Licensing R - Guidelines and tools

Colin FAY & Miles McBain 2018-10-31

The Challenge

Licensing is a vital part of Open Source. It provides guidelines for interacting with a program, and for making code accessible and reusable (or not). It provides a way to make code open source, in a way one wants to share it, protecting how it will be used and reused.

Licensing is also challenging and complex: there are a lot of available licenses, and the choice is influenced by how you import and interact with elements from other packages and/or programs.

It's even harder when it comes to combining licenses, as it can lead to potentially incompatible interaction. For example:

> "By combining GPL-licensed code with code under any but the most unrestrictive licenses, the creator of the putative"new program" is imposing a restriction (compliance with the terms of that license) that is not present in the GPL License and which accordingly violates the GPL"

Understanding Open Source and Free Software Licensing, Andrew M. St. Laurent

As stated in Miles McBain proposal (1) for an unconf project around this topic, a number of developers use license without a clear knowledge of what these licenses precisely imply. When one wants to choose a license, a lot of questions can arise. For example:

- How do the dependencies to a package impact the choices of a license?
- Can we use any license we want in an R package?
- Can we use any license we want with R in general? R is *"is distributed under the terms of the GNU General Public License, either Version 2, June 1991 or Version 3, June 2007"*(2). How does this impact the code one is writing?
- Do `Depends`/`Suggests` have the same impact on license choice?
- How can we include and license a data set in a package?
- What are one rights when contributing to a package?
- What are the obligation of a package maintainer when it comes to changing the license of a package? Can one change the license at any time?
- What does each license really imply when using a package?
- How does a dependency license influence the licensing of a package?
- What should be done if a package depends on two non-compatible open source licenses?
- Are there global vs local states of open sources licenses (country-specific)? Can country-specific licenses be used in another country? See for example the CeCILL license (3), used by 18 packages on the CRAN (4).
- How do license impact the writing and publication of books/articles/blog posts/instructional materials?

Licensing and R

A quick dive into the CRAN package database can give us an overview of how complex licensing is and of how the community is making a lot of different choices.

```
# Done on
Sys.Date()
```

```
## [1] "2018-10-31"
```

```r
db <- tools::CRAN_package_db()
```

Let's ask us a simple question: how many licenses are currently used on the CRAN?

```r
# How many different licenses?
length( unique( db$License ) )
```

```
## [1] 155
```

Inside this list, various licenses: GPL, Apache, LGPL, AGPL, CC0, MIT...

A lot appears to be a variation of the GPL license:

```r
# How many GPL based licenses ?
length( unique( grep("GPL", db$License, value = TRUE) ) )
```

```
## [1] 83
```

And some are used just once:

```r
# How may licenses are used just once?
sum( table( db$License ) == 1 )
```

```
## [1] 56
```

From these two numbers, two questions arise:

- Why are there that many different licenses in use? Are there really 155 different configurations that require the use of 155 different licenses?
- When it comes to the packages that are the only one using a specific license, what makes these packages so special/different from the others so that they need such a rarely used license?

One guess could be that developers choose a license without a deep knowledge about this license, leading to a choice which might not be the optimal one.

The plan

Licensing: documentation & consulting

The first part of the plan is to gather documentation and notes about current state of open source licenses, and to decipher compatibility and incompatibly elements inside these licenses.

This first steps will include:

- gathering information from current state of open source licenses.
- (if needed) consulting one/several lawyer(s) specialized in open source to gather advice about our findings.

Licensing: guidelines

Online book

Once the first step is completed, we plan to write an online book (written with bookdown) containing all the results from our findings.

This book will be distributed as open source (under a license we will choose based on our findings). The idea is to provide a simple but comprehensive overview of open source licenses and how to use them in R.

This book will be focused on licensing R-related development, but the first part of the book will be more general, hence it could be of interest to a broader audience. In other words, even someone coding in another open source language will find relevant information in the first part of the book.

Here's a draft of the general outline of the book (subject to change):

1. Introduction

- What are open source licenses?
- Why are they important?
- A short history of open source licenses
- Key concepts of open source licensing (copyleft, open source, ...)

2. General Overview of Open Source licenses

- Standard and widely-used open source licenses
- International & Country Specific licenses
- Purpose-specific licenses
- Non-reusable licenses
- Uncategoized
- (inspired by https://opensource.org/licenses/category)

3. Licensing R Code: packages

- Dependencies
- Datasets
- Including external code or programs
- Contributing to someone's package

4. Licensing R Code : publication

- Publishing book
- Blogging about R
- Teaching materials & conferences

5. Conclusion

Licensing: tools

Package

The last step will be the development of a package that will be able to give guidelines about the possibility of licensing for a package.

This package will allow a developer to parse the skeleton of another package, and to get a quick report about licensing of this specific package. The idea is to answer these questions:

- Is my license choice compatible with the dependencies I've chosen?
- Is my license choice compatible with R in general?
- Can I include this dataset in my package?
- Can I include this program in my package?
- What conditions does my license place on people who wish to use or depend on my package?

The Team

The research will be made by:

- Colin Fay (5) - Data Scientist, R Hacker & Trainer at ThinkR, Open Source developer, Blogger, Speaker.
- Miles McBain (6) - Research Associate at Queensland University of Technology (QUT), Statistician & Open Source Developer, and Blogger.

The results will be hosted on GitHub, open to external contributions. The repository will have a Code of Conduct, and will be completed with a contribution guide. Every contribution will be welcome, be it from a beginner or a more experienced developer.

Milestones

Documentation

The first step of the project will be to read and gather as much information as needed around open source licensing. We have started to gather resources on the GitHub repository (7).

Estimated time: 3 to 4 months

Guidelines & Writing

Estimated time: 3 to 4 months

Tooling

Estimated time: 1 months

How Can The ISC Help

We are asking for a grant to support the working days spent to investigate, gather information, to develop the tools and guidelines, and to promote them. We estimate the documentation and writing to take at least 30 days (15 days each), so around 210 hours. The package development should take around 6 days (3 days each), so around 42 hours.

We are asking for the support of half of these hours from the RConsortium, based on a rate of 100$ / hour. The other part will be covered by ThinkR & QUT, in their effort to support Open Source Software, by freeing time for us to work on this project.

We would also need a "floating budget" of 4K for external legal opinion (to be used or not), and of 1K for documentation (to be used or not). These two budgets will allow us to buy documentation if we need to, and to get advice from an expert in case of need.

Below is a summary of our needs:

What	How long	How much
Documentation	4 months	5250
Guidelines	4 months	5250
Tooling	1 month	2100
Floating	—	5000

Total:

- Fixed: 12600, to be shared between Colin Fay & Miles McBain.
- Floating: 5000, to be used in case of need.

Dissemination

Communicating

A big part of the success of the project will be communication. We hope that the community will grasp this opportunity to contribute and help, either by providing inputs or feedback.

We hope this book will be of interest to a broad audience, that is to say to people not developping in R but looking for information about open source licensing.

We will publicise our work through several channels:

- Blogposts on R related blogs (on ThinkR and others)
- Article proposals on R and software engineering journals (R Journal, Journal of Statistical Software, ...)
- Talks at meetups and conference presenting our findings (in particular useR).

Footnotes

1) https://github.com/ropensci/unconf17/issues/32
2) From `base::license()`
3) http://www.cecill.info/index.fr.html
4) https://github.com/ThinkR-open/isc-proposal-licence/issues/2
5) https://colinfay.me/
6) https://milesmcbain.xyz/
7) https://github.com/ThinkR-open/isc-proposal-licence/issues/1