Open Science and Verifiability

J. Daniel Gezelter
gezelter@openscience.org

Associate Professor, Department of Chemistry & Biochemistry, University of Notre Dame, and Director, The OpenScience Project (openscience.org)

What, exactly, is *Open Science*? It is often easiest to answer this question by listing the four issues which garner the largest attention. These are (in no particular order): Open Source, Open Data, Open Access, Open Notebook. These banner issues are really just shorthand for a few fundamental goals:

- Transparency in experimental methodology, observation, and collection of data.
- Public availability and reusability of scientific data.
- Public accessibility and transparency of scientific communication.
- Using web-based tools to facilitate scientific collaboration.

At the OpenScience project, the idea we've tackled most often is the goal of transparent methodology. It is our view that granting access to source code is equivalent to publishing scientific methodology when the kind of science being done involves numerical experiments. Without access to the source code for the programs we use, we rely on faith in the coding abilities of other people to carry out our numerical experiments. In some cases (i.e. when simulation codes or parameter files are proprietary or are hidden by their owners), numerical experimentation *isn't even science*. A secret experimental design doesn't give skeptics the ability to repeat (and hopefully verify) your experiment, and the same is true with numerical experiments. Science has to be "verifiable in practice" as well as "verifiable in principle".[1]

Good science is *verifiable*

The difference between *falsifiability* and *verifiability* in science deserves a bit of elaboration. It is not always obvious (even to scientists) what principles they are using to evaluate scientific theories.[2] We'll start a discussion of verifiability by thinking about *Popper's asymmetry*.[3] Consider a scientific theory (T) that predicts an observation (O). There are two ways we could approach adding the weight of experiment to a particular theory. We could attempt to falsify or verify the observation. Only one of these approaches (falsification) is deductively valid:

[1] The distinction between *verifiable in principle* and *verifiable in practice* was originally made in: Ayer, A. J. *Language, Truth and Logic*, (New York: Dover, 1946) p. 32.

[2] This discussion closely follows a treatment of Popper's asymmetry in: Sober, Elliot *Philosophy of Biology* (Boulder: Westview Press, 2000), pp. 50-51.

[3] Popper, Karl R. "The Logic of Scientific Discovery" 5th ed. (London: Hutchinson, 1959), pp. 40-41, 46.

Falsification	*Verification*
If *T*, then *O*	If *T*, then *O*
Not-*O*	*O*
Not-*T*	*T*
Deductively Valid	Deductively Invalid

Popper concluded that it is impossible to know that a theory is true based on observations (*O*); science can tell us only that the theory is false (or that it has yet to be refuted). He concluded that meaningful scientific statements are *falsifiable*.

A more realistic picture of scientific theories isn't this simple. We often base our theories on a set of *auxiliary assumptions* which we take as postulates for our theories. These auxiliary assumptions show us that real science is very often not a deductively valid exercise. The Quine-Duhem thesis [4] recovers the symmetry between falsification and verification when we take into account the role of the auxiliary assumptions (*AA*) of the theory (*T*):

Falsification	*Verification*
If (*T and AA*), then *O*	If (*T and AA*), then *O*
Not-*O*	*O*
Not-*T*	*T*
Deductively Invalid	Deductively Invalid

That is, if the predicted observation (*O*) turns out to be false, we can deduce only that something is wrong with the conjunction, (*T* and *AA*); we cannot determine from the premises that it is *T* rather than *AA* that is false. In order to recover the asymmetry, we would need our assumptions (*AA*) to be independently verifiable:

Falsification	*Verification*
If (*T and AA*), then *O*	If (*T and AA*), then *O*
AA	*AA*
Not-*O*	*O*
Not-*T*	*T*
Deductively Valid	Deductively Invalid

Falsifying a theory requires that auxiliary assumption (*AA*) be demonstrably true. Now, since auxiliary assumptions are often highly theoretical, if we can't verify *AA*, we will not be able to

[4] Gillies, Donald. "The Duhem Thesis and the Quine Thesis", in Martin Curd and J.A. Cover ed. *Philosophy of Science: The Central Issues*, (New York: Norton, 1998), pp. 302-319.

falsify *T* by using the valid argument above. Contrary to Popper, there really is no asymmetry between falsification and verification. *If we cannot verify theoretical statements, then we cannot falsify them either.*

Since verifying a theoretical statement is nearly impossible, and falsification often requires verification of assumptions, where does that leave scientific theories? What is required of a statement to make it *scientific*?

Carl Hempel came up with one of the more useful statements about the properties of scientific theories:[5] "The statements constituting a scientific explanation must be capable of empirical test." More explicitly, a theory is scientifically objective if it is both publicly verifiable and testable. Public verification minimizes subjective misperception and increases the probability of objective perception. Good scientific theories must be verifiable in both a *conceptual* and an *operational* sense.[6]

Computational science and verifiability

Modern science relies to a very large degree on computer simulations, computational models, and computational analysis of very large data sets. These methods for doing science all have underlying theoretical assumptions that are *verifiable in principle*. For very simple systems, and small data sets this is nearly the same as *verifiable in practice*. As systems become more complex and the data sets become too large to reproduce, calculations that are verifiable in principle are no longer verifiable in practice without public access to the code (or data). If a scientist makes a claim that a skeptic can only verify by spending three decades writing and debugging a complex computer program that exactly replicates the workings of a commercial code, the original claim is really only *verifiable in principle*. If we really want to allow skeptics to test our claims, we must allow them to see the workings of the computer code that was used. It is therefore imperative for skeptical scientific inquiry that software for simulating complex systems be available in source-code form and that real access to raw data be made available to skeptics.

Our position on open source and open data in science was arrived at when an increasing number of papers began crossing our desks for review that could not be subjected to verifiability tests in any meaningful way. Paper A might have used a commercial package that comes with a license that *forbids people at university X from viewing the code*![7] Paper 2 might use a code which requires parameter sets that are "trade secrets" and have *never been published in the scientific literature*. Our view is that it is not healthy for scientific papers to be supported by computations that *cannot* be verified except by a few employees at a commercial software developer. *Should this kind of work even be considered Science?*

Other "Opens" in Open Science

The emphasis at openscience.org has always been about promoting and recognizing developers of open source scientific software. We do this because closed-source scientific software raises real questions about universality and verification. Open source is not the only "Open" debate in the scientific community, however.

---

[5] C. Hempel. *Philosophy of Natural Science* 49 (1966).

[6] Lett, James, *Science, Reason and Anthropology, The Principles of Rational Inquiry* (Oxford: Rowman & Littlefield, 1997), p. 47

[7] See, for example www.bannedbygaussian.org

Open Notebook science is a way of working that makes available the entire record of a research project as it is recorded. This can include publicly searchable lab protocols, raw data, and incomplete experiments with warts and all. There are now more than 100 individual labs publishing their protocols on openwetware.org. Some of the pioneers of Open Notebook science include Jean-Claude Bradley and Cameron Neylon.

Open Data is the idea that primary scientific data should be available to anyone without restrictions from copyright, patents, or other mechanisms of control. Alex Kandel, one of my colleagues at Notre Dame, puts all of the raw data from his Scanning Tunneling Microscopes online as soon as it is acquired. Some open data sets are very large (i.e. genomes, meta-genomes, proteomes, and databases of chemical structures). Pub3D (created by Rajarshi Guha at Indiana University, Bloomington) now hosts over 17 million 3D chemical structures! The RCSB protein data bank is perhaps the best-known Open Data project, but there are many more data sets made available every year.

Open Access is the idea that scientific research should be published in such a way that the findings of a study are accessible to all potential users without any barriers. The Open Access debate pits two things that are good for science against each other:

- The public funding of science implies that the public can expect (and should expect) to be able to see the results of that scientific effort in a timely manner.
- Peer review and editing are good for science. Publishers provide a valuable and under-appreciated service. This service costs money.

Why isn't all science "Open"?

In general, we're moving towards an era of greater transparency in a number of areas (methodology, data, communication, and collaboration). The problems we face in gaining widespread support for Open Science are really about incentives and sustainability. How can we design or modify the scientific reward systems to make these four activities the natural state of affairs for scientists? Right now, there are some clear disincentives to participating in these activities. Scientists are motivated by most of the same things as normal people:

- Money, for themselves, for their groups, and to support their science.
- Reputation, which is usually (but not necessarily) measured by citations, h-indices, download counts, placement of students, etc.
- Sufficient time, space, and resources to think and do their research (which is, in many ways, the most powerful motivator).

Right now, the incentive network that scientists work under seems to favor "closed" science. Scientific productivity is measured by the number of papers in traditional journals with high impact factors, and the importance of a scientist's work is measured by citation count. Both of these measures help determine funding and promotions at most institutions, and doing open science is either neutral or damaging by these measures. Time spent cleaning up code for release, or setting up a microscopy image database, or writing a blog is time spent away from writing a proposal or paper. The "open" parts of doing science just aren't part of the incentive structure.

Recognition and Attribution

The table that follows summarizes two examples of “open” scientific projects. ***Jmol*** is a molecular visualization tool that is widely used in chemical instruction, journals (ACS & Royal Society), numerous protein and materials science, and the RCSB protein databank. It has been an unfunded “hobby” project that has been passed down to 5 separate lead developers over the years. **OpenMD** (OOPSE) is a single-group research code that has been developed while carrying out funded research on other topics.

	Jmol	OpenMD / OOPSE
Started:	1998	2004
Purpose:	Molecular Visualization	Molecular Dynamics Simulations
Languages:	Java	C++, F95, Python
Developers:	29	7 (mostly graduate students)
Lead Developers:	5	1
Code base:	262,143 lines	134,700 lines
Person-Years:	67	18
Estimated Development Costs:	\$3.7 M	\$400,000
Explicitly-funded Costs:	\$0	\$0
Downloads:	Over 150,000 at SourceForge alone, (possibly millions more)	29,000
External Citations:	13	3
Citations to lead developers:	3	3

The problems above are two-fold. Both of these scientific codes were expensive, complex projects which have *never had explicit funding*. These are tools that are useful to the scientific community, but *there has never been commensurate recognition of these contributions in the form of citations in the scientific literature*. Nearly all developers of open source scientific codes will have similar stories to tell about the level of monetary support and academic recognition of their contributions.

In the Open Source community, code re-use is *encouraged*. That is, a few lines of code from one project can diffuse into yours. Code re-use allows us to avoid reinventing the wheel in each project. Even if attributions are left in the code comments or the various license files, the *users* of your code may never know (or care) who wrote which bits. Developers of open source codes gain community recognition for their skills and can point to their contributions when looking for jobs or making connections to other developers. As Eric Raymond pointed out, the open source community is a *gift economy* in which developers are paid in prestige for their contributions.

In the scientific community, using a paragraph from my paper in your paper without attribution is *plagiarism*, and is considered serious misconduct. Re-use *requires* quotation and citation. Over the course of the past century, we have developed a number of methods for tabulating the interdependence of scientific papers in order to determine the importance of a body of work. The common metrics of publication are: 1) paper count, 2) citation count, and 3) h-index. The traditional methods for tabulating and measuring contributions to science have a "quantum of effort" which is the traditional research paper and a "quantum of recognition" which is the citation. These measures have not translated well into a more granular world of rapid electronic communications.

Scientists who are considering contributing to or starting an "open" science effort face a number of questions about how this contributions will affect their careers: What happens if time and effort exerted in pursuit of open science projects reduces the ability to publish? Why aren't open projects treated more like publications? How should you cite an online STM image database? A blog? An open source data visualization project? How does that citation get tied to a particular researcher?

The attribution metrics should (but currently don't) take into account:

- Amount of scientific effort,
- Complexity of the work,
- Importance of the work to the scientific community,
- Externalities of the work beyond the scientific community.

The main problem of "open" forms of scientific reputation-building is that there's no way to tie these efforts (those outside of traditional publications) into a metric that can be used by institutions. More than any other factor we've considered, the inability of institutions to measure scientific or scholarly impact of contributed software, contributed data sets, blogging, and other forms of non-standard publishing is what limits participation in open science.

Sustainability

Another large issue facing open science projects is the issue of sustainability. We might define 'sustainability' as having a mechanism in place for generating, or gaining access to, the economic resources necessary to keep the intellectual property or the service available on an ongoing basis.[8]

Why is sustainability so difficult? Many of the open science efforts involve not only original research, but continuing service to the users and the community. PIs in academia usually have experience that is limited to: 1) doing research, 2) writing papers, 3) teaching, and 4)

[8] Guthrie, K; Griffiths, R; and Maron, N. "Sustainability and Revenue Models for Online Academic Resources" A report to the Strategic Content Alliance (SCA).

securing funding. This expertise is quite different from what is required of the leader of a service enterprise. Some suggested models of sustainability from the SCA report include:

1. Direct beneficiaries pay
   a. Subscription or one-time payment
   b. Pay per use
   c. Contributor pays

2. Indirect beneficiaries pay
   a. Host institution's support
   b. Corporate sponsorships
   c. Advertising
   d. Philanthropic funding
   e. License content

In the current funding model for scientific research, open projects may have an easy time securing funding for new features and applications, but have difficulty securing stable funding for maintenance and improving usability. Our attempts to define sustainability models for scientific software projects have included the following suggestions (some of which are more applicable to commercial interests attempting to contribute to open projects):

- Sell something physical (i.e. a spectrometer)

- Sell services (i.e. support for a complicated program)

- Sell advertising

- Dual-license (i.e. academic vs. commercial software licensing)

- Differentiate between single-run and high-throughput versions

- Philanthropic funding

We know of only a handful of open science projects that have secured sufficient funding or institutional support to graduate to true sustainability. Many other promising projects languish as PIs, post-docs, and graduate students move on to the next fundable topic.

Outlook for Open Science

It is vital that the scientific community think through some of the implications of complex simulations and large data sets, but because of the attribution and recognition and sustainability issues, we're not particularly sanguine about the outlook for Open Science.

Any solution to these problems is going to have to work within the established behaviors of the various communities. In our opinion it is time to make some changes to Michael Faraday's advice to his junior colleague to: "Work. Finish. Publish." It shouldn't be enough to publish a paper anymore. If we want open science to flourish, we should raise our expectations to: "Work. Finish. Publish. Release." That is, scientific research shouldn't be considered complete until the data and meta-data is put up on the web for other people to use, until the code is documented and released, and until the comments start coming in to the blog post announcing the paper. If the general expectations of what it means to complete a project are raised to this level, the scientific community will start doing these activities as a matter of course.