Overview of Data Properties

Introduction

In environmental studies, the data collection process can be divided into three phases: the instrument-calibration phase, the data-gathering phase, and the post-processing phase. In the *instrument-calibration phase*, the scientist may calibrate the data-sensing equipment based on knowledge about the site being studied. Anomalies may be introduced into the collected data. Some causes of anomalies in sensor data include noise from external sources, hardware noise, inaccuracies and impressions in sampling methods and derived data, and various environmental effects [5]. In the *data-gathering stage* the scientist collects the data using calibrated equipment; at this stage, scientists typically rely on their knowledge and experience in the field to manually identify and distinguish one type of anomaly from another given the equipment and prevailing climatic conditions. In the *post-processing* stage, data are generally processed and stored in a database by the scientist. In this stage, scientists typically define and use site-specific anomaly detection processes to identify and differentiate anomalies in the processed data. Even though anomaly detection in sensor data is important to draw reliable environmental conclusions, scientists rarely share or reuse knowledge about their data processes with other scientists until a process is established by the scientific community mostly because of the lack of a well-defined methodology for doing so and lack of tool support for such sharing.

The challenges that contribute to the limited sharing of scientific data processes and knowledge include:

- The lack of documented properties and associated information, e.g., statements about who contributed the property and the applicability of the property. Properties of interest include checks on the following: upper and lower limits of variables; limit on the rate of change between data; detection of subsequence of data with the same value; analysis of two or several parameters at the same point in time; spatial continuity or consistency checks in which the values of adjacent stations are allowed to differ within a certain bound; and diagnostic equations to which data are expected to adhere;

- Differences in checking criteria and processes across the scientific community;
- Ambiguity in natural languages when describing properties;
- Complexity in properties when dealing with time and multiple criteria;
- Technical knowledge required by scientists, especially when dealing with relationships among properties of various types and issues related to data repositories; and
- Use of embedded or hard-coded property checking in many existing systems, making it difficult to reuse and refine properties.

Data Property Categorization

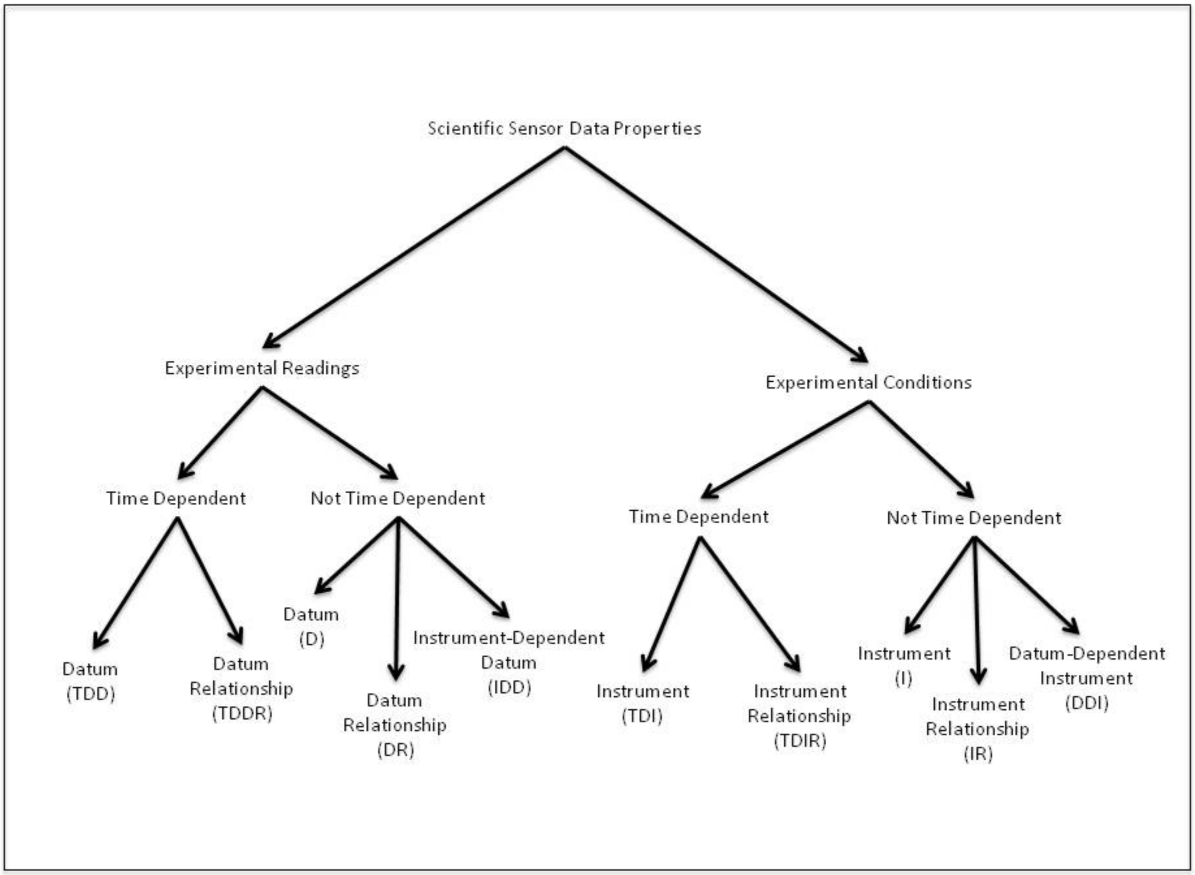

Figure 1. Scientific sensor data properties categorization.

Fig. 1 provides a category of types of properties of interest. Properties labeled experimental readings are divided into the following five subcategories:

Datum[1]*:* A datum (D) property specifies the expected value of a single sensor reading. A sensor reading is compared against a pre-defined or historical value. Example: *The relative humidity percentage should always be greater than or equal to 0 and less than or equal to 1* [38].

Time-Dependent Datum: A time-dependent datum (TDD) property specifies the expected value(s) of a single type of sensor, where the readings are filtered by date and time. The selected sensor readings are compared against a predefined value or a historic value. Example: *During daylight on May 12th, the dry bulb temperature should be less than or equal to 103°F* [39].

Datum Relationship: A datum-relationship (DR) property specifies the relationship between two or more types of sensor readings. A DR property can be used to compare sensor readings against readings from other types of sensors, against a predefined constant value, or against an historic value. Example: *Temperature < Wet-Bulb-Temperature < Dew-Point-Temperature* [40].

Time-Dependent Datum Relationship: A time-dependent datum relationship (TDDR) property specifies the relationship between two or more related sensor readings that are filtered based on time. The selected readings may be compared against each other, against a predefined value, or an historic value. TDDR properties capture relationships within time series data and datasets behaviors dependent on time. Example*: No two measurements of the consensus subset can differ by more than 1/8 of the maximum measurable velocity, where the consensus subset is created each hour by applying the consensus algorithm from the ten 6-minute radial velocity measurements on each antenna beam* [41].

Instrument-Dependant Datum: An instrument-dependant datum (IDD) property is one that specifies a property about an instrument that influences behavior of the sensor readings. Example: *If the profile lies close to land and the depth is less than 50 meters, the observed value should lie within 5 standard deviations from the mean value*.

[1] For this work “Datum” is used to denote the singular form of “Data”

Experimental conditions properties are divided into the following five subcategories:

Instrument: An instrument (I) property specifies the expected behavior of an instrument by describing an attribute of the instrument. The attribute is compared against either a predefined value or an historic value. Example: *The collection-time sensor voltage should fall inside the expected range*.

Time-Dependent Instrument: A time-dependent instrument (TDI) property captures the expected behavior of a single instrument that is dependent on time. The instrument reading is compared against a predefined constant value, a historic value, or a time entity in a given time constraint. Example: *Based on the time since last scanned, each radar must scan a 360-degree sector at the lowest two elevations every 2.5 minutes*.

Instrument Relationship: An instrument relationship (IR) property captures the relationship between one or more related instruments. An IR property can be used to compare the behavior of the instrument. Example: *If a current meter is used, at least one of the HCSP/HCDT or NSCT/EWCT sensor couples must be present*.

Time-Dependent Instrument Relationship: A time-dependent instrument relationship (TDIR) property captures the relationship between two or more related instruments and expected behavior based on time. A TDIR property can be used to compare instrument behavior dependent on a time. Example: *Based on the time since last scanned, perform sector scans of storms with 2 or more radars every 1-minute*.

Data-Dependant Instrument: A datum-dependant instrument (DDI) property captures a known datum or datum relationship whose value influences instrument behavior, or causes an instrument's action. DDI properties capture continuity problems. Example: *If there is no change in current direction data, the system must generate an error alert*.

Data Property Specification and Pattern System (D-SPS).

Data property specification using the Data Property Specification and Pattern System (D-SPS) is composed of the following components: *patterns, scopes*, *Boolean statements,* and *data* *property categories*. D-SPS data property specifications use *patterns,* which are common occurring data properties, to define how *Boolean statements* are evaluated over data subsequences defined by *scopes*. A *data property category* refines the pattern selection based on a data property's dependency on time, number of sensors, and sensor relationships.

D-SPS Patterns and Scope

The D-SPS uses *patterns* and *Boolean statements* to specify data properties. A ***property pattern*** is a high-level abstraction describing a commonly occurring property about a scientific dataset. Property patterns for this work were defined based on the commonly occurring properties extracted from the data property categorization. The SPS definitions, without the collection-time extension, were adapted to create the initial D-SPS. The original D-SPS included the quantitative, not time-constrained, patterns: *Universality*, *Absence*, *Existence*, *Precedence* and *Response*. ***Boolean statements*** use relational operators (<, ≤, =, ≠, ≥, >) applied to data readings to establish relationships between sensors. For this work, a Boolean statement is classified as type "single sensor reading" (denoted by S) when a sensor reading is compared to a constant numerical value. A Boolean statement is classified as type "multiple sensor reading" (denoted by M) when a sensor reading is compared to another sensor reading. For example, *temp* < 20 denotes a single sensor reading type and *temp*=*age_temp* denotes a multiple sensor reading type. It is assumed that in both cases that the sensor readings are of the same measurement type.

The data property categorization makes a distinction between properties that are dependent on time and those that are not, and within these two categories, there are sub-categories that capture the dependencies based on the number of sensors and sensor relationships associated with the property. Given the similarities between the categories in the data property categorization and the patterns supported by the D-SPS, it is possible to relate the data patterns from the D-SPS to the categories in the data property categorization. For example, the data property category of type "time dependent datum" captures an expected behavior of a single sensor readings dataset that depend on time. The D-SPS provides patterns to evaluate a single data property's occurrence (*Maximum Duration*, *Minimum Duration*) or recurrence (*Bounded Recurrence*) that are dependent on time. Similar relationships can be built for all of the data property categories and data patterns. The relationships are constructed based on the type and number of Boolean statement(s) to be evaluated and on whether the Boolean statement(s) evaluation depend(s) on time.

Time Dependent: Data property categories that depend on time are of two types: properties that evaluate a single sensor readings dataset over time, i.e., *Time-Dependent Datum* and *Time-Dependent Instrument*, and properties that evaluate multiple sensor readings datasets over time, i.e., *Time-Dependent Datum Relationship* and *Time-Dependent Instrument Relationship*.

For *time-dependent datum* and *time-dependent instrument* properties, the D-SPS patterns are restricted to *Minimum Duration, Maximum Duration,* and *Bounded Recurrence*; these patterns assess a single Boolean statement of type *S* over time.

For time-related categories with multiple sensor readings datasets over time, i.e. *time-dependent datum relationship* and *time-dependent instrument relationship*, the D-SPS patterns are restricted to: *Minimum Duration, Maximum Duration,* and *Bounded Recurrence* for which a pattern that assess a single Boolean statement of type *M* over time is used, and *Bounded Response* and *Bounded Invariance* that assess a combination of two Boolean statements of any combination of type *S* and *M*.

For timed patterns, units of time are based on a sequence of readings and indexing values in a dataset of interest are assumed to be ordered time stamps, with equal constant time resolution, i.e., the time stamps are all of the same time measurement unit and change at a same constant time rate. The scientist is expected to align the unit of time used in the dataset to the unit of time used in the data property specification. For example, consider a scientist that wants to specify a property *P* that captures a Boolean statement *B* that must be evaluated every 5 minutes over dataset *D*. If the time resolution for *D* is minutes, *B* is evaluated every 5 units (data readings) in *D*. If the time resolution for *D* is seconds, B is evaluated every 300 units (data readings) in *D*.

Not Time Dependent: Similarly, data property categories that do not depend on time are of three types: properties that evaluate a single sensor readings dataset, i.e., *Datum* and *Instrument*; properties that evaluate *multiple sensor readings* within the same data property category, i.e., *Datum Relationship* and *Instrument Relationship*; and properties that evaluate multiple sensor readings from different data property categories, i.e., *Instrument-Dependent Datum* and *Datum-Dependent Instrument*.

For *datum* and *instrument* properties, D-SPS patterns are restricted to: *Absence*, *Universality,* and *Existence*; these patterns assess a single Boolean statement of type *S* over the sensor readings dataset. For *datum relationship* and *instrument relationship*, D-SPS patterns are restricted to: *Absence, Universality, Existence, Precedence,* and *Response*. The *Absence*, *Universality* and *Existence* patterns evaluate a single Boolean statement of type *M* over the sensor readings datasets. The *Precedence* and *Response* patterns combine two Boolean statements over the datasets. The Boolean statements can be one combination of type *S* and *M*.

For *Instrument-Dependent Datum* and *Datum-Dependent Instrument*, the D-SPS patterns are restricted to *Precedence* and *Response;* these properties assess a combination of two Boolean statements over the sensor readings dataset. The Boolean statements can be of any combination of type *S* and *M*.

A different number of Boolean statements are evaluated by the patterns over a set of scopes. ***Scopes*** are classified as *Global*, *Before (del)*, *Between (del, del)*, *AfterUntil (del,del)*, where *del* can be a numeric value (*number.number*) or a time stamp (*monthy-day-year hours:minutes:seconds:milliseconds*). Every pattern definition includes the name of the pattern, a pattern type, the Boolean statement type associated with the property, a natural language description of the pattern, the data categories from the data categorization that use the pattern, a formal representation of the pattern, and a pseudo-code representation of the pattern. You will be given the code that corresponds to the patterns and cope. The number of scopes associated with every pattern corresponds to the number of sensor datasets that can be evaluated by the pattern's Boolean statement(s).

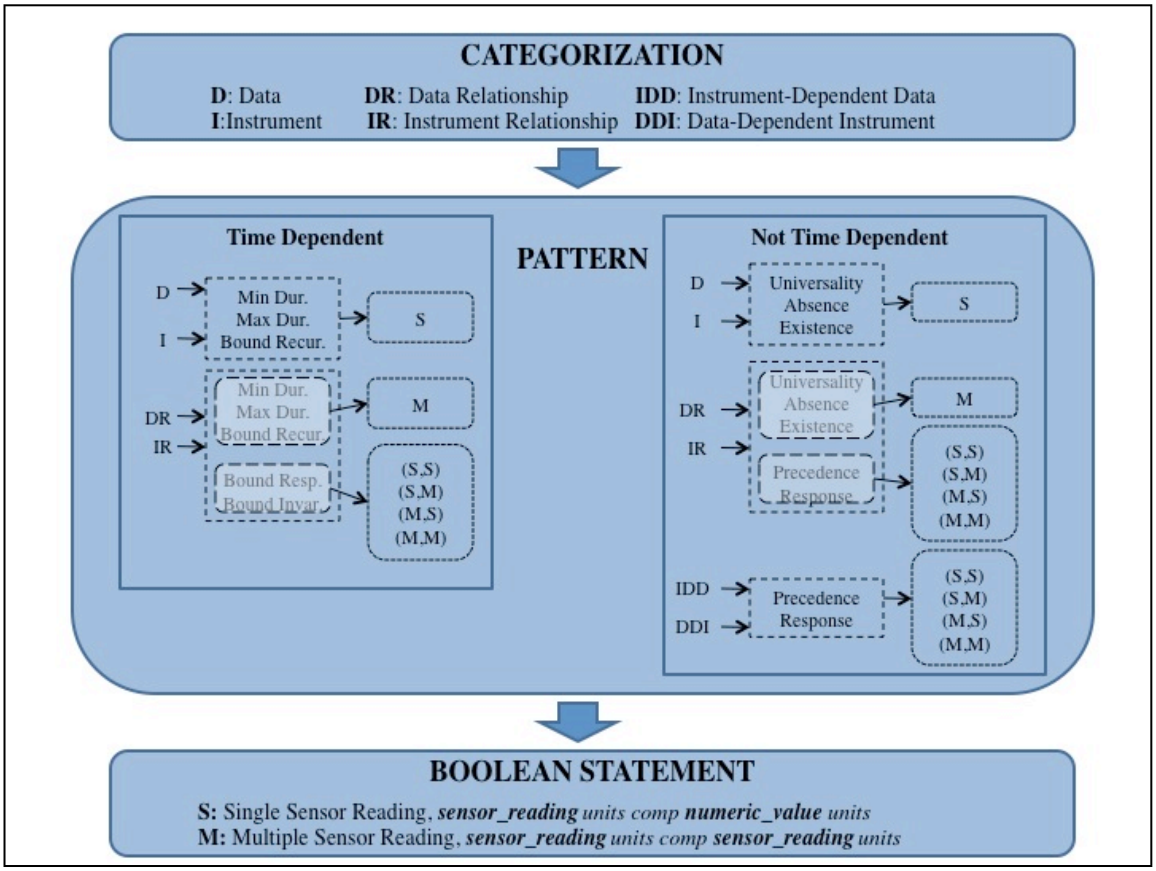

Figure 2. The relationship between the categorization, patterns and Boolean statement

Examples

Example 1:

> *"For May 12th at daytime, the dry bulb temperature should be smaller than or equal to 35.0 degrees Celsius" [48].*

A scientist using the D-SPS to specify the data property selects *Datum* as the data property category because she is interested in evaluating only one set of sensor data. The scientist then specifies the scope for the data property. The desired scope encompasses the data readings recorded on May 12th during *daytime*, i.e., between 6:15:00 AM and 8:00:00 PM, as defined by the project's documentation available to the scientist. The scientist uses *Between L and R* as the scope, where *L* stands for time stamp *05-12-2010 06:15:00.0* and *R* stands for time stamp *05-12-2010 20:00:00.0*. The scientist examines the available documentation associated with the sensor readings of interest, and identifies *temp* as the variable containing the dry bulb temperature reading in degrees Celsius. The scientist defines a Boolean statement to be *temp* ° ≤ *35°C*. The scientist then selects $Universality_S$ as the pattern to evaluate if it is always the case, over the scope of data reading values, that *temp* ° ≤ *35°C*. The high level specification using D-SPS is presented in Table 1.

Table 1. High-level D-SPS specification for a dry-bulb temperature property.

Document Specification:	"On May 12th during the daytime, the dry bulb temperature should be smaller than or equal to 35.0 *°C* "[73].
D-SPS Representation:	*Datum, Between(05-12-2010 06:15:00.0, 05-12-2010 20:00:00.0), UniversalityS(temp:°C,≤ ,35:°C)*
Assumptions:	*Daytime:* time-frame between 05:12:2010:06:15:00 and 05:12:2010:20:00:00 from scientific expert knowledge. *temp:* dry bulb temperature sensor dataset.

Example 2:

D-SPS also allows scientists to specify properties that capture problems in the equipment. For example, consider the data property that captures the effect of temperature over an equipment diagnostic value (agc) that is miscalculated whenever high temperatures occur during the summer:

"For all dry bulb temperature dataset values, it is always the case that, If temp>35 °C, then agc<70".

The data property can be captured using D-SPS as follows. The scientist is interested in evaluating two related sensor datasets, thus she selects *Data Relationship* as the data type category for the property. The property will evaluate all of the *temp* readings in the dataset, thus the scientist selects *Global* as the scope. The scientist realizes that whenever dry bulb temperature temp is greater than 35°C, the *agc* diagnostic flag should be less than 70; otherwise, if temp is greater than 35 and the *agc* value is also greater than 70, then agc is likely being miscalculated possibly due to the effect of the high temperature over the equipment in the remote research location. The scientist uses the *Response(T,P)* pattern to capture the ordering of the Boolean statements T and P. Boolean statement T evaluates whether the *temp>35 ℃* and Boolean statement P evaluates whether *agc<70*. The *Response(T,P)* pattern evaluates if is always the case that when *temp>35 ℃* then *agc<70* is also true. The high level specification using D-SPS is presented in Table 2 and the derivation tree:

Table 2. High-level D-SPS specification for a diagnostic sensor property.

Document Specification:	*"For all dry bulb temperature dataset values, it is always the case that, If temp>35 °C, then agc<70.0."[73]*
D-SPS Representation:	*Global,ResponseS(temp:°C,>,35:°C, agc:n/a,<,70.0:n/a)*
Assumptions:	*temp:* dry bulb temperature sensor dataset. *agc:* diagnostic flagging sensor dataset.