

Why open source?

Several research colleagues have criticized me for what they call my mania, that I only use open-source software and nothing else. They claim it's just to buck the tide. After all, why not use what everybody else uses? If they all prefer Windows, why should I insist on software as abstruse as Linux?

For starters, let's revisit the meaning of "free". Free software means freedom to: (1) run the program for any purpose; (2) study how the program works, and change it to make it do what you wish; (3) redistribute copies so you can help your neighbor; and (4) improve the program, and release your improvements (and modified versions in general) to the public, so that the whole community benefits (http://en.wikipedia.org/wiki/Free_software). The definition of open source is more pragmatic, but it guarantees that the software's source code is readable and intelligible to any programmer (http://pt.wikipedia.org/wiki/Código_aberto). In practice, both terms – free and open – mean that dozens and even hundreds of programmers can participate in developing the software.

But what's the advantage? Why is this so important? After all, I can sum up everything I know about programming in two words: almost nothing. So what's the reason for this nearly ideological battle? First, because science is built collectively. Hiding the code of a statistical package, for example, prevents peer validation, the most important validation of science. To explain better, we can even compare research results in two different software packages, but to study why they produced different results (and thus improve the techniques) is impossible in a proprietary environment. The reason is obvious: the software was developed for the company to make a profit, not for the advancement of science. That's why I prefer the definition "free software".

The essential idea of free software is the same as that used by CSP for copyright, and is discussed in the editorial *Free Access* ¹: CC-BY-NC (<http://creativecommons.org/licenses/by-nc/4.0/deed.en>), which allows others to "*remix, transform, and build upon the material*". This proposal is brilliantly described in the article on how RecLink ² migrated to Open-RecLink. Right from the start the authors intended to develop a "Free and Open-Source Software" – FOSS (http://en.wikipedia.org/wiki/Free_and_open-source_software), more in keeping with the scientific ethos, but for various reasons they were unable to keep the code open, which is now possible. And why this huge effort? Because of RecLink's intrinsic proposal, as a software package widely used in probabilistic record linkage between secondary databases, the potential of which is achieved by expanding the community of users and new developers.

Developing software this way – freely and cooperatively – is sure to succeed when it involves participation by the user community. And the quality is especially guaranteed. Of all the examples, R is the one that most amazes me as a user. From a project that started in 1993 with two researchers and just 1,000 lines of code, R has now become a consistent statistical software environment with 6,216 additional libraries and more than 150 published books. The proposal's vitality is obvious. Regardless of individual taste, broad participation by the community of statisticians and users in general is what ensures the project's quality and scope.

We at CSP prefer everything that's free: from access to the articles themselves to articles that provide free access to the questionnaires ³ and codes they use. We also plan to start discussing a topic that has begun to take shape in scientific publications: free access to

the data. In practice, we use Linux in most of our computers and accept articles in formats generated with free software. That's why I'm so proud that this issue of CSP is publishing the article *Going Open Source: Some Lessons Learned from the Development of OpenReclink*. We wish OpenReclink a huge success, and within a few years, may the authors celebrate the tenth version with participation by many, many collaborators. Health research thanks you!

Marilia Sá Carvalho

Editor

-
1. Carvalho MS, Travassos C, Coeli CM. Acesso livre. *Cad Saúde Pública* 2013; 29:213-5.
 2. Camargo Jr. KR, Coeli CM. *Reclink*: aplicativo para o relacionamento de bases de dados, implementando o método *probalistic record linkage*. *Cad Saúde Pública* 2000; 16:439-47.
 3. Carvalho MS, Travassos C, Coeli CM, Reichenheim ME. Um passo à frente na política de acesso aberto de CSP: instrumentos de aferição. *Cad Saúde Pública* 2014; 30:1357-9.