed for testing summarizability [15,18], and our framework can directly leverage these approaches.

4 Summary and Future Work

We have presented an approach for modeling observational data that extends existing conceptual-modeling frameworks by adding new constructs for representing observations, measurements, measurement standards, and context. The benefits of our approach include a formal and generic treatment of observation context, the ability to decouple observational models from conceptual descriptions (allowing observational data to be described via shared ontologies), an approach for simplifying data annotation (e.g., based on key, weak-entity, and context constraints), and support for data summarization. These benefits directly address challenges in interpreting and integrating heterogeneous observational data that are critical for supporting broad-scale scientific analyses.

We have implemented a number of prototype tools within the SEEK project based on an earlier version of the framework presented here. These tools support semantic annotation and discovery of observational data sets described in the EML[6] metadata language. We intend to extend these tools to support the constructs and annotation approach presented here. We are also exploring summarization capabilities and the merging of multiple data sets via observational models. Finally, we are developing a number of domain-specific ecological ontologies to support the annotation of ecological data within our framework.

[6] `http://knb.ecoinformatics.org/software/eml`

References

1. Andelman, S., Bowles, C., Willig, M., Waide, R.: Understanding environmental complexity through a distributed knowledge network. BioSciences 54(3), 240–246 (2004)
2. Ellison, A., et al.: Analytic webs support the synthesis of ecological datasets. Ecology 87, 1345–1358 (2006)
3. Madin, J., Bowers, S., Schildhauer, M., Jones, M.: Advancing ecological research with ontologies. Trends Ecol. Evol. 23(3), 159–168 (2008)
4. Cox, S.: Observations and measurements. Technical Report 05-087r4, OGC (2006)
5. Tarboton, D., Horsburgh, J., Maidment, D.: CUAHSI community observations data model (ODM), version 1.0 (2007), http://water.usu.edu/cuahsi/odm/
6. Cushing, J., Nadkarni, N., Finch, M., Fiala, A., Murphy-Hill, E., Delcambre, L., Maier, D.: Component-based end-user database design for ecologists. J. Intell. Inf. Syst. 29(1), 7–24 (2007)
7. McGuinness, D., et al.: The virtual solar-terrestrial observatory: A deployed semantic web application case study for scientific research. In: AAAI (2007)
8. Williams, R., Martinez, N., Goldbeck, J.: Ontologies for ecoinformatics. J. of Web Semantics 4, 237–242 (2006)
9. Raskin, R.: Enabling semantic interoperability for earth science data (2004), http://sweet.jpl.nasa.gov
10. Madin, J., Bowers, S., Schildhauer, M., Krivov, S., Pennington, D., Villa, F.: An ontology for describing and synthesizing ecological observation data. Eco. Inf. 2, 279–296 (2006)
11. Tu, S., Wang, R.: Modeling data quality and context through extension of the ER model. In: Workshop on Information Technologies and Systems (1993)
12. Henricksen, K., Indulska, J., McFadden, T.: Modelling context information with ORM. In: OTM Workshops (2005)
13. Gregersen, H., Jensen, C.: Temporal entity-relationship models – a survey. TKDE 11, 464–497 (1999)
14. Stevens, S.: On the theory of scales of measurement. Science 103, 677–680 (1946)
15. Lenz, H., Shoshani, A.: Summarizability in OLAP and statistical data bases. In: SSDBM (1997)
16. McCarthy, J.: Notes on formalizing context. In: IJCAI (1993)
17. Beeri, C., Levy, A., Rousset, M.: Rewriting queries using views in description logics. In: PODS (1997)
18. Hurtado, C., Mendelzon, A.: OLAP dimension constraints. In: PODS (2002)
19. Guha, R., McCarthy, J.: Varieties of contexts. In: International and Interdisciplinary Conference on Modeling and Using Context (2003)
20. Analyti, A., Theodorakis, M., Spyratos, N., Constantopoulos, P.: Contextualization as an independent abstraction mechanism for conceptual modeling. Inf. Syst. 32(1), 24–60 (2007)
21. Petit, J., Toumani, F., Boulicaut, J., Kouloumdjian, J.: Towards the reverse engineering of denormalized relational databases. In: ICDE (1996)
22. Alhajj, R.: Extracting the extended entity-relationship model from a legacy relational database. Inf. Syst. 28(6), 597–618 (2003)
23. Davis, K., Aiken, P.: Data reverse engineering: A historical survey. In: WCRE (2000)
24. An, Y., Borgida, A., Mylopoulos, J.: Discovering the semantics of relational tables through mappings. J. Data Semantics VII, 1–32 (2006)