

Subject: Scholes cabin 5 (p): Number crunching

4 February 2010 1255 UTM -70.5233, -8.1854

Dear Stirling,

We are loading cargo at the Atka ice-port, a section of ice-shelf only about 8 m above the sea level (in other words, there is about 25 m below the sea level, for a total thickness of at least 30 m - plenty to support even heavy loads). There are many containers lined up at a safe distance back from the ice shelf edge, in case a big chunk decides to crack off and become an iceberg. Yesterday we finally unloaded the two heavy bulldozers, each about the maximum our crane can hoist. The whole ship heeled over when the weight was swung over the side!

The oceanographers are using this quiet time to tidy up and analyse the huge amount of data we have collected so far. My lab alone collects 35.5 million numbers every day, and we have been going for 56 days! That is way too much information for any human brain to process unaided, so I have to use a computer to do a lot of the initial sorting and summarising.

I need to deal with five different types of data files, coming from three different data-capturing computers. In some cases, there is only one file of a given type per day (that is why I go to the lab at midnight - to change the file name). More often there have been interruptions during the day, which means that the instruments were restarted and there are up to five files of a given type in the day.

The first priority is keeping track of all that data. 'Data about data' is called metadata, and it is essential to make sure that the metadata is correct and easy to understand. You can't just call each file any old name, and hope to remember which was which the next day! My first line of defence against confusion is the written logbook in the lab - a bound volume on acid-free paper (so that it will last), written in ink that will not run if it gets wet. Every important event gets recorded there as it happens, along with the names of all the files that are created. I transfer that logbook into an electronic spreadsheet every day, and back-up the files from the original computers where they were collected onto a flash drive.

The second line of defence is following a set of rules, called file conventions, when naming the files. Each has a unique name that identifies what it contains and when it started. That is not enough, because filenames are easily changed and are usually too short to adequately describe the contents. So the first three lines of each file are devoted to a full description: line one says, for instance 'Agulhas 148 Buoy Run northbound mass spectrometer RJ Scholes CSIR 4 February 2010'; line two lists the variables in the file; and line three has the units of measurement. Although a lot of the

data sorting is done using a spreadsheet, the main files are not stored in spreadsheet format, because those have a nasty habit of changing every few years and being incompatible with one another. So I store them as 'flat ascii' files, which you can think of as the 'plain vanilla flavour' of data storage.

The first data clean-up job is to stitch the files of a given type together, end-to-end in the right order, and patch any missing bits as best as possible - either with '-9999', which means 'missing data', or with fixes from some other source. I often make a quick graph to look for obvious mistakes, and either fix them or make a note of them. I don't try to be too picky at this stage, because you don't want to throw out interesting results just because they look funny!

The next job is to merge the files from different sources together. That is easier said than done, because each data set is collected with a different time-step. Rather like two different armies trying to join up, but one is marching fast, and the other slow! I have to slow them all down to the same pace, then get them to fall into line row by row. The second problem is that the three computers are not necessarily following the same clock – to continue my marching analogy, the soldiers arrive at different times. So I have to get some to mark time until the others arrive. Then I have to make clear notes about what I did, so that if I made a mistake, someone can fix it up later - the original data is never altered or deleted. Finally, I have one nice, tidy, well-documented dataset, and the real analysis can begin!

But now I am exhausted, so I will tell you about that some other time. And that is just my data! All the other team members have to do the same thing, and it needs to be finished before we get back to Cape Town.

Love,

Dad