Madagascar Open-Source Software Project

School on Reproducible Computational Geophysics
Austin, Texas, July 20-21, 2012

Sergey Fomel
Jackson School of Geosciences
The University of Texas at Austin
Software as a Research Tool

Connections

Power
## Friday, July 20

<table>
<thead>
<tr>
<th>Time</th>
<th>Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>9:00 – 10:15</td>
<td><strong>Introduction</strong></td>
</tr>
<tr>
<td>10:30 – 12:00</td>
<td>Introductory exercise</td>
</tr>
<tr>
<td>10:30 – 12:00</td>
<td>Lunch @ Commons</td>
</tr>
<tr>
<td>1:00 – 2:30</td>
<td><strong>Fundamentals</strong></td>
</tr>
<tr>
<td>2:45 – 3:30</td>
<td>Vplot graphics</td>
</tr>
<tr>
<td>3:30 – 5:00</td>
<td>Field data processing</td>
</tr>
<tr>
<td>3:30 – 5:00</td>
<td>Dinner @ Maggiano’s</td>
</tr>
</tbody>
</table>
### Saturday, July 20

<table>
<thead>
<tr>
<th>Time</th>
<th>Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>9:00 – 12:00</td>
<td>Wavefield Imaging</td>
</tr>
<tr>
<td>1:00 – 1:45</td>
<td>Lunch @ ROC</td>
</tr>
<tr>
<td>1:00 – 1:45</td>
<td>Programming/C++</td>
</tr>
<tr>
<td>1:45 – 2:30</td>
<td>Python/GUI</td>
</tr>
<tr>
<td>2:45 – 3:30</td>
<td>Parallel computing</td>
</tr>
<tr>
<td>3:30 – 4:15</td>
<td>Contributing</td>
</tr>
<tr>
<td>4:15 – 5:00</td>
<td>General discussion</td>
</tr>
</tbody>
</table>
Obtaining Madagascar

- Download a stable version from sourceforge.net
  - currently madagascar-1.3

- Check out a current development version using Subversion
  - svn checkout
  - svn update
  - svn commit *(for developers)*

Installing from Source

$ ./configure --prefix=/dir/name
$ make
$ make install
$ source /dir/name/share/madagascar/etc/env.sh

http://www.ahay.org/wiki/Installation
http://www.ahay.org/wiki/Advanced_Installation
Outline

- Reproducible research
- History of the project
- Look inside Madagascar
Science or Black Magic?

Volume 1  JANUARY, 1936  Number 1

GEOPHYSICS

A Journal of General and Applied Geophysics

BLACK MAGIC IN GEOPHYSICAL PROSPECTING

L. W. BLAU

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.
SEG Mission

Article I. Name
The Society shall be called the “Society of Exploration Geophysicists.”

Article II. Objectives
The objectives of this Society shall be to promote the science of geophysics, especially as it relates to exploration and research, to foster the common scientific interests of geophysicists, and to maintain a high professional standing among its members.

SEG Constitution
Duke Cancer Scandal

Retracted

The Economist

Misconduct in science
An array of errors

Investigations into a case of alleged scientific misconduct have revealed numerous holes in the oversight of science and scientific publishing

Sep 10th 2011 | from the print edition
Human Language Technology in the US

Mark Liberman
Human Language Technology in the US

Stopped in 1966-1969 by John R. Pierce, the director of acoustics research at Bell Labs
Pierce Report

J. R. Pierce, 1969, Whither Speech Recognition?, Letter to the Editor of JASA.

“We are safe in asserting that speech recognition is attractive to money. The attraction is perhaps similar to the attraction of schemes for turning water into gasoline, extracting gold from the sea, curing cancer, or going to the moon. One doesn't attract thoughtlessly given dollars by means of schemes for cutting the cost of soap by 10%. To sell suckers, one uses deceit and offers glamor.

It is clear that glamor and any deceit in the field of speech recognition blind the takers of funds as much as they blind the givers of funds. Thus, we may pity workers whom we cannot respect.”
“Most recognizers behave, not like scientists, but like mad inventors or untrustworthy engineers. The typical recognizer gets it into his head that he can solve "the problem." The basis for this is either individual inspiration (the "mad inventor" source of knowledge) or acceptance of untested rules, schemes, or information (the untrustworthy engineer approach).

The typical recognizer ... builds or programs an elaborate system that either does very little or flops in an obscure way. A lot of money and time are spent. No simple, clear, sure knowledge is gained. The work has been an experience, not an experiment."
Human Language Technology in the US

- Stopped in 1966-1969 by John Pierce, the director of acoustics research at Bell Labs
- Restarted in 1986 by Charles Wayne, DARPA program manager

- Protect against “glamour and deceit”
  - A limited and objective evaluation metric
  - Applied by a neutral agent (NIST)
- Ensure that “simple, clear, sure knowledge is gained”
  - Participants must reveal their methods
  - To the sponsor and to one another
  - At the same time that the evaluation metric is applied.
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 expose their ideas and results to independent testing and replication by other scientists. This requires the complete and open exchange of data, procedures and materials.
Fig. 1. The seismic experiment, conducted over a simplified subsurface with one dipping reflector. Applying the law of cosines to triangle $s's'rs$, one may express the travel time $t$ from source $s$ to receiver $r$ in terms of zero-offset time $t_0$, half-offset $h$, velocity $v$, and dip $\theta$. The result is equation (3) in the text, the

Defining

$$A \equiv \frac{dt_n}{dt_0} = \frac{t_0}{t_n} = \left[ 1 + \left( \frac{\Delta t_0}{\Delta y} \frac{h^2}{t_n^2} \right)^{1.2} \right],$$

and using equation (10) to replace $p_n(\sqrt{t_n^2 + (\Delta t_0/\Delta y)^2 h^2}, y, h) = p_n(t_n, y, h)$, the Fourier transform becomes

$$P_0(\omega_0, k, h) = \int dt_n A^{-1} e^{i\omega_0 t_n} \int dy e^{-iky} p_n(t_n, y, h). \quad (12a)$$
(Hale, 1984)

(a) CMP STACK with DIP-MOVEOUT

(b) CMP STACK without DIP-MOVEOUT
“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.”

(Buckheit and Donoho, 1995)
“It is a big chore for one researcher to reproduce the analysis and computational results of another [...] I discovered that this problem has a simple technological solution: illustrations (figures) in a technical document are made by programs and command scripts that along with required data should be linked to the document itself [...] This is hardly any extra work for the author, but it makes the document much more valuable to readers who possess the document in electronic form because they are able to track down the computations that lead to the illustrations.” (Claerbout, 1991)
Reproducible Research

Addressing the Need for Data and Code Sharing in Computational Science

By the Yale Law School Roundtable on Data and Code Sharing
Reproducible Research in 2011-2012

- The Digitization of Science: Reproducibility and Interdisciplinary Knowledge Transfer at AAAS
- Verifiable, Reproducible Research and Computational Science at SIAM CS&E
- Reproducible Science and Open-Source Software in the Geosciences at SIAM Geosciences
- Reproducible Research at Interface
- Reproducible Research: Tools and Strategies for Scientific Computing at AMP
- Reproducible Research in Computational Science: What, Why, and How at ICIAM
- Publishing Reproducible Science and Software at SIAM Imaging Science
Reproducible Research: Madagascar Lessons

- Reproducibility is not the goal
- The principal beneficiary is the author
- Each computation is a test
- Reproducibility requires maintenance
- Maintenance requires an open community
"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
Outline

- Reproducible research
- History of the project
- Look inside Madagascar
“Run a program as you wish, for any purpose you wish, not limited to any narrowly defined application.”

“Help yourself by improving the program (which requires access to source code).”

“Help your neighbor by sharing a copy of the program with them.”

“Help community by sharing the improved copy at large.”
“Given a large enough beta-tester and co-developer base, almost every problem will be characterized quickly and the fix obvious to someone.”

"Given enough eyeballs, all bugs are shallow."
Free/Open-Source Software Examples

- Linux kernel
  - GPL, 1991, Unix
- GIMP
  - GPL, 1996, Photoshop
- R project
  - GPL, 1993, S
- Sage
  - GPL, 2005, MATLAB/Mathematica
- Madagascar
  - GPL, 2006
In a Nutshell, Madagascar...

... has had **7,697 commits** made by **57 contributors** representing **451,655 lines of code**

... is **mostly written in C** with **an average number of source code comments**

... has **a well established, mature codebase** maintained by **a large development team** with **increasing year-over-year commits**

... took an estimated **120 years of effort** starting with its **first commit in May, 2003** ending with its **most recent commit 3 days ago**
Languages

- C: 63%
- TeX/LaTeX: 25%
- Python: 5%
- 13 Other: 7%

Lines of Code

- 1000k
- 500k
- 0k

Year:
- 2004
- 2005
- 2006
- 2007
- 2008
- 2009
- 2010
- 2011
- 2012
Contributors

Vancouver-2006

photo by Joe Dellinger
Beijing-2011
Web Traffic Sources

- google.com
- direct
- sourceforge.net
- wikipedia.org
- stanford.edu
- seg.org
- other
Website Visits from USA
Stable Versions

- **0.9** – released in June 2006 (Vienna)
- **1.0** - released in July 2010 (Houston)
- **1.3** – released in April 2012
- **1.4** – preparing for release

**Goals for version 2.0**

- Seismic field data processing examples
- Large-scale parallel computing examples
- Non-seismic applications
Outline

- Reproducible research
- History of the project
- Look inside Madagascar
Madagascar Pyramid

- Implement
- Test
- Publish
Madagascar Pyramid

- Papers
- Examples
- Programs
Data arrays as file objects on disk

“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
Objectives

- To make computational research efficient
- To make it easy to share computational results
- To promote an open community