Why are indicators on open source software needed?

Software for which source code is public and can be freely copied, shared and modified is called "open source software" (OSS).[1] It is often co-authored using online version control repositories such as GitHub, and may also be bundled into a "package" and uploaded to a "package manager" platform, to be downloaded and re-used by others. There is an incentive to make code as abstract and re-useable as possible, be it within a single program, an organisation or even worldwide as it is inefficient to rewrite code repeatedly (Hunt and Thomas, 1999).

Open source innovation has become a ubiquitous element of digital innovation. Today, open source tools such as Apache servers, Linux operating systems and countless machine learning libraries underpin the functioning of the digital economy. Even market actors famous for proprietary software now see value in OSS. In 2018, Microsoft was the largest contributing organisation to open source projects on the GitHub platform (GitHub, 2018), and acquired it for USD 7.5 billion, while IBM bought Red Hat, an open source operating system, for USD 34 billion.

Despite its contribution to productivity gains in firms (Nagle, 2014), OSS, like other free assets, is a product provided at zero cost, and as such not recorded in the System of National Accounts. Accordingly, the capital services provided by these free assets are also valued with a zero price. Equally, an increasing number of academic outputs take the form of impactful software, which are not accounted for either.[2]

To better understand and measure how the digital transformation is shaping the economy, it is essential to gain insights on OSS. For this reason, the Digital Supply and Use Tables (see page 2.11) include a line for the product category "free services and assets", and as a consequence invite countries to develop methods to estimate the monetary value of these products.

What are the challenges?

Measuring OSS is fraught with conceptual and practical difficulties. Since it is generally the product of collaboration between a wide variety of actors, attributing credit for its creation is difficult, as is estimating its value. In addition, the data available from online sources may at times be incomplete or difficult to interpret.

Statistical frameworks such as the System of National Accounts typically require the identification of a producer and a consumer of an output, but this distinction is often blurred in the case of OSS. Open source developments rely on consumers being able to modify and improve software. Collaborative coding sites generally display projects hosted on a user or organisation's page, but the code itself may be authored by many other users, or even authored by one user and "committed" (approved) by another.

Furthermore, it is difficult to measure the quality of OSS contributions. Weighting them by the number of lines modified may to some extent help, but relies on the assumption that more is better (while it may in fact reflect less efficient programming). Possible alternatives consider the popularity suggested through users "starring" repositories to bookmark them and signal interest (as is done in GitHub); and by the numbers of times a software package is downloaded. Additional quality indicators could be built using information on dependencies between packages (i.e. packages requiring other packages to run), or by analysing actual coding scripts for the use of different packages.

Assigning a monetary value to code is also fraught with difficulty, given the potential diversity of software use and developers' profiles. Robbins et al. (2018) use a combination of average wages, intermediate inputs, capital service costs and lines of code, to estimate that OSS in four languages (R, Python, Julia and JavaScript) is worth USD 3 billion worldwide.

Additional measurement challenges include the sheer volume and quality of data available, and the fact that available data are unstructured, often incomplete, and require computing power and advanced programming skills to be collected and exploited. For instance, many platform users only make public a username rather than a full name, often without complementary information on their geographical location or affiliation.[3] In addition, geographical data obtained from IP addresses may not accurately reflect the location of users or producers due to the use of remote servers.

As an illustration, data suggests that downloaders of Python packages are most frequently located in the United States (over 65%), followed by Ireland and China. However, data on operating systems suggests that a significant share of downloads may come from remote cloud servers. This is most evident in the case of the Amazon Linux AMI distribution (over 6% of downloads), which is used on Amazon Web Services cloud servers. It is likely that the location of cloud servers contributes to making such country-level statistics inaccurate.

1. See the Open Source Initiative for a more comprehensive definition of open source software, https://opensource.org/osd-annotated.
2. Some efforts to track the academic contributions are conducted (see http://depsy.org).
3. Although this may sometimes be possible using data from package managers, as in OECD (2018).