Open Source Software for Public Health Epidemiology

From: Andrew G. Dean, MD, MPH
agdean9@hotmail.com

Originally Submitted: November 1, 2002
Revisions incorporated January 2003 after decision to proceed with JavaScript version of "Statcalc"

Summary

Epi Info, CDC's public-domain program for public health, has followed or pioneered many of the principles of the Open Source software movement since 1984. Now that Open Source software development methods have been recognized as productive and useful, it is proposed that selected programs or modules be developed with conscious attention to Open Source methods, using a website external to CDC to facilitate communication.

An open framework for statistical calculation, a Windows and Web version of the DOS program Statcalc, will be the first Open Source project, to be known as OpenEpi. It is intended to extend the statistical capabilities of Epi Info to include many procedures currently offered on the web and to offer a uniform development platform for new statistical programs that can be used either on the Web or from the desktop.

Background

Both the DOS version of Epi Info (Epi 6) and the Windows Version (Epi 2002) have many characteristics of Open Source software.

Epi Info Development Open Source Concepts

Open Source Concept	Epi Info development	Shortfalls
Source code available	Source code provided to anyone making a request, but only a dozen or two requests received over the years	Embedded commercial library modules require separate license. Source not available on Web
Programs available without charge or for minimal copying charges	Installable programs available on CDC web site	None
Users participate in programming	Translators recompiled and altered code for Epi 6	With occasional exceptions, most programming is done by the core development team.
Users participate in testing and debugging	Hundreds of users participated in beta tests	Lack of staff limits 2-way communication with

		testers, but this has generally been a strong feature of Epi Info development.
Frequent releases provide common access to fixes and improvements	Patches have been released every few months as final production releases	In beta testing, source is not available, and internal team has versions not available on beta ftp site
Development includes a worldwide group of user/developers	A very strong feature of Epi Info	More systematic support of translators could be done
Two-way communication among user/developers	One of the reasons for Epi Info's success	Communication has varied with resources available. Support and communication staff have been cut to the point where "keeping up" and critical mass are in question.
Product incorporates user ideas and responds to needs	A strong area in general	Programmer's are allowed to "own" source code and are seldom challenged by other programmers on claims that xyz feature is "technically" impossible, dangerous, or expensive. Decisions are made by vote of a programming team not always in touch with user needs.
Users feel a sense of ownership and participation	Generally a strong feature.	Government environment may inhibit free discussion, particularly of resource issues
User support is provided	Yes, by CDC, and, to some extent, by a ListSurv group	ListSurv mechanism has deteriorated, now requires a password, loses settings, and could benefit from more intensive care and feeding.
Both released and beta versions available on web site	Released versions on website. Beta via ftp for selected group with passwords. Beta source	Beta and source versions could be publicly available with careful

	code not available.	labeling

Statcalc, a Flexible Statistical Framework

The DOS version of Statcalc, a statistical calculator, is still distributed with Epi Info for Windows. Aggregate numbers are entered into a table on the screen and the program produces epidemiologic statistics similar to those offered by the Analysis program for Microsoft Access and other tabular data.

One reason that Statcalc has not been rewritten for Windows is that there are thousands of free calculators on the Web, and a Google search for "calculator statistics epidemiology" turns up 2280 links. It is clear that "Yet another calculator" is not an emergency requirement. A framework for unifying various calculators, or for entering data that could be used in many, or for allowing many participants to write new statistics, could well be useful, however.

It is proposed that the new Statcalc be developed as an Open Source project under the general name of OpenEpi. At least two versions for Windows have been developed as partially finished Visual Basic programs by Vic Sahai, MSc, of Northern Health Information Partnership in Ontario, and by Dr. Hemant Kulkarni of Nagpur, India, and San Antonio, Texas, and Andrew Dean, author of this proposal. Kevin Sullivan, Dept. of Epidemiology, Rollins School of Public Health, Emory University has developed prototype JavaScript modules for simple proportions and 2 x 2 tables.

Proposed Method of Open Source Development

It is proposed that the core group combine the best features of previous efforts and generalize the program as much as possible so that others can add features. It would then be released as beta Open Source, using a Website to be constructed under this grant.

A website called www.openepi.com will be developed and maintained under this proposal. It will provide facilities for discussion of Open Source epidemiologic software projects, for downloading of final and provisional source code and for participation in analysis, design, programming, testing, and evaluation. It will also encourage participants to post needs for new software and ideas for improvement of existing programs. Another section of the site will provide information on current Open Source software for public health epidemiology, including statistical functions available on the Web.

The website will be open to anyone who wants to participate, with a minimum number of procedural barriers to its use, and, for most purposes, no passwords. It will be organized

as a convenience and communication mechanism for those developing Open Source software for public health.

The site will be located on a commercial web hosting company, www.brinkster.com. The first year will be considered a test, after which the project will be evaluated for productivity and a proposal for continuation prepared.

The author of this proposal will manage the design and development of the site and participate actively on a daily basis to make it a success, acting as Master of Ceremonies for the site until other suitable voices emerge. He served for 16 years as chief of the Epi Info development team. As a State Epidemiologist, he supervised and edited monthly newsletters for 8 years. In 1998, he served as Acting Editor of CDC's Morbidity and Mortality Weekly Report (MMWR) for 4 months. He has skills in epidemiology, software development, programming, and communication.

Risks

Collaboration with the Epi Info development team at CDC will be required. If other programs, such as Epi Info 6, are to be included in the open source effort, commercial software sources of libraries incorporated in Epi 6 would be asked to give permission for selected modules of source code to be distributed without royalties, as these products of the previous decade are difficult even to purchase. It is likely that most of this code could be replaced by an Open Source effort, but doing so might slow development. Very little or none of the statistical code in Epi Info is commercial in origin.

Since the first version of this proposal was submitted, subsequent discussions have determined that the first version of OpenEpi would be entirely in HTML and JavaScript, so that it can be downloaded and run on a local computer as well as from a server. Since JavaScript, for security reasons, is limited in its ability to save data to disk, programming in other languages such as Java will be required later.

Epidemiologists are usually not also computer programmers, although an increasing number have such skills and interests. Demand for Epi Info's source code has been only occasional in the past, and then usually by translators. Making source easily available should attract additional participants, such as classes of computer science students, but the extent to which this is true remains to be demonstrated.

Timeline

Immediately

Proceed with development of OpenEpi, using email and occasional telephone conversations.

By February 1, 2003

Implement the OpenEpi website with preliminary ideas and prototypes for comment, including a configurable menu in JavaScript (adapted from available Web Open Source menus).

Features for storing and downloading source code, and for interaction among developers will added during the first three months. The Statcalc project will be used to guide the development of resources for other Open Source development if appropriate.

By April 1, 2003

Develop a Data Entry module for use by those writing statistical routines.

By June 1, 2003

Develop an Output formatting module so that statistical modules need produce only statistical results.

By July 1, 2003

First statistical modules using the input and output facilities of OpenEpi available. Validation procedures developed and applied.

One year after approval

Records will of site usage, features developed, and software produced will be summarized and the project generally evaluated and presented to interested participants, both online and offline. If the site is considered useful, a second proposal for continuation and enhancement will be prepared.