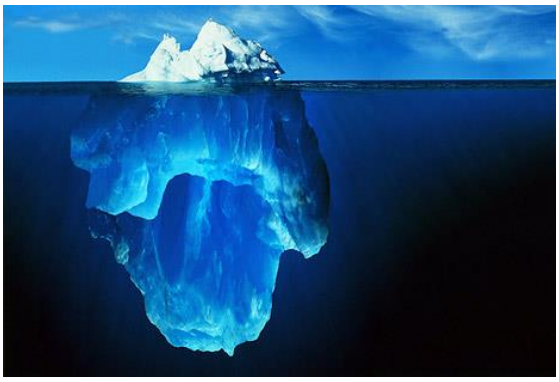
Return of the Notorious Iceberg

I recently had a discussion about data in a particular company, "of course 70% of our data is unstructured" said my companion. I understood what he meant, but the phrase still struck me as at the very least questionable. There are two things that make me nervous of this increasingly quoted statistic, both related in different ways to the question of what is being measured.



First of all it is not clear what this is 70% of, in most modern subsurface data sets something like 90% of the actual bits are occupied by seismic, so the 70% does not refer to data volume. But the things being counted here are clearly also not files, because then the number is way too low. In most oil companies there are big files, containing SEG-Y data, Petrel projects, Oracle databases and so on, and then there's a shared drive or two with a few million files (most of which are near duplicates). Measuring "file system objects" would surely result in 99.9% of them counting as unstructured.

So the statistic clearly isn't counting data volume or files but rather "things". I took part in a review of data within an oil company, one of the questions we were asked was what percentage of the wells had up-to-date completion diagrams. This turned out to be quite a contentious issue, first of all because (as the guys at PPDM have spent a lot of time pointing out) there is little agreement about what a "well" is. The exploration department's definition said there were 400 wells and, for the same data, the operations guys said there were 1,200. Secondly the drilling department's definition of "up to date" for a completion diagram was the last one they issued, whereas operations said that to be up to date it had to include annotations about the latest activities. In the end the answer to these types of questions depend more on your definitions than on the actual data held by the company. Those with no experience of dealing with E&P data welcome the fiction of "business objects", it allows them to say things like "if it takes two geoscientists one month to interpret 20 well logs, how long will it take six geoscientists to build a single static model?". Well maybe that is an exaggeration, but not by much. I am often asked by oil companies to count "well logs" and report on their quality, as though business objects were a real thing that could be agreed on (and ignoring the issue that "quality" can only be assessed by those that utilise the data).

The second challenge is the phrase "unstructured data", PowerPoint files are, of course, structured, otherwise we wouldn't know where to find the contents of slide 73. The thing is that while this "format data" tells us that we have a presentation, it doesn't help distinguish between a presentation about developing Alberta's tar sands and a collection of pictures of the company picnic. Of course there are a variety of tricks that everyone uses to "tag" important files: picking a meaningful name; placing it in a particular directory; editing its properties; keeping an index; and so on. So all data has degrees of "structured-ness".

I suspect this statistic about 70% of E&P data being "unstructured" is just the latest version of the "Data Iceberg" picture that was so popular 20 years ago (and which disappeared from presentations 10 years ago because no-one could take it seriously any more).