File 20101209.1110: Notes from Probabilistic Redaction meeting this morning. Meeting was in Larry Brown's office; in attendence were Joe Loughry, Jeff Dutoit, Larry Brown, Tom Marso and Kevin R. Miller.

Jeff began the meeting by relating that we have $500,000 to spend in Phase III, and only nine months to do it. Jeff saw a demo of NetBeans last week from David Neal and feels that Mr Neal will be perfect to do the GUI work. Requirements are due in March. Kevin suggested to consider Brandon Henderson. What other tasks are there that need doing? Tom Marso suggested a SOA interface. I suggested the MSCG; Jeff countered that it is not core functionality, but would be a nice-to-have. Jeff felt that the MSCG could almost drop out of the setup processing that Tom Marso's system already does; Tom concurred in principle: that is a good chunk of it. I recommended dropping the Hadoop cluster unless another project in TSS could use it. We just do not have enough time right now, and Tom concurs that the tool is working sufficiently well with the existing back end. If we need more performance later, then implement it. Right now, the performance is acceptable. There was some discussion of performance, Kevin asking if the current hardware is large enough; I explained that at first we thought it would be memory-bound, but it developed that while it uses more disk space than we thought, performance is actually CPU-bound. Tom noted that the Perl we are using now ties up only one core anyway when it runs; I suggested that he might get multicore balancing for free just by installing a new version of Perl compiled for 64-bits; Kevin pointed out that someone (Brandon?) was working on multi-threaded Perl programmes and might be able to help out. Tom was much interested in that thought. Hadoop is Phase IV now.

I gave the thesaurus to Tom. He seemed to like it an felt it was in a good format to use. I suggested applying the cosine correlator to a rapid high-water-mark classification determination on input; Jeff asked if I could do that, but I demurred, saying I was rusty, and recommended Hawk. Jeff seemed okay with that idea. I will be responsible for all the documentation. Tom suggested facetiously to use Wikileaks for test data; I mused that seriously we ought to look at it, because the Wikileaks data dump was redacted before it was released, and it should be possible, even easy, to get hold of both the original and redacted versions and use them as test data. Everyone thought that was a very bad idea, that it would be hard to get approval, and Tom said he would not touch it without signatures from Security. Jeff and Kevin came around eventually and asked me to contact Lyle Wilson to solicit approval to get and use it. I said I would in my adopted rôle of trouble-maker.

Jeff said that one of the things he brought back from the AFRL meeting a few weeks ago is that AFRL have tried this several times before, and failed each time. If we can do it well, there would be important work for us in future. I asked about other file types, including text and PDF and Word documents. The concensus was that we should concentrate on text only, and get really good at it. Concentrate on our core strength functionality. Tom Marso agreed with this, as did Jeff, Kevin and Larry. Tom reported that the speed of the correlator is about half a day to ingest the full corpus and get ready to run, then a few minutes to redact a reasonable sized text document (a few hundred kilobytes), with subsequent documents processing faster. Most of the time is taken in initialisation. The tool is fast enough for now.

The following people will do the following tasks:

- Joe: all documentation
- Hawk: integrate the cosine correlator
- Tom: tech stuff
- Brandon: will help
- David Neal: GUI

We need a test asset (Scott?) in the July or August timeframe. '5.01 better be ready by then', said Jeff.

I will ask Lyle Wilson who to talk to to get permission to use Wikileaks as test data for the probabilistic redaction tool.

What keeps Tom Marso awake at night is the presentation given by the other redaction guys at AFRL a few weeks ago. In their presentation, the corpus was a single bullet point. They are doing a lot more algorithmic analysis. It was a classified presentation, although very little of it was classified. Jeff noted that another redaction tool presented was little more than a GUI; it did not do any redaction, as far as he could tell.

The UCDMO, NSA, DODIIS and AFRL meetings will continue, although we will not likely be invited. We were invited to the last one (Jeff and Tom attended) only because Probabilistic Redaction was on the agenda. Boyd Fletcher is now at NSA and he likes RM. He praised RM highly at that meeting, surprisingly. NSA likes RM; as Larry put it, Radiant Mercury is now a generic term for CDS; lots of people say they want a Radiant Mercury, and then they ask what kind. They say RM when they mean CDS.

References