Data Quality

Accuracy

Accuracy is defined as the closeness between a value v and a value v', considered as the correct representation of the real-life phenomenon that v aims to represent.

Syntactic accuracy is the closeness of a value v to the elements of the corresponding definition domain D.

Syntactic accuracy is measured by means of functions, called comparison functions, that evaluate the distance between v and the values in D. Edit distance is a simple example of a comparison function.

Semantic accuracy is the closeness of the value v to the true value v'.

Suppose we swap the names in two tuples. The exchange is an example of a semantic accuracy error: indeed, for record 1, a value named that way would be admissible, and thus it is syntactically correct. Nevertheless, it does not represent the reality; therefore a semantic accuracy error occurs.

Completeness

Completeness can be generically defined as *the extent to which data are of sufficient breadth, depth, and scope for the task at hand.*

Schema completeness is defined as the degree to which concepts and their properties are not missing from the schema.

Column completeness is defined as a measure of the missing values for a specific property or column in a table.

Population completeness evaluates missing values with respect to a reference population.

The closed world assumption (*CWA*) states that only the values actually present in a relational table r, and no other values represent facts of the real world.

In *CWA*, We can define:

- *a value completeness*, to capture the presence of null values for some fields of a tuple;
- *a tuple completeness*, to characterize the completeness of a tuple with respect to the values of all its fields;
- *an attribute completeness*, to measure the number of null values of a specific attribute in a relation;
- *a relation completeness*, to capture the presence of null values in a whole relation.

Time-Related dimensions

Currency concerns how promptly data are updated.

Volatility characterizes the frequency with which data vary in time.

Timeliness expresses how current data are for the task at hand. The time- liness dimension is motivated by the fact that it is possible to have current data that are actually useless because they are late for a specific usage.

Consistency

The consistency dimension captures the violation of semantic rules defined over (a set of) data items, where items can be tuples of relational tables or records in a file. With reference to relational theory, integrity constraints are an instantiation of such semantic rules.