

Subject: Scholes cabin 5 (p): Metadata blues

17 February 2010 0004 UTM -56.3329, -0.0033

Dear Stirling,

I have been working all day and most of the night, getting sorted out for hand-over. It will be worked on by many other people, some perhaps many years in the future. I won't be there to explain it (and I probably would not remember anyway, you know what a scatterbrain I am), so it needs to be very clear what everything is, where it is, where the problems are, and what has been fixed.

Have your teachers ever moaned at you for handing in work without your name, the date and a heading on the paper? They should. That information is called metadata, and without it the work itself is quite useless. 'Meta' is a Greek word, which in this context means 'data about data' - the information that unlocks the dataset.

When I was a student, scientists still kept laboratory notebooks. They were sturdy books, on acid-free paper, and you had to write neatly. If you made a mistake, you were not allowed to rub it out or tear the page out - you had to draw a neat line through the wrong work and start again. We were not allowed to leave the experiment until the lab-book had been signed off by the supervisor. I hated that, because I often used to miss the last bus home. But there was a point to it - it was a way of keeping all the information together in a way that you (or someone else) could retrace your steps. Being able to repeat the experiment is fundamental to science.

Then along came computers. That should have made keeping track of the data easier, but instead it made it much worse. First of all, we all started to produce much more data than before. In lots of versions, with funny filenames like 'final' and 'very final'. Second, we scribbled our work down temporarily on scraps of paper rather than in notebooks, meaning to enter it in a neat table sometime, and either forgot to do so or made a mistake while doing so. Meanwhile, the scrap of paper was lost or unintelligible. Third, the software and the hardware changed so often that we could not read the data we had entered on the computer anyway. You know that big box under the desk in the basement? That is full of computer tapes from my PhD work that nobody in the world can read today.

So in this brave new world, we are having to learn a new set of metadata rules. In general we will not go back to pen-and-paper, although there is still a place for that. There is just too much data to write it all down, and too much risk of making a mistake while doing so. But we have to be even more disciplined than before. My students know better than to give me a flash memory without a name on it. Or a file with a silly filename like 'AM27j'. Or a data file where the first lines do not explain exactly what is in the file,

who produced it and when, how it is structured and what the units are. Or a computer programme without comments. After a while it just becomes a habit.

People like me (and you) who are not naturally tidy and organised have to make a special effort, otherwise we spend our lives hunting for stuff that we know we put somewhere. Or worse still, picking up the wrong stuff and making a big mistake. Trust me, it is less work in the end.

It is snowy and cold, but the sea is not rough. There is a prion (a type of bird) roosting in the hydraulics under the UCTD. I don't know if it is tired, or just hitching a ride.

Love,

Dad