Literate programming and reproducible research

Darlene R. Goldstein

Institut de mathématiques, EPFL

30 Sept 2010

The reproducible research principle

- wavelet community, Stanford University
- Buckheit and Donoho: When we publish articles containing figures which were generated by computer, we also publish the *complete software environment* which generates the figures.

Wavelab anecdotes

- Final versions of figs for publication
- Lost or stolen work
- Communication
- Applying old methods on new data
- Reconstructing work of others

Claerbout slogan

- 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.

Research practice

- Discipline in software building
- From the start, *expect* it to be made available to others as part of the publication of their work
- (Also think in terms of program re-use)

But does it go far enough?

- Jan de Leeuw says no!
- No reason to single out figures -principle should equally apply to tables, SEs, etc. any form of computer-generated output
- No reason to limit to published articles can apply to teaching, lectures
- Should not violate the freeware principle

Management

- File management
  - terminal subdirs have all the files associated with a project or sub-project
- Script management and documentation
  - version control system
- Reproducible research
  - make, perl
  - literate programming practice

Literate Programming (Knuth)

- Combining the use of a text formatting language (such as TeX) and a conventional programming language (like C or R) so as to maintain documentation and source code together, the art of writing computer programs for the human reader
- may use *inverse comment convention*
- A kind of literate programming where the program code is marked to distinguish it from the text, rather than the other way around as in normal programs

WEB (not www)

- WEB (Donald Knuth), noweb (Norman Ramsey)
- a WEB system consists of two processors, called *WEAVE* and *TANGLE*
  - WEAVE "weaves" the document for a human reader, producing TeX output
  - TANGLE "tangles" the document for a computer, producing a plain programming language file to be compiled, linked and executed
- WEB (and variants) are not the only environments for Literate Programming

Tools

- Compendium concept
  - dynamic document
  - data
  - auxiliary software
- Tools for use with R
  - ESS (Emacs Speaks Statistics)
  - Sweave
  - LaTeX