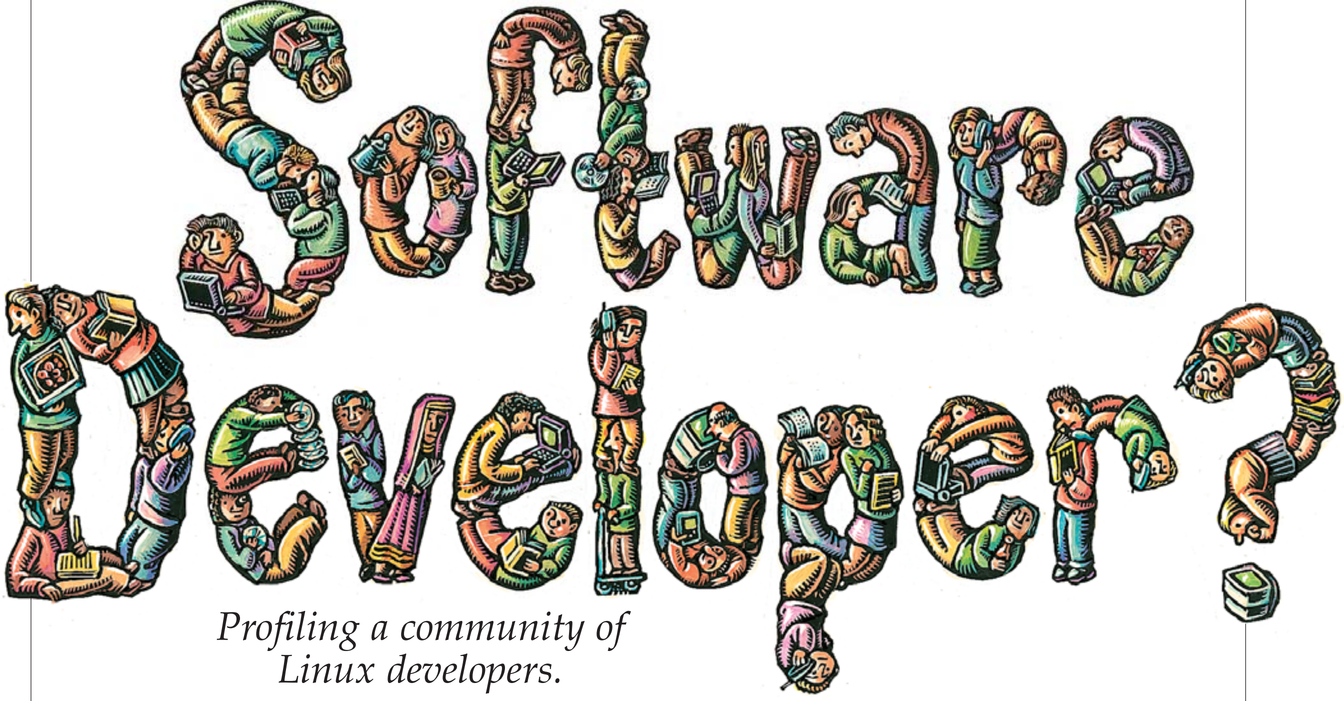
Who Is an Open Source Software Developer?



Profiling a community of Linux developers.

< Bert J. Dempsey, Debra Weiss, Paul Jones, and Jane Greenberg >

The genesis of the open source model for software development and distribution goes back to the earliest days of software in university environments when software was developed to solve problems and be freely shared. The term "free software" was popularized by the seminal Free Software Foundation—the parent organization for the GNU (GNU's Not Unix) project—founded in 1984 by MIT researcher Richard Stallman. Stallman's vision was to develop a free operating system, complete with standard software tools such as compilers, interpreters, text editors, mailers, and so forth, in order to re-create a community of cooperating hackers he felt had been lost as Unix was commercialized.

The Free Software Foundation explained its notion of free software in the now-classic distinction, free as in "free speech," not "free beer." That is, free software may or may not be distributed with a monetary cost, but the knowledge that underlies the program—the source code—should be freely available in order to empower future innovation. Software source code is a form of scientific knowledge, and just as scientists publish so that other scientists can build on their results, computer scientists must publish their source code in order to foster continued innovation in

LISA HANEY

computing. In Stallman's words, whether it has an economic advantage or not, free software has a "social advantage, allowing users to cooperate, and an ethical advantage, respecting the user's freedom." [5]

A key component of Stallman's effort in developing a successful free software organization was to formulate a licensing agreement that would prevent businesses from taking free software and using it in binary-only redistributions for commercial gain. Stallman developed the GNU General Public License, known as the GPL or "copyleft," to address this issue. Once code was GPLed, anything derived from that code or included that code must also be covered by GPL and so be "free."

The tone adopted by Stallman, the most prominent free software advocate for some time, was distinctly antibusiness, and as a result, the term "free software" gained some negative connotations for many in the commercial computing world. In early 1997, a group of leaders in the free software community decided to address this problem head-on with a marketing campaign designed to "argue for 'free software' on pragmatic grounds of reliability, cost, and strategic business risk." [4] They were goaded to action largely by frustration over what they felt was the unrecognized potential of free software as a driver of innovation and the basis for the development of commercial-grade software, despite the successes of Apache, Linux, and other projects. An initial decision of the group, which would become the Open Source Initiative (OSI), was to choose the term "open source" for its campaign. They meant to extend the meaning of open source to include a variety of software licenses from those such as the Sun Community License that exposes the code but releases no rights to software placed in the public domain, and even to software covered by the term "free software." Accordingly, OSI adopted a set of criteria, titled "The Open Source Definition," required for a distribution agreement to be designated an OSI-certified license (see opensource.org/osd.html).

In recent years, vocal proponents of open source have effectively engaged the mainstream computing community over practical arguments for the adoption of open source methods (see the April 1999 *Communications*). Some proponents claim free software methods leveraging the Internet represent an alternative economic model for engendering and managing robust software that will ultimately reshape the multibillion-dollar commercial software industry. Skeptics challenge the idea that the technical and organizational approach represented by open source development can really scale up in the coming years and produce the robust software required for large-scale mainstream computing [3].

Despite the importance of the open source movement, little beyond the anecdotal has been published on open source developers and their collective dynamics in large-scale projects. Here, we provide a quantitative profile of a community of open source Linux developers, individuals submitting non-juried contributions to a repository of Linux materials.[1] Our primary methodology is to analyze the contents of over 4,500 contributor-generated metadata files embedded in the actively managed UNC MetaLab[2] Linux repository over the past six years. The results offer insight into the contributor demographics and repository dynamics in this non-juried, broad-based effort of individual contributions to the Linux community.

Linux: Open Source Development on a Global Scale

Perhaps the most influential open source project to date has been and continues to be the Linux operating system. Linux is playing an increasingly significant role in the business plans of established computing companies, in university research labs, and in the development of new companies focused on Linux support and integration issues. According to an April 1999 survey conducted by the Internet Operating System Counter (leb.net/hzo/ioscount/), Linux was the operating system at over 30% of Internet server sites, and many sources have shown evidence of a rapidly growing worldwide user base for Linux.

Begun in 1991 as a personal project of Finnish graduate student Linus Torvalds, the Linux Kernel Project continues today as a loosely coordinated team of core volunteers with Linus as the central coordinator and ultimate decision maker on architectural issues. As with other open source projects, the Kernel Project leverages the power of Internet communications to bring together a large number of developers in a coordinated effort. The credits file accompanying a recent release of the kernel (Linux 2.2.10) lists 190 names, though one estimate has placed the total number of contributors at approximately 1,200 [3].

While the Kernel Project continues, application-level development projects have assumed increasing importance with the rising tide of users and installed systems. Torvalds has commented that in the near future "the most exciting developments for Linux will happen in user space, not kernel space." [7] Unaffiliated individuals, academic groups, and commercial

[1] This work was supported in part by funding from the Council on Library and Information Resources, which awarded Debra Weiss the 1999 A.R. Zipf Fellowship in Information Management. Support was also provided by the Intel Corporation through an equipment grant under the Technology for Education 2000 program. Correspondence regarding this work should be sent to the UNC Open Source Research Team: osrt@metalab.unc.edu.

[2] MetaLab has been renamed ibiblio.org—see www.ibiblio.org.

programmers produce and port to Linux a wide variety of applications, development tools, system components, games, and other software that spreads the use and usefulness of the Linux kernel work. Like the Kernel Project, larger application projects organize through loose coordination with a few central developers driving the overall process. Examples include the Linux Documentation Project (www.linuxdoc.org) and the desktop environment effort, the GNOME project (www.gnome.org/). Smaller contributions are also distributed through actively managed Linux repositories such as the Linux Center (www.portalux.com/) and MetaLab Linux Archives (metalab.unc.edu/pub/Linux). The most popular contributions are often selected for inclusion in the various distributions of Linux assembled, tested, and packaged by Linux groups or companies such as Red Hat Linux, Infomagic, Debian, and others.

Linux Software Maps (LSMs)

Contributions to the MetaLab Linux Archives are required to be accompanied by a small metadata file in a format called the Linux Software Map (LSM). This convention arose naturally from the Linux community and as such is widely adhered to by contributors. From their beginning, Linux Software Map entries were designed to help developers make their contributions highly available to users and to other developers by serving as finding aids as well as a standardized means of announcing new software (to comp.os.linux.announce and other newsgroups). Many, but certainly not all, contributors of Linux software use LSMs to describe their software as they send announcements to comp.os.linux.announce and other newsgroups. Several LSM-based search utilities have been developed to assist users and developers in finding open source contributions, such as linsearch at the MetaLab site. Finally, LSMs also ensure authors are properly credited if and when their software is integrated into Linux distributions.

LSMs are created according to the LSM metadata template consisting of 14 metadata elements, five of which are mandatory. The five mandatory fields are: Title, Version, Entered-date, Description, and Primary-site fields. Based originally on the IAFA (Internet Anonymous FTP Archives) metadata schema developed for Archie [2], the LSM metadata schema has undergone a series of revisions initiated and overseen mainly by Jeff Kopmanis, with input from the Linux community. It is now in its fourth revision and an annotated template is available at metalab.unc.edu/pub/Linux/docs/linux-software-map/lsm-template.

LSMs permit authors to record their expert knowledge about the resource they created, rather than having a secondary party create the representation, as is practiced with many other metadata schemes. Contributors submitting software to the MetaLab Linux Archives place the software and an associated LSM into an incoming directory. Using a program called "keeper" (originally written by open source advocate Eric Raymond), the (human) Linux archivist reviews the LSM information and places the software and LSM in their correct location in the Archives. LSMs help the archivist replace older obsoleted versions of software by use of standard names and version numbers. The LSM is then forwarded to the LSM maintainer for inclusion in the definitive LSM list at execpc.com. Major Linux sites worldwide including MetaLab regularly mirror the LSM list at execpc.com.

> < Open source empowers individual programmers to participate in a large programming community in a meaningful way. >

LSM-accompanied contributions generally represent small contributions of specific applications or utilities, though the size and complexity runs the full gamut from a single GIF file, to complete applications, to entire subsystems for the Linux platform. Since the contribution process at MetaLab is nonjuried, the LSM authors constitute a broad range of developers cutting across many segments of the Linux community. The collective characteristics of LSM authors and their ongoing efforts to create and update open source software are examined here.

Analysis of LSM Metadata

The Linux Software Maps represent a large collection of author-generated metadata. We have analyzed the body of all extant LSMs at a comprehensive Linux site in order to obtain quantitative information on the nature of Linux contributions and their contributors, as seen through this lens. Specifically, we aggregated the information in all well-formed LSMs to provide overall views of contributions across time, authors' (cyber)demographics, and the licensing information provided by authors.

The data here represents analysis using all well-formed LSMs (over 4,500) found in the Linux Archives on the UNC MetaLab server on June 19, 1999. As one of the largest and oldest continuous

Linux repositories, the MetaLab site has collected virtually all Linux-related materials available on the Web, including all major distributions of the base kernel code, the Linux Documentation Project (coordinated and hosted by MetaLab), and a large archive (/pub/Linux on the MetaLab server) of contributed software and auxiliary materials actively managed since 1992. Our study drew from this latter portion of the MetaLab Archives, which contains all the LSM files. Reachable through both FTP and HTTP access, the /pub/Linux repository is large (over 50GB when counting the popular distributions mirrored) and very widely used (for example, access counts often exceed 100,000 transactions per 24-hour period).

Figure 1. LSMs by last-modified date (3842 LSMs, through June 1999).

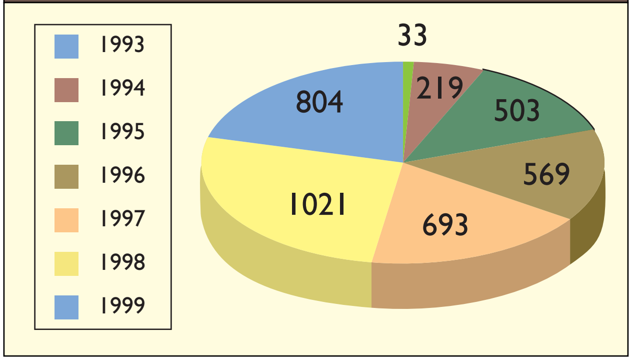
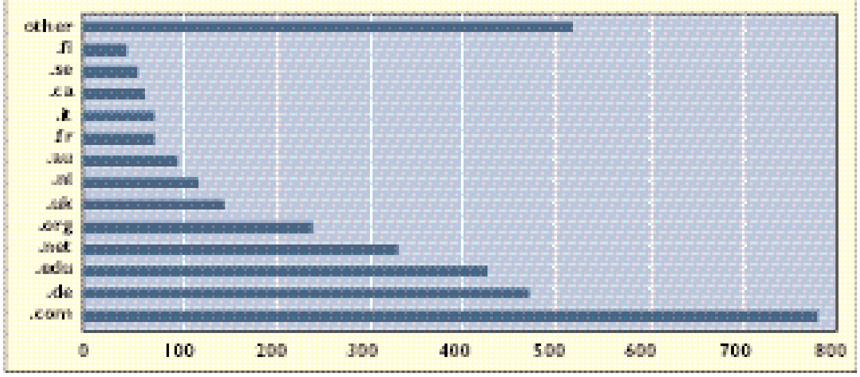


Figure 2. LSM authors' email suffixes.



Except where noted, all data sets are derived from statistical summaries performed by automated processing of relevant fields in the LSM files. Note that the number of LSM files varies across statistics since some LSMs contain missing or unusable field entries (a date field as "Thursday"). For more details on methodology, see [1].

Collection Structure. The table here shows the six top-level directories in the MetaLab collection with the most LSM files. As shown in the accompanying table, the directories for application programs and system utilities contain over one-half of the LSM files. These large directories contain a diverse set of LSM-annotated contributions. The directory apps contains a wide variety of applications for Linux systems (graphics libraries, components for the K Desktop Environment (KDE), multimedia applications, Java utilities, and ports of popular Unix tools). The directory system similarly includes an extensive range of contributions. Examples include device drivers, networking software, NFS, and other file system implementations, and so forth. Finally, another 30% of the LSMs are spread through the directories for contributions related to the X11 windowing system, utilities (shell utilities, file manipulation, and terminal-specific software), games, and development tools such as Perl and Python interpreters, debuggers, and the like.

Figure 1 breaks down the LSMs in the Linux Archives by the date field in the LSM indicating when the LSM was last modified. In interpreting this graph, it is important to remember the policy of the MetaLab Archives has been to replace old LSMs when a new version of a software package arrives. Thus, this data is not an accurate longitudinal study of how many contributions have been made in which years. Rather, it shows that portions of the existing archive extend back to 1993, but many of the contributions have been added or updated recently. Almost one-fourth of the LSMs are additions or updates submitted during the first six months of 1999.

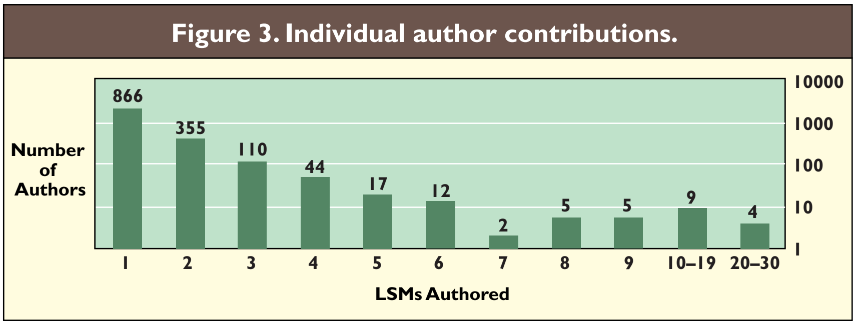
In a separate study (see [8]), we used a mirror of the /pub/Linux portion of the MetaLab server to monitor changes of all file types within the Linux Archives. We found that over this month-long period (April–May 1999), one-third of all activity involved modifying existing files. Since most change under /pub/Linux is driven by LSM submissions, this data shows a significant portion of LSM submissions are updates to existing LSM-accompanied software packages already at the repository. Closer examination of file types revealed about 1.5% of the LSM files (59) were updated and 4% (179) added over this time period. This data gives clear evidence some LSM-based software is being actively maintained and/or evolving over time. A more detailed investigation is needed, however, to determine the exact nature of these updates (for example, bug fixes, versions with new functionality, or other reasons).

Contributor Demographics. The LSM format requires the creator of the package to provide his or her email address. Figure 2 gives a summary of this information by email suffix in order to investigate the demographics of LSM contributors. Linux has been a global phenomenon, and participation has come from a broad-based community. Our data shows the extent of this widespread community. First, we see that contributors indeed come from both the commercial and the nonprofit world domains, with .com as the single largest domain. Moreover, the global nature of the community is clear in the remarkable 71 different country suffixes found here.

Six subdirectories in MetaLab Linux archives containing the most LSMs.

Top-level directory in /pub/Linux on MetaLab	Number of LSMs	Total Number of Files and Bytes
apps	1312	3904 files (994MB)
system	1301	3865 (391MB)
X11	397	2495 (567MB)
utils	373	1670 (218MB)
games	297	896 (130MB)
devel	288	1433 (1.3MB)

Figure 3. Individual author contributions.



Perhaps most striking is the proportion of contributors who have European addresses. As seen in Figure 2, Europeans are well represented in the leading country codes with Germany (.de) appearing more often than any other suffix except .com. An aggregation of all suffixes representing European countries reveals 37% of all LSMs in our data set have authors with European country codes. Of course, this calculation underrepresents the true European participation in Linux development since some authors with geographically unspecific email suffixes such as .com and .net are presumably located in Europe. We conclude that the European roots of the Linux project appear to run deep indeed.

Number of Contributors. We next consider the distribution of the number of LSM contributions per LSM developer to answer the question: Were the LSM contributions created by a few prolific developers or by a large number of individuals who submit one or two pieces of software? Figure 3 shows the frequency count of authors' last names taken from the Author field in the LSMs. The data reveals 2,429 distinct contributors. The vast majority of LSM authors (91.4%) have contributed only one or two items, with only a very small number of developers (2.2%) having produced five or more contributions. Only 13 individual contributors have 10 or more contributions to their credit. This data indicates the breadth of the Linux open source community, revealing that the LSM-accompanied contributions come from many participants adding isolated contributions over time, and are not limited to contributions from a few very prolific developers.

Copyright Information. The LSM record also contains a field by which authors identify the distribution agreement for their software. Reflecting the informal attitude of many contributors, the information here runs the gamut from authors who claim a copyright on their software (rare) to "beerware," "freely distributable," and many variations along these lines. However, by far the most prevalent license cited is the GNU General Public License. Over half of all LSMs use the GPL license in the copyright information field of the LSM. The majority of them appeal only to the GPL, though the rest add qualifications, comments, or include the GPL reference in a mixed license with other distribution policies.

Conclusion

Four key conclusions emerge from our study:

- LSM contributions span a range of software functions. The majority are found in the MetaLab Archives under directories holding application and system-specific software, with only a relative few devoted to games.
- The rate of LSM-based submissions is growing. In addition to new contributions, many LSM submissions to the MetaLab Archives are updates to existing packages.
- LSM authors come from a truly worldwide community spanning many organizations. Europeans have been especially prolific contributors.
- Contributions are spread widely across a base of over 2,400 individuals. Three-quarters of contributors appear as the author of exactly one LSM-annotated submission, and only a handful (2.2%) of application contributors have contributed five or more submissions.

Our study shows the systems and applications categories are by far the largest areas of contribution and games has relatively few contributions. This indicates that in this most open archive with a very low technical barrier, contributors participate in challenging and rewarding technical solutions rather than simple diversions. Linux developers who contribute to the Archives are generally a serious type of individual.

< Open source developers are signaling a bright future for open source communities as a basis for developing and evolving software for the global Internet. >

The Archives continue to grow at a very fast rate (since 1993), showing there is ongoing dedication to the project even if that dedication comes from a shifting set of contributors. These findings confirm that a broadly defined open source project, such as the MetaLab Linux Archives, is sustainable over a long period. Its accelerating growth after six years, an extremely long life for a volunteer project by any standard, is encouraging evidence that long-term, sustainable open source communities can be organized around loose cooperation between volunteers.

Contributors to the Archives are European by a very large margin—14% greater than the next nearest contributor group (.com). Both commercial (.com) and European contributors outnumber U.S. academics and students (.edu) by a very wide margin. Conventional wisdom suggests "free software" and open source developments are driven by academics and students, that when the reality of competition and economics is faced university ideals must be tossed aside, but here in the most open of the open source communities, we see that the .edu contributors account for a mere 12% of the total.

Although some commentators, notably Eric Raymond, have claimed that open source contributors are motivated by "going for the glory," our study shows most people contribute only one or two objects (programs and so forth) to the Archives, rather than a small amount of people contributing many objects. In other words, in this the most open of open source archives, there are few, if any, "great programmer heroes," but rather many individuals contributing single items to the Archives. These numbers support the hypothesis that open source empowers individual programmers to participate in a large programming community in a meaningful way. They also indicate that Raymond is more on target when he says "People do their best work when they are passionately engaged in what they're doing." [6]

With the continuing success of Linux, we conclude that this passionate engagement has resulted in very good and very widely used code. Obviously it is sustainable and produced by a broad community. The community producing the Archives is not an academic one separate from commerce, nor is it U.S.-centric; it is global—although very German—and commercial. The widespread impact thus far from the work of contributors to the Archives from around the globe speaks especially to the remarkable power of a global Internet in connecting communities of people with common interests and goals. Open source developers are taking advantage of that transforming power today, signaling a bright future for open source communities as a basis for developing and evolving software for the global Internet.

References

1. Dempsey, B., Weiss, D., Jones, P. and Greenberg, J. A quantitative profile of a community of open source Linux developers. SILS Tech. Rep. TR-1999-05, School of Information and Library Science, University of North Carolina at Chapel Hill, October 1999; metalab.unc.edu/osrt/.
2. Deutsch, P., Emtage, A., Koster, M., and Stumpf, M. Publishing Information on the Internet with Anonymous FTP (Working Draft), 1995.
3. Lewis, T. Asbestos pajamas: An open source dialogue. *IEEE Computer 32*, 1999, 112.
4. OpenSource.org. Open Source Initiative launch announcement, 1998.
5. Stallman, R. The GNU operating system and the free software movement. OpenSources: Voices from the *Open Source Revolution,* C. DiBona, S. Ockman, and M. Stone, Eds., 1998, 53–70.
6. Taylor, W. Inspired by work: An interview with Eric Raymond. *Fast Company 29*, 200; www.fastcompany.com/online/29/inspired.html.
7. Torvalds, L. The Linux edge. *Commun. ACM* 42, 4 (Apr. 1999), 38–40.
8. Weiss, D. Towards an efficient, scalable replication mechanism for the Internet2 distributed storage infrastructure (I2-DSI) project. Master's Thesis, School of Information and Library Science, University of North Carolina at Chapel Hill, May 1999.

Bert J. Dempsey (dempsey@ils.unc.edu) is an associate professor in the School of Information and Library Science and an adjunct associate professor in Computer Science at the University of North Carolina at Chapel Hill.

Debra Weiss (weisd@ils.unc.edu) is a Ph.D. candidate in the UNC School of Information and Library Science and the recipient of the 1999 A.R. Zipf Fellowship in Information Management from the Council on Library and Information Resources.

Paul Jones (pjones@ibiblio.org) founded ibiblio.org in 1992 as sunsite.unc.edu and has served as the director of the project since that time. He holds appointments in the School of Journalism and Mass Communication and in the School of Information and Library Science at the University of North Carolina at Chapel Hill.

Jane Greenberg (janeg@ruby.ils.unc.edu) is an assistant professor in the School of Information and Library Science at the University of North Carolina at Chapel Hill.