About the proposal for software indicators in OSM

Open Science Committee - France
Free and Open Source Group

September 15th, 2018

Table of Contents

This report follows the request for contributions concerning the indicators proposed in the document Preliminary Open Science Monitor (OSM).

1. Summary of the proposed indicators

The indicators proposed by the OSM for software are the following:

1. Number of code projects with DOI (Mozilla Codemeta)
2. Number of scientific API (Programmableweb)
3. % of journals with open code policy (Stodden 2013, Stodden, V., Guo, P. and Ma, Z. (2013), "Toward reproducible computational research: an empirical analysis of data and code policy adoption", PLoS One , Vol 8 No. 6, pp. E67111, doi: 10.1371 / journal.pone.0067111)
4. Number of scientific projects on Github (Github)

2. Analysis of the proposal

2.1. Terminology and proposal 3)

In terms of terminology, it should first be noted that "open code "is an unknown neologism in the world of software. It has no precise definition, it is not used by software developers, nor by the lawyers and experts in the domain. Thus, the term "open code" could be interpreted as relating to a

code which can simply be browsed on the Internet, without any guarantee that it comes with additional rights of modification and redistribution of changes that characterize open source and free software licenses.

It is obvious that, in the absence of a precise definition of what one wants to measure, it will not be possible to have effective measures reliable or to recommend explicit policies in this field. We therefore strongly discourage the use of this neologism, that is undefined and subject to misinterpretation.

We recommend using the terms "free software" or "open source software", which are well defined and have been used for decades. The definition of these terms is maintained by specialized NGOs (OSI, Open Source Initiative, and FSF, Free Software Foundation). They correspond to a real use in the world of software development and IP law. Some institutions of the European Union promote the aggregative term of "FLOSS", for Free / Libre and open-source software (see https://cordis.europa.eu/publication/rcn/9015_en.html), that can also be used.

In the framework of Open Science, the term "Open Source Scientific Software" could therefore be used.

Consequently, proposition 3) is, of course, to be reviewed.

We will now detail below our comments on proposals 1), 2) and 4).

2.2. 1) using DOIs to count / identify software

DOIs are totally unknown identifiers in the world of software. Using them as indicators would not only be inoperative, but also ***very dangerous***, for the following reasons:

An incomplete measure

There are tens of thousands of scientific software programs openly available, but hardly any of them has an associated DOI: the scientific community, and the developer community, do not use DOI for software. The number of DOIs associated to software would therefore not measure by any means the amount of research software distributed under a free license.

An incorrect measure

Some data repositories, such as Zenodo, propose mechanisms to issue DOIs associated to each ***version*** of a software hosted on GitHub (and GitHub only). If the number of DOIs were to become an indicator, it would be trivial to create dozens, or hundreds of versions of the same software, with the aim of generating a lot of DOIs and therefore inflate artificially this indicator. This one would not measure the amount of software projects, but the number of versions, without identifying the reasons why these are created (major improvement or minor correction), making this indicator irrelevant.

Distortion of competition

In the field of software, DOIs are not suitable (see the article "Identifiers for Digital Objects: the Case of Software Source Code Preservation ", iPres 2018), nor justified by the current practice. HAL, for example, does not issue DOIs, but its own identifiers; the BNF, which is in charge of the deposit software in France, uses Ark identifiers, which are free unlike the DOIs. Adopting DOIs as the only recognized measure of scientific software would not only promote an inappropriate technology: it would create monopoly position, with and annuities (the issuance of DOIs being charged for) to a single actor, in the new field which is the measurement of scientific software production and use.

Systemic and strategic risks

The only platform currently able to provide DOIs on software is the GitHub silo, via its connection with Zenodo. Promoting the use of DOIs would therefore boil down to encouraging all scientific software developers to migrate to GitHub, to the detriment of platforms that could provide in the longer term more innovative and relevant services. At a time when GitHub has just been bought by a major player in North American software publishing is not in the interest of Europe to encourage its developers to use a single non-European platform, controlled by private actors that have no obligation nor commitment to maintain a public service in the long term, as the recent shutdown of Gitorious and Google code has made self-evident.

2.3. 4) GitHub as the only reference for counting software projects

This proposal is to be rejected absolutely, for the following reasons:

An incomplete indicator

There is a large amount of scientific software available on public forges different from GitHub (which has only existed for 10 years) or institutions; many are only accessible from the web page of their authors. It is precisely for collecting in one place all of this scattered software that the French national open access portal, HAL, has partnered with the Software Heritage Foundation to offer a scientific software deposit that is going to be open to all at the end of the month of September 2018.

So, counting only the scientific software that is present on GitHub would amount to ignoring a large mass of existing scientific software available under an open source license.

An indicator hard to build

It is very unclear how one can identify all scientific projects by just looking at GitHub, which is already (as seen above) introducing a huge bias.

Distortion of competition and systemic risks

Using GitHub as the only official repository would harm competing platforms and institutional repositories, introducing systemic and strategic risks in the long term (see the arguments detailed in the previous points).

3. Alternative Proposals

There are several alternatives available to build way better indicators for the scientific software that is made available under an open source license. Some are immediately actionable, others are more in the medium term.

3.1. Actionable alternatives

3.1.1. Leverage the OpenAire European research infrastructure

OpenAire is a 10+ year European union-funded infrastructure that aggregates all the open access publications available in the many open access portals across Europe, and beyond. In collaboration with Software Heritage, OpenAire is building a catalogue of software projects mentioned in the scientific articles indexed in its database, which provides a concrete first step to build a transparent, independently verifiable indicator.

Considering that OpenAire is an EU-funded infrastructure, it would be difficult not to leverage it for building the software indicators for the OSM.

3.1.2. Leverage the Software Heritage universal software archive

Identifying open source research software is a preliminary step to building a software indicator for the OSM. Ensuring that the identified software is properly available to other researchers requires an extra essential step: archiving it.

Software Heritage (www.softwareheritage.org) is building the universal archive of software source code, and has already been identified as an essential infrastructure in the French National Plan for Open Science.

It is essential that the OSM counts the amount of open source research software that is safely archived in Software Heritage.

Software Heritages provides also a high quality harvesting process of many software repositories. It is can be seen as a better starting point than GitHub.

3.2. Medium term initiatives

3.2.1. Leverage existing catalogues of scientific software

There are tens of thousands of important scientific software that has been available under an open source license for years and even decades. Many scientific communities have built manually curated catalogues whose quality is way superior to any automated index one can imagine today. Building an aggregate index for these catalogues is the way to go if we want to create a meaningful indicator in the medium to long term.

4. Conclusion and recommendation

The analysis provided in this report shows that the set of indicators proposed under the OSM for the software activity of the researchers is very far from satisfactory, and cannot even be considered as a baseline, because of the methodological and economic bias they entail.

Furthermore, we strongly remark that scientific software is *a small part of a much bigger ecosystem of software development* that involves industries and developer communities well outside the research sector. Hence, any proposal made concerning research software must take into account the practice and standards of the developer community at large.

Hence, we strongly advises the creation of a working group involving a representative panel, including researchers having a long experience in software and hardware development, actors of the transfer of software and research materials to industry, representatives of open source developer communities, legal experts from the world of software and free hardware, and scientists with the necessary skills to assess the feasibility of constructing the proposed indicators.

Contacts :

- Roberto Di Cosmo roberto@dicosmo.org
- François Pellegrini francois.pellegrini@u-bordeaux.fr
- Marin Dacos marin.dacos@recherche.gouv.fr