Reproducible Research: Experience of the MADAGASCAR Open-Source Project

Sergey Fomel

Jackson School of Geosciences
The University of Texas at Austin

March 23, 2011

Second Paper in Geophysics

ON SEISMIC COMPUTATIONS, WITH APPLICATIONS, I[1]

M. M. SLOTNICK[2]

INTRODUCTION

The problem of depth determinations from data obtained by the method commonly called "reflection shooting" is an interesting one and one which, very often, assumes an importance that justifies a good deal of study. It is the purpose of this paper to indicate a method

[1] Published by permission of the Board of Directors, Humble Oil and Refining Company.

[2] Mathematician, Geophysics Department, Humble Oil and Refining Company.

9

VOLUME 1 JANUARY, 1936 NUMBER 1

GEOPHYSICS

A Journal of General and Applied Geophysics

BLACK MAGIC IN GEOPHYSICAL PROSPECTING[1]

L. W. BLAU[2]

EDITOR'S NOTE: *The term "doodle-bug" is coming more and more to mean proposed methods of geophysical prospecting that are neither based upon scientific fact nor upon known or proven properties of oil, minerals and geologic formations. The geophysicist is often consulted concerning the reliability of such a proposed method, and his task then is to explain scientifically just why the proposed method fails and is unsuitable for the intended purpose.*

What is Science?

What is Science?

Science *is the systematic enterprise of gathering knowledge about the universe and organizing and condensing that knowledge into testable laws and theories. The success and credibility of science are anchored in the willingness of scientists to* ***independent testing and replication*** *by other scientists. This requires the* ***complete and open exchange of data, procedures and materials****.*
American Physical Society, What is Science?

Communicating to a Skeptic

DIALOGO
DI
GALILEO GALILEI LINCEO
MATEMATICO SOPRAORDINARIO
DELLO STVDIO DI PISA.
E Filosofo, e Matematico primario del
SERENISSIMO
GR. DVCA DI TOSCANA.
Doue ne i congressi di quattro giornate si discorre
sopra i due
MASSIMI SISTEMI DEL MONDO
TOLEMAICO, E COPERNICANO;
Proponendo indeterminatamente le ragioni Filosofiche, e Naturali
tanto per l'una, quanto per l'altra parte.

CON PRI VILEGI.

IN FIORENZA, Per Gio: Batista Landini MDCXXXII.
CON LICENZA DE' SVPERIORI.

From Science to Open-Source Software

Abandoning the habit of secrecy ***in favor of process transparency and peer review was the crucial step by which alchemy became chemistry. In the same way, it is beginning to appear that open-source development may signal the long-awaited maturation of software development as a discipline.***

Eric Raymond, TAUP, 2004

What is Reproducible Research?

- Attaching software code and data to publications
- Communicating computational results to a skeptic

An article about computational science in a scientific publication is not the scholarship itself, it is merely advertising of the scholarship. The actual scholarship is the complete software development environment and the complete set of instructions which generated the figures. J. Buckheit and D. Donoho, WaveLab

Reproducible Research Discussions

ICASSP 2007

Berlin-6 2008

CiSE 2009

- Donoho et al.
- LeVeque
- Ping & Eckel
- Stodden

IEEE Signal Processing Magazine 2009

- Vandewalle et al.

Yale Roundtable 2009

NSF Archive Workshop 2010

- http://www.reproducibleresearch.net

Reproducible Research Discussions

AAAS 2011

- The Digitization of Science: Reproducibility and Interdisciplinary Knowledge Transfer

SIAM CS&E 2011

- Verifiable, Reproducible Research and Computational Science

SIAM GS 2011

- This minisymposium

AMP 2011

- Reproducible Research: Tools and Strategies for Scientific Computing

ICIAM 2011

- Reproducible Research in Computational Science: What, Why and How

Outline

Reproducible Research

History of Madagascar

Jon Claerbout's Story

1987 Sunview experience
- Interactive programs are slavery

1992 LaTeX + `cake`
- Building books by a single command

1990s Ph.D. students
- `cake` to `make`, CD-Rom to WWW

2001 Reproducible research paper in *CiSE*
- The principal beneficiary is the author

Lesson 1

The principal beneficiary is the author.

http://reproducibility.org/
http://ahay.org/

Ohloh.net about MADAGASCAR

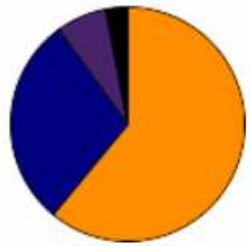
Ohloh Analysis Summary

- Mostly written in C
- Mature, well-established codebase
- Increasing year-over-year development activity
- Large, active development team

Updated 03 Mar 2011 15:52 UTC

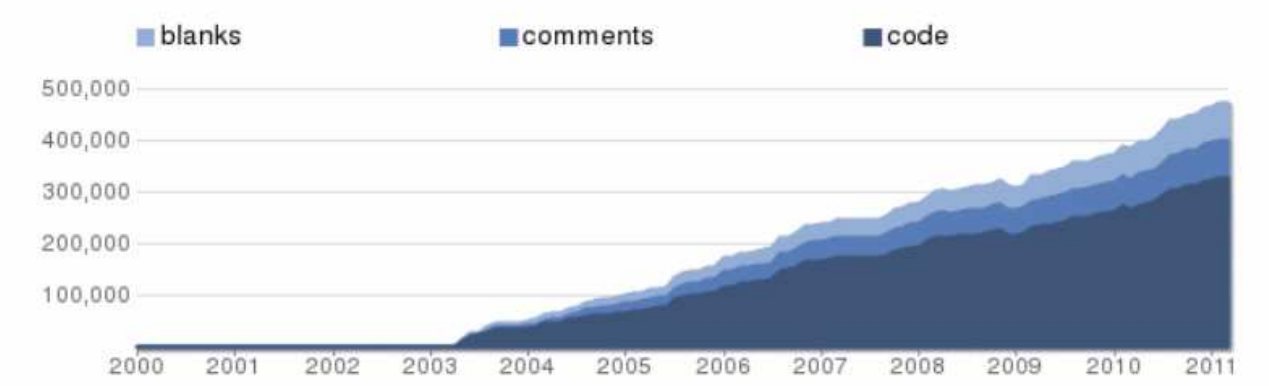
GNU General Public License 2	1482 files
New BSD License	13 files
GNU Library or "Lesser" GPL (LGPL)	3 files

C	61%
TeX/LaTeX	29%
Python	7%
Other	3%



Language	Code Lines	Comment Lines
C	199,655	51,414
TeX/LaTeX	94,806	8,573
Python	21,907	10,744
Fortran (Fixed-format)	4,002	8
C++	2,996	688
Matlab	1,366	709
Make	740	13
Java	704	31
Fortran (Free-format)	370	41

Lines of Code



Lesson 2

- http://www.ahay.org/wiki/Reproducible_Documents

Each computation is a test.

Thanks

- Tariq Alkhalifah, Vladimir Bashkardin, Jules Browaeys, William Burnett, Cody Brown, Maria Cameron, Lorenzo Casasanta, Joseph Dellinger, Jeff Godwin, Gilles Hennenfent, Trevor Irons, Jim Jennings, Long Jin, Roman Kazinnik, Siwei Li, Guochang Liu, Yang Liu, Doug McCowan, Henryk Modzelewski, Colin Russell, Paul Sava, Jeffrey Shragge, Xiaolei Song, Eduardo Filpo Silva, Ioan Vlad, Jia Yan, Lexing Ying

School and Workshop: Vancouver 2006

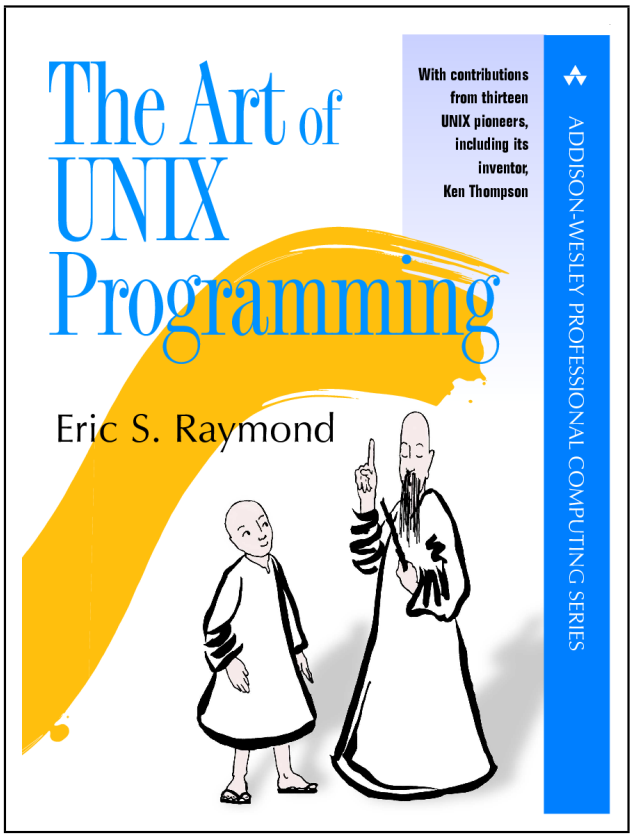
School and Workshop: Houston 2010

Lessons 3 and 4

Reproducibility requires maintenance.

Maintenance requires an open community.

MADAGASCAR Design



- **Multidimensional arrays as files**

Write programs that do one thing and do it well. Write programs to work together. Write programs to handle text streams, because that is a universal interface.
Doug McIlroy, Unix

MADAGASCAR filter in C

```c
#include <rsf.h>

int main(int argc, char* argv[])
{
    int n1, n2, i1, i2;
    float clip, *trace;
    sf_file in, out;

    sf_init(argc,argv);
    in  = sf_input("in");
    out = sf_output("out");

    sf_histint(in,"n1",&n1); /* trace length */
    n2 = sf_leftsize(in,1);  /* number of traces */
    if (!sf_getfloat("clip",&clip)) sf_error("Need clip=");

    trace = sf_floatalloc (n1);
    for (i2=0; i2 < n2; i2++) {
        sf_floatread(trace,n1,in);
        for (i1=0; i1 < n1; i1++) {
            if      (trace[i1] >  clip) trace[i1]= clip;
            else if (trace[i1] < -clip) trace[i1]=-clip;
        }
        sf_floatwrite(trace,n1,out);
    }

    exit(0);
}
```

MADAGASCAR filter in Python

```
#!/usr/bin/env python

import numpy
import m8r

par = m8r.Par()
input  = m8r.Input()
output = m8r.Output()

n1 = input.int("n1") # trace length
n2 = input.size(1)   # number of traces

clip = par.float("clip")

trace = numpy.zeros(n1,'f')
for i2 in xrange(n2): # loop over traces
    input.read(trace)
    trace = numpy.clip(trace,-clip,clip)
    output.write(trace)
```

MADAGASCAR SConstruct script

```
from rsf.proj import Flow

Flow('spike',None,'spike n1=1000 n2=100 | bandpass fhi=10')
Flow('cliped','spike','clip clip=0.5')
```

```
bash$ scons
scons: Building targets ...
/usr/bin/sfspike n1=1000 n2=100 | /usr/bin/sfbandpass fhi=10 > spike.rsf
< spike.rsf /usr/bin/sfclip clip=0.5 > cliped.rsf
scons: Done building targets.
bash$ sed -i'' -e's/0.5/0.25/' SConstruct
bash$ scons -Q
< spike.rsf /usr/bin/sfclip clip=0.25 > cliped.rsf
```

- http://www.scons.org/

- **Reproducible research**
  - Attaching software and data to publications
  - Computational experiments communicated to a skeptic
- **MADAGASCAR Lessons**
  1. The principal beneficiary is the author.
  2. Each computation is a test.
  3. Reproducibility requires maintenance.
  4. Maintenance requires an open community.