software carpentry (http://software-carpentry.org)

Open Science

Objectives

- Explain how the GNU Public License (GPL) differs from most other open licenses.
- Explain the four kinds of restrictions that can be combined in a Creative Commons license.
- Correctly add licensing and citation information to a project repository.
- Outline options for hosting code and data and the pros and cons of each.

> The opposite of "open" isn't "closed". The opposite of "open" is "broken".
> — John Wilbanks

Free sharing of information might be the ideal in science, but the reality is often more complicated. Normal practice today looks something like this:

- A scientist collects some data and stores it on a machine that is occasionally backed up by her department.
- She then writes or modifies a few small programs (which also reside on her machine) to analyze that data.
- Once she has some results, she writes them up and submits her paper. She might include her data—a growing number of journals require this—but she probably doesn't include her code.
- Time passes.
- The journal sends her reviews written anonymously by a handful of other people in her field. She revises her paper to satisfy them, during which time she might also modify the scripts she wrote earlier, and resubmits.
- More time passes.
- The paper is eventually published. It might include a link to an online copy of her data, but the paper itself will be behind a paywall: only people who have personal or institutional access will be able to read it.

For a growing number of scientists, though, the process looks like this:

- The data that the scientist collects is stored in an open access repository like figshare (http://figshare.com/) or Dryad (http://datadryad.org/) as soon as it's collected, and given its own DOI.
- The scientist creates a new repository on GitHub to hold her work.
- As she does her analysis, she pushes changes to her scripts (and possibly some output files) to that repository. She also uses the repository for her paper; that repository is then the hub for collaboration with her colleagues.
- When she's happy with the state of her paper, she posts a version to arXiv (http://arxiv.org/) or some other preprint server to invite feedback from peers.
- Based on that feedback, she may post several revisions before finally submitting her paper to a journal.

- The published paper includes links to her preprint and to her code and data repositories, which makes it much easier for other scientists to use her work as starting point for their own research.

This open model accelerates discovery: the more open work is, the more widely it is cited and re-used. However, people who want to work this way need to make some decisions about what exactly "open" means in practice.

Licensing

The first question is licensing. Broadly speaking, there are two kinds of open license for software, and half a dozen for data and publications. For software, people can choose between the GNU Public License (http://opensource.org/licenses/GPL-3.0) (GPL) on the one hand, and licenses like the MIT (http://opensource.org/licenses/MIT) and BSD (http://opensource.org/licenses/BSD-2-Clause) licenses on the other. All of these licenses allow unrestricted sharing and modification of programs, but the GPL is infective (../../gloss.html#infective-license): anyone who distributes a modified version of the code (or anything that includes GPL'd code) must make *their* code freely available as well.

Proponents of the GPL argue that this requirement is needed to ensure that people who are benefiting from freely-available code are also contributing back to the community. Opponents counter that many open source projects have had long and successful lives without this condition, and that the GPL makes it more difficult to combine code from different sources. At the end of the day, what matters most is that:

1. every project have a file in its home directory called something like `LICENSE` or `LICENSE.txt` that clearly states what the license is, and
2. people use existing licenses rather than writing new ones.

The second point is as important as the first: most scientists are not lawyers, so wording that may seem sensible to a layperson may have unintended gaps or consequences. The Open Source Initiative (http://opensource.org/) maintains a list of open source licenses, and tl;drLegal (http://www.tldrlegal.com/) explains many of them in plain English.

When it comes to data, publications, and the like, scientists have many more options to choose from. The good news is that an organization called Creative Commons (http://creativecommons.org/) has prepared a set of licenses using combinations of four basic restrictions:

- Attribution: derived works must give the original author credit for their work.
- No Derivatives: people may copy the work, but must pass it along unchanged.
- Share Alike: derivative works must license their work under the same terms as the original.
- Noncommercial: free use is allowed, but commercial use is not.

These four restrictions are abbreviated "BY", "ND", "SA", and "NC" respectively, so "CC-BY-ND" means, "People can re-use the work both for free and commercially, but cannot make changes and must cite the original." These short descriptions (http://creativecommons.org/licenses/) summarize the six CC licenses in plain language, and include links to their full legal formulations.

There is one other important license that doesn't fit into this categorization. Scientists (and other people) can choose to put material in the public domain, which is often abbreviated "PD". In this case, anyone can do anything they want with it, without needing to cite the original or restrict further re-use. The table below shows how the six Creative Commons licenses and PD relate to one another:

	Licenses that can be used for derivative work or adaptation

Original work	by	by-nc	by-nc-nd	by-nc-sa	by-nd	by-sa	pd
by	X	X	X	X	X	X	
by-nc		X	X	X			
by-nc-nd							
by-nc-sa				X			
by-nd							
by-sa						X	
pd	X	X	X	X	X	X	X

Software Carpentry (http://software-carpentry.org/license.html) uses CC-BY for its lessons and the MIT License for its code in order to encourage the widest possible re-use. Again, the most important thing is for the `LICENSE` file in the root directory of your project to state clearly what your license is. You may also want to include a file called `CITATION` or `CITATION.txt` that describes how to reference your project; the one for Software Carpentry states:

```
To reference Software Carpentry in publications, please cite both of the foll
owing:

Greg Wilson: "Software Carpentry: Lessons Learned". arXiv:1307.5448, July 201
3.

@online{wilson-software-carpentry-2013,
  author      = {Greg Wilson},
  title       = {Software Carpentry: Lessons Learned},
  version     = {1},
  date        = {2013-07-20},
  eprinttype  = {arxiv},
  eprint      = {1307.5448}
}
```

Hosting

The second big question for groups that want to open up their work is where to host their code and data. One option is for the lab, the department, or the university to provide a server, manage accounts and backups, and so on. The main benefit of this is that it clarifies who owns what, which is particularly important if any of the material is sensitive (i.e., relates to experiments involving human subjects or may be used in a patent application). The main drawbacks are the cost of providing the service and its longevity: a scientist who has spent ten years collecting data would like to be sure that data will still be available ten years from now, but that's well beyond the lifespan of most of the grants that fund academic infrastructure.

Another option is to purchase a domain and pay an Internet service provider (ISP) to host it. This gives the individual or group more control, and sidesteps problems that can arise when moving from one institution to another, but requires more time and effort to set up than either the option above or the option below.

The third option is to use a public hosting service like GitHub (http://github.com), BitBucket (http://bitbucket.org), Google Code (http://code.google.com), or SourceForge (http://sourceforge.net). All of these allow people to create repositories through a web interface, and also provide mailing lists, ways to keep track of who's doing what, and so on. They all benefit from economies of scale and network effects: it's easier to run one large service well than to run many smaller services to the same standard, and it's also easier for people to collaborate if they're using the same service, not least because it gives them fewer passwords to remember.

However, all of these services place some constraints on people's work. In particular, most give users a choice: if they're willing to share their work with others, it will be hosted for free, but if they want privacy, they may have to pay. Sharing might seem like the only valid choice for science, but many institutions may not allow researchers to do this, either because they want to protect future patent applications or simply because what's new is often also frightening.

Key Points

- Open scientific work is more useful and more highly cited than closed.
- People who incorporate GPL'd software into theirs must make theirs open; most other open licenses do not require this.
- The Creative Commons family of licenses allow people to mix and match requirements and restrictions on attribution, creation of derivative works, further sharing, and commercialization.
- People who are not lawyers should not try to write licenses from scratch.
- Projects can be hosted on university servers, on personal domains, or on public forges.
- Rules regarding intellectual property and storage of sensitive information apply no matter where code and data are hosted.

Find out whether you are allowed to apply an open license to your software. Can you do this unilaterally, or do you need permission from someone in your institution? If so, who?

Find out whether you are allowed to host your work openly on a public forge. Can you do this unilaterally, or do you need permission from someone in your institution? If so, who?

Email (admin@software-carpentry.org) **Twitter (https://twitter.com/swcarpentry)**
RSS (http://software-carpentry.org/feed.xml) **GitHub (https://github.com/swcarpentry)** **IRC (irc://moznet/sciencelab)**
License (../../LICENSE.html)
Bug Report (mailto:admin@software-carpentry.org?subject=bug%20in%20novice/git/04-open.md)