The Scope and Impact of Open Source Software as Intangible Capital: A Framework for Measurement with an Application Based on the use of R Packages

Carol A. Robbins[1], Gizem Korkmaz[2], Claire Kelling[3], Sayali Phadke[3], Stephanie Shipp[2], Sallie Keller[2]

[1]NSF, National Center for Science and Engineering Statistics
[2]Virginia Tech, Social and Decision Analytics Lab
[3]Penn State University

Question

Open source software (OSS) is everywhere, both as specialized applications nurtured by devoted user communities and as digital infrastructure underlying platforms used by millions daily. This type of software is developed, maintained, and extended both within the private sector and outside of it, through the contribution of people from universities, government research institutions, nonprofits, and individuals.

In this paper, we present a framework for measurement of production and use of open source software based on a sectoral analysis: universities, government institutions, nonprofit organizations, households, and businesses. Our goal is to estimate the contribution of universities and government institutions to investments in open source software. Our research question is to test the feasibility of using nonsurvey data to measure the scope and impact of open source software, focusing on the R language and its associated packages as a prototype. The measures we aim to develop include the amount of open source software that is in use, the amount created each year, the sectors where this software is created and used, along with indicators of the relative impact of individual open source packages.

We view OSS as fitting within the definition of an own-account fixed asset because it is expected to be a source of economic benefits to the owner over a period of years. While no fee is charged for issuing OSS licenses, the creators derive benefits by using the software.

Methods

We have a two-pronged approach to measuring the scope and impact of OSS. Drawing from the methodology used for aggregate estimates of own account software in national economic accounts, we estimate the resource cost for these packages by Identifying the number of people involved in each package's development and estimating the time spent on software development. We use average compensation for computer programmers and computer system analysts from Bureau of Labor Statistics wage data. We estimate intermediate inputs based on BEA and OECD methodologies to develop a per package resource cost. Contributor information allows us to identify the sector where the OSS has been created.

We can compare these package costs with the relative impact of each package that can be estimated using methods currently applied to two other outputs of research spending: peer-reviewed publications and patents. For OSS we collect data on the number of complete R packages available, downloads per package, reuses in other packages (dependencies), and citations to OSS. Network analysis provides additional information about the relationships between packages.

Data

Open source software is disseminated online, providing a wealth of information to be collected from archives, repositories and in the code and headers of the software programs themselves. Our data collection strategy combines data from multiple locations. These include the registries that host and distribute OSS, such as the Comprehensive R Archive Network (CRAN); the individual package files that contains metadata such as author, license, version, and dependencies; and the repositories that contain different versions of applications.