Data Quality, Fish and Ska music

When the topic of "Data Quality" comes up most people would, I suspect, list a set of quality measures, something like: completeness; consistency; uniqueness; currency; precision and reasonableness. It is rare to find two experts that agree on the exact set, for example many would have different names, perhaps timeliness rather than currency, or add some extra ones like say: conformity; availability; and coverage.

With just one exception, which I'll come to shortly, the exact list used is not overly important. Of course, everyone wants their list of metrics to be sufficiently long to cover all the categories of blunder that may befall the data, but not so long that they get confused between them. A good set of dimensions provides a checklist of validations to think about, each item inspiring a new range of tests. The fact that these dimensions are somewhat loosely defined is really nothing to worry about.

During the Middle Ages good Christians were allowed to eat Beaver or Barnacle Goose on Fridays, since they were both considered to be fish. These days we are much more rigorous, we know that an octopus is a type of mollusc, and a crab is a crustacean, we would not call either a fish. Well, except, what is a "fish" exactly? Stephen Jay Gould once pointed out that the modern definition is not quite consistent either, some types of fish are genetically more closely related to mammals than they are to other fish. In strict biological terms there might not be such a thing as a fish, but I still know one when I see one (most of the time). The concept may not be precisely defined, but it is still valuable.

There's a similar thing with musical genres, terms like Progressive Rock, Cool Jazz and Ska are quite hard to define. Indeed work contrasting some automated classification techniques with how good human experts are seems to suggest that the boundaries between genres are not generally agreed. Yet there are some artists that everyone would agree are Ska and that information would help many people to either select or avoid them.

So, in many other topics we can be fairly relaxed about the labels used. As long as everyone agrees on the core definition the fact that the edges are a bit fuzzy doesn't usually worry us. The data quality metrics are like that, mostly. I mentioned that in my opinion there is one exception. Which metric is so bad that I would consider its use harmful?

For me the completely illogical data quality dimension is the one that is usually called "accuracy", or "correctness", or "exactness", or "truth". This is because any one set of data can only be compared with other sets, in any comparison there will be some probability that either side could be wrong. We might use a list of client names and addresses that we believe are 99% accurate to validate a larger less trusted database, but this is checking consistency between the sets, a mismatch doesn't prove the database is wrong, it could be that this is one of the 1%. Each assertion of "correctness" is just a comparison and no real set of data can ever be guaranteed to be 100% accurate. If you think you know of such a set you just haven't been imaginative enough in your testing yet.