1 Annotation constraints

How to annotate the above tdataset using "key yes", "distinct yes" and "identifying yes".

- The constraint "key yes" is applied to a measurement type in the annotation. If a measurement type of an observation type is denoted as *key yes*, it means the measurement value of one observation instance is the key for this observation's entity instance. If two observation instances have the same measurement value, they are two different different observation instances from the *same entity instance*. The measurement types of one observationo type which are specified with "key yes" are also called the **key measurement**(s) of this observation type.
- The constraint "distinct yes" is applied to an observation type in the annotation. It need to be used together with "key yes" to catch the semantics of *same observation instance*. I.e., if an observation type is specified with "distinct yes" but it does not have any key measurement types. The annotation is invalid. Besides this, we also have the following **Rule 1** between the "key yes" and "distinct yes" constraints. I.e., when an observation type is denoted with "distinct yes", all of its measurement types should be specified with "key yes". Otherwise, there would be inconsistency. We show it using an example.
- The constraint "identifying yes" is applied to an context type description in annotation. It is used to identify that the uniqueness of one entity instance is not only decided by its own key measurements, but also decided by its context key mensurements. Based on its semantic measning, the following **Rule 2** is applied.
- Rule 1: If an observation type is specified with *distinct yes*, all its measurement types are automatically marked with *key yes*.
- Rule 2: If the context of an observation type is specified with *identifying yes*. This observation type MUST have some key measurement types. And the context observation type also MUST have key measurement types too.

Given the data in Table 1(b) and the annotation in Figure 1, the following example shows why **Rule** 1 is needed.

According to the annotation, if two plots have the same value for "PlotName", they represent the same plot observation. Obviously, it has problem to interpret the data in Table 1 with this annotatioin. E.g., the first and second rows catch information about plot with EntityName A. According to the annotation, they should be the same plot obervation. However, the data shows that this one plot has two different areas 1.0 and 1.1, so, there is confusion here.

For this case, we can have several ways to make the annoation consistent.

- The first way is: when an observation type is denoted by "distinct yes", all its measurement types should be denoted with "key yes" automatically.
- The second way is: get ride of the onstraint "distinct yes". This way, when all the measurements are denoted with "key yes", implicitly, the same entity instance only corresponds to the same observation instance. [FROM HP: I PREFER THIS ONE.]

For the above example, the right way to annotate the data is to get rid of the *distinct yes*. So that the same PlotName can points to the same plot entity, but the different areas mean that the two different observations get two different values.

2 A more complex use case

Given the dataset in Table ??, users have different situations to catch.

• Requirement 1: *Plot* with label "1" should refer to the same one physical plot (i.e., the Plot in the first 4 row means the same thing); similarly, *plot* with label "2" should refer to the second physical plot (i.e., the Plot in the last 4 row means the same thing).

plt	spp	dbh		plt	area	spp	dbh	
Α	piru	35.8		А	1.0	piru	35.8	
Α	piru	36.2		А	1.1	piru	36.2	
В	piru	33.2		В	2.0	piru	33.2	
В	abba	34		В	2.0	abba	34	
(a)				(b)				

Table 1: Dataset

observation "o1" distinct yes entity "Plot" measurement "m1" key yes characteristic "PlotName" standard "Nominal" observation "o2" entity "Tree" measurement "m3" key yes characteristic "TreeName" standard "TaxonomicName" measurement "m4" characteristic "DBH" standard "Centimeter"' context identifying yes "o1" map "plt" to "m1" map "spp" to "m3" map "dbh" to "m4"

observation "o1"									
entity "Plot"									
measurement "m1" key yes									
characteristic "PlotName"									
standard "Nominal"									
measurement "m2"									
characteristic "area"									
standard "sqft"									
observation "o2"									
entity "Tree"									
measurement "m3" key yes									
characteristic "TreeName"									
standard "TaxonomicName"									
measurement "m4"									
characteristic "DBH"									
standard "Centimeter"'									
context identifying yes "o1"									
map "plt" to "m1"									
map "area" to "m2"									
map "spp" to "m3"									
map "dbh" to "m4"									

(a)

(b)

Figure 1: Annotation of Table 1

- This can be captured in annotation by putting "Distinct yes" for observation type *Plot* and "key yes" for its measurement type *PlotLabel*.
- Requirement 2: *sub-plots* with the same lable should refer to the same physical sub-plot if they are within the same plot; but the sub-plot with the same label with different *Plot* label are different sub-plots. E.g., Row 1 {Plot=1, sub-plot=A} refers the same sub-plot as that in Row 2, but is different from the one in Row 5 {Plot=2, sub-plot=A}.
 - This can be captured by putting "Distinct yes" for observation type SubPlot, "key yes" for its measurement type SubPlotLabel. We need to denote Plot is its context and with identifying yes specified on this context.
- Requirement 3: *Tmnt* with the same lable should refer to the same treatment process (So that we can aggregate on different treatment process, e.g., on "X" or on "C".) But the treatment at different sub-plot should refer to different treatment.
 - The first requirement can be captured by treating all the Treatment with value "X" as the different entity instances with the same type (TmntType). The second requirement can be captured by treating the treatments in different sub-plots as observations of the different entity instances. I.e., treatments in row 1 and row 3 are of but are different observation instances which are of different entity instances.

Tmnt has the context sub-plot identified with "identifying yes".
Summarization questions:
Give me the average weight of the individuals with treatment "X". How can this question be answered after the annotation and materialization? This need to be answered after we annotate Sp, Ind, and wt. The group by should be only on Tmnt. This can be expressed using the OM query with a condition in the avg(Tmnt), but not avg(distinctTmnt). constraint.

• Requirement 4: *Ind* with the same label and the same species name (Sp) should refer to the same bird observation. But the individual (with the same label and the same species name) at different sub-plot should refer to different bird observations.

The annotation can be done as follows.

Plot	sub-plot	Tmnt	Sp	Ind	wt
1	А	Х	Aus	1	10
1	А	С	Bus	1	20
1	В	Х	Aus	3	10
1	В	С	Bus	4	10
2	А	Х	Aus	1	20
2	А	С	Bus	4	10
2	В	Х	Aus	5	20
2	В	С	Bus	4	10

observation "o1" entity "Plot" measurement "m1" key yes characteristic "PlotLabel" standard "Nominal" observation "o2" entity "SubPlot" measurement "m2" key yes characteristic "SubPlotLabel" standard "Nominal" context identifying yes observation "o1" observation "o3" entity "Treatment" measurement "m3" key yes characteristic "Procedure" standard "Nominal" context identifying yes observation "o2" observation "o4" entity "Bird" measurement "m4" key yes characteristic "Species" standard "TaxonomicBirdName" measurement "m5" key yes characteristic "Individual" standard "Nominal" measurement "m6" characteristic "weight" standard "kg" context identifying yes observation "o2" map "Plot" to "m1" map "Sub-plot" to "m2" map "Tmnt" to "m3" map "Sp" to "m4" map "Ind" to "m5" map "wt" to "m6" Annotation

A dataset with more complex information