The dragon of data quality

Last week I suggested that it is important to keep both a readable version and an executable version of all of your quality tests. Those, like me, who have a naturally cynical bias will have spotted that I was a bit vague about exactly what sorts of things a "quality test" should contain. Partially this was because the expert users should be allowed to dictate what form the intent statement takes. If they want to incorporate a new ideas they should be allowed to do so, as data handling experts our role is to ensure we understand (and document of course) any novel concepts employed. Suppose an expert user wishes to add a test stating "all toves that are brillig should have a gimble value of more than 3", well that English sentence might be completely clear to the right type of practitioner but the rest of us will need a bit of explanation. As I said the need to allow domain experts to have the final say is absolute, but it is a bit of a feeble excuse. As data handling experts we should be able to articulate the majority of the concepts that are encountered in lists of quality tests before interviewing our first technical data expert.

So, what would I suggest needs clarification before starting a "Data Quality" initiative? Well, the requirement to agree the "dimensions of quality" is obviously a good starting point: completeness; uniqueness; consistency; timeliness; measurement accuracy; and reasonableness for example would be a realistic list of dimensions to focus on. The exact list can be discussed as long as everyone has agreed before we start writing rules. Some concept of the different data groups, both so we agree where to place relevant quality tests and what to call business objects. It's no good striving to detail what names of wellbores look like if we haven't first agreed what a "wellbore" is. We also need to have a rough idea of what the data is actually for, that is the business reason for keeping it, after all the goal here is to have data that is "fit for purpose" spending extra time or resources to refine values beyond the necessary is a waste. Understanding exactly where the data is to be found is important, it's hard to execute tests if we don't know have a path to what is being tested. There's also a concept of the "role" that each data location has, some data may be scratch information, used for trying out ideas, while other data may be the "official", "approved", "corporate" version and hence subject to more restrictions (and, one hopes, more trustworthy as a result). Having an agreed list of "data location types" is an essential step to controlling the quality. Once we've found the data, and we know how much we should be able to rely on it, then we can look at its structure (which elements are present, what attributes they have and so on). Often one key type of test is to check the conformance with predefined lists of valid values (usually called "reference lists"). Finally there is the concept of the "profile", that is some kind of description about different "ways of working", perhaps asset teams use different software, or national affiliates need to employ distinct business processes to conform with regulatory agencies. Often there will be a "corporate profile" that sets quality tests across the whole organisation and additional supplementary rules that apply in particular locations.

You need quite a bit of preparation to be properly equipped to face the dragon of data quality.