Distr.
GENERAL

CES/AC.71/2004/18 (Summary)
1 March 2004

ENGLISH
Original: RUSSIAN

UNITED NATIONS STATISTICAL COMMISSION AND ECONOMIC COMMISSION FOR EUROPE

EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)

CONFERENCE OF EUROPEAN STATISTICIANS

ORGANISATION FOR ECONOMIC COOPERATION AND DEVELOPMENT (OECD)
STATISTICS DIRECTORATE

Joint ECE/Eurostat/OECD meeting
on the management of statistical
information systems (MSIS)
(Geneva, 17-19 May 2004)

Topic III: The use of open source software in statistics

Open source software and the management of statistical institute information technologies

Supporting paper

Submitted by the State Committee on Statistics of Azerbaijan*

Summary

* Author: Fayg Zhaliov (ssc@azstat.org)

GE.04-30487 (E) 240304 260304

Introduction

1. The open source software movement is a trend in collective software development whereby programs are distributed free of charge to everyone, everywhere, together with the source code. As distinct from the classic, proprietary development model, open source implies:

- Software creation without a single, detailed, pre-ordained and approved plan;
- Software creation by a virtual community of independent programmers open to all comers;
- Development documentation and program code completely accessible to everyone, everywhere, not just those involved in the development process.

2. The purpose of this paper is not to discuss the open source movement or its psychological, moral, social and commercial aspects. Much has been written about these already, not least by leading figures in the movement such as Richard Stallman, Eric Raymond and Tim O'Reilly.

3. Instead, it will be limited to questions relating to the use of open source software at statistical institutions and the prospects for using the same approach to develop applications for the processing of statistical data.

Experience in the use of open source software

4. Our experience has consisted, first, in the use of system products to construct a fully-fledged corporate network at Goskomstat:

- The Linux operating system. Since 2001 we have migrated completely from Novell OS and Windows NT to Linux. All servers on our corporate network are now running under Linux;
- Internet-Intranet components. These are: a BIND (Berkeley Internet Name Daemon)-based DNS (domain name system) server, a Samba file server, an Apache Web-server integrated with Perl, PHP and MySQL (our site, www.azstat.org, is hosted on Apache), a Sendmail mail server, a Squid proxy server, an iptable netfilter and firewall etc.

5. Second, freely distributed programming languages and MySQL have been used to develop applied client-server statistical routines, which themselves make use of a large number of freely distributed programs and modules. It does not yet make sense to renounce commercial products entirely, of course. We chiefly use resources and program development systems from Microsoft, Borland, Oracle etc. for application development.

Prospects for using the open source method to develop statistical programs

6. Can the open source method be used to tackle problems in the processing of statistics? If so, what sort of problems, and how should one go about it? Statistical institutions do not tend to have many programming staff, and they cannot develop for all the problems encountered. Either the staff must be increased or part of the work must be contracted out. Either option requires significant expenditure and has substantial drawbacks. Open development drawing on a virtual community of like-minded individuals could be a partial or even a complete solution. Let us consider some of the attendant issues:

- Kinds of problem. The problems proposed by statistical institutions for open source development must be sufficiently general-purpose and suggest original, interesting solutions. This is essential in order to enlist a large community of independent programmers in tackling them. It should be borne in mind that programmers will not be joining the project out of altruistic motives: they will be potential users who, like us, have a stake in the ultimate outcome. We cannot count on enlisting a large virtual community if the problem proposed has only a very narrow, technical sphere of application.

- Project management. A project may be announced on a working site, but it cannot be started with just a bald proposal. From the outset it is important to put forward a real program with the source code. It may be at an early stage, it may contain mistakes and function poorly. That does not matter. The idea is the thing. The project must be managed by a highly qualified expert at the institution who has a gift for communicating with the community.

Corporate development of statistics applications

7. This consists in the joint development of applications by a number of organizations with a stake in the outcome. It may be arranged by contract between the organizations concerned. The problems themselves may be tackled in the traditional way or following an open source approach. This method can be used for problems that are not sufficiently general-purpose and apply only in narrow technical fields, and thus cannot be counted on to attract a large community of independent programmers, but are of interest to a number of statistical institutions which can try to tackle them together.

8. Given the limits on the length of this paper, we merely allude here to certain issues and trends. We hope to offer a more thorough discussion soon in a full version of the paper.

9. Open source is an exercise in collective program development whereby programs are distributed to all comers along with the source code. The open source movement did not arise in a vacuum. Early in the development of Unix and the Internet, programmers exchanged their source code. When AT&T was forbidden to market computer technology commercially, Ken Thompson and Dennis Ritchie began to send the source code for Unix to anyone who asked

for it. Thanks to this, thousands of programmers had the opportunity to develop the operating system unhindered, adding their own ideas and routines. Later the development of Unix shifted onto a commercial footing, and this has not been highly beneficial to its subsequent fortunes. Beginning in about the late 1970s, an information, file and news exchange network began to take shape among the independent developers and users of Unix and some other programs: this ultimately evolved into the Internet.

10. The open source software movement is now a widespread phenomenon. It denotes a pattern of development whereby:

- Software is created without a single, detailed, pre-ordained and approved plan;
- Software is created by a virtual community of independent programmers open to all comers;
- Development documentation and program code are completely accessible to everyone, everywhere, not just those involved in the development process.

11. Linux. Built on top of Linus Torvalds' kernel, Linux distributions typically include hundreds of other open-source packages. However, most of those packages also run on other platforms, including just about every dialect of UNIX, Windows and Windows NT, MacOS, and many others.

- FreeBSD, OpenBSD and other Berkeley UNIX derivatives. While Linux gets the lion's share of the attention, BSD UNIX systems also have a significant following.
- Programmer's Tools. The Free Software Foundation's GNU project has created a high-quality set of programmer's utilities, including the gcc C compiler, the g++ C++ compiler, the emacs editor, the gdb debugger. Two other programming tools developed outside the GNU project that are an absolutely indispensable part of the open-source culture are Larry Wall's patch program, which allows developers to exchange small fixes to programs rather than having to ship around the source code for an entire program, and CVS, the Concurrent Versioning System, which allows developers to maintain multiple versions of the source tree.
- Languages. Larry Wall's Perl language is the undisputed king of the open-source programming languages, but John Ousterhout's tcl and Guido van Rossum's Python language also have thriving communities.
- Apache. As noted earlier, the Apache Group has dominant web server market share. The most recent Netcraft Web server survey (www.netcraft.co.uk/survey) shows Apache with 54% of all visible web servers, followed by Microsoft's IIS at 23% and Netscape at 7%. Apache continues to be developed and maintained by a group of about 12 core developers, plus a large and active user community.

- Samba. Developed by a worldwide team headed by Andrew Tridgell in Australia, Samba allows UNIX and Linux systems to act as file and print servers on NT and Windows 95/98 networks. This is a “stealth” technology that has allowed administrators to work Linux into their networks without the knowledge of management.

- Sendmail. Originally developed as part of Berkeley UNIX, sendmail is the backbone of the Internet’s e-mail server infrastructure.

- - - - -