



The Open Science Project: OpenScience.org

J. Daniel Gezelter

Associate Professor, Department of Chemistry and Biochemistry
University of Notre Dame, Notre Dame IN, 46556



What is Open Science?

- Transparency in methodology and collection of data.
- Availability and re-use of scientific data.
- Public accessibility to scientific communication.
- Using social media to facilitate scientific collaboration.

Open Science is the idea that scientific knowledge of all kinds should be openly shared as early as is practical in the discovery process.

Science must be reproducible

The statements constituting a scientific explanation must be capable of test by reference to publicly ascertainable evidence.¹ One thing we expect from empirical tests is the universality of the results. Independent scientists should be able to subject theories to similar tests in different locations, on different equipment, and at different times and get similar answers. *Reproducibility* of scientific experiments is one of the foundations of science.

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

Reproducible computational science

Modern science has come to rely on computer simulations, computational models, and computational analysis of very large data sets. *Numerical experiments* are relatively new.

As simulations and models become more complex and data sets become larger, calculations that are reproducible in principle are no longer reproducible in practice without public access to the code, data and meta-data.

To insure reproducibility and universality, reports of numerical experimentation should include:

1. All **source code** necessary to reproduce the calculation.
2. All **input data** used to perform the calculation.
3. All **meta-data** required to allow other codes to use the input data

These are equivalent to the methodology section of an experimental paper. This standard requires **Open Source**, **Open Data**, and **Open MetaData** for computational science to be considered reproducible.

Openness in methodology

Open Source and **Open Data** science are important for reproducibility. **Open Notebook** science makes available the entire record of a research project as it is recorded. These encourage public engagement and re-use in other fields.

Availability of data

The sharing of data and materials is not a new concept in science. **Open Data** is the idea that primary scientific data should be available to anyone without restrictions from copyright, patents, or other mechanisms of control. Some data sets are very large (e.g. genomes, meta-genomes, proteomes, and databases of chemical structures).

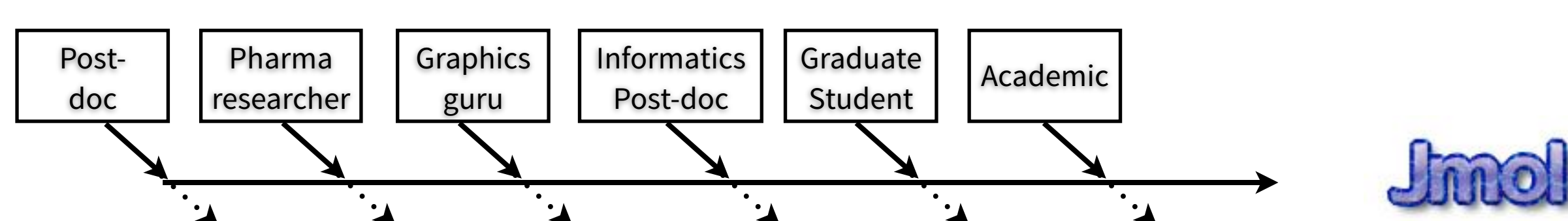
Open Data encourages re-use outside the original field of study, and leads to unexpected discoveries.

Access to results

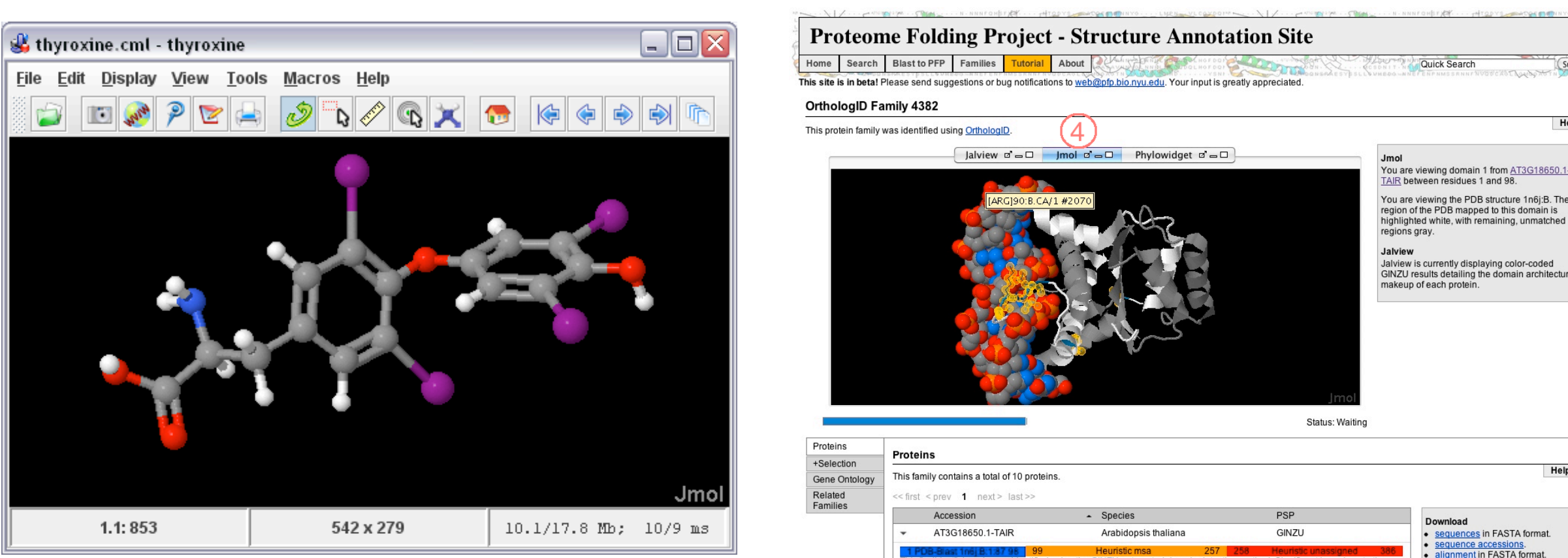
Open Access is the idea that scientists should be releasing their findings in ways that are accessible to all potential users without any barriers.

The recent announcement by the Office of Science and Technology Policy on expanding public access to the results of federally funded research is a welcome development. A mandate by the funding agencies is one of the best ways to make sure this aspect of open science becomes reality.



An OpenScience case study: Jmol



- Filled a void created by the death of a closed-source tool.
- Developed by a series of project leads and their *geographically-distributed teams*. The lead developers *hand off the code* when they become too busy.
- Application focus changed dramatically over 10 years.
- External users of the code tend to *run the application* rather than re-use algorithms.
- Jmol has become the standard tool for embedding chemical structures in web pages:
 - RCSB Protein Data Bank (PDB)
 - Inorganic Crystal Structure Database
 - Viewer for Folding @ Home projects
 - Can be directly included in Sakai, Moodle, WebAssign and LON-CAPA sites
- Jmol is now the structure viewer for *many* academic journals.

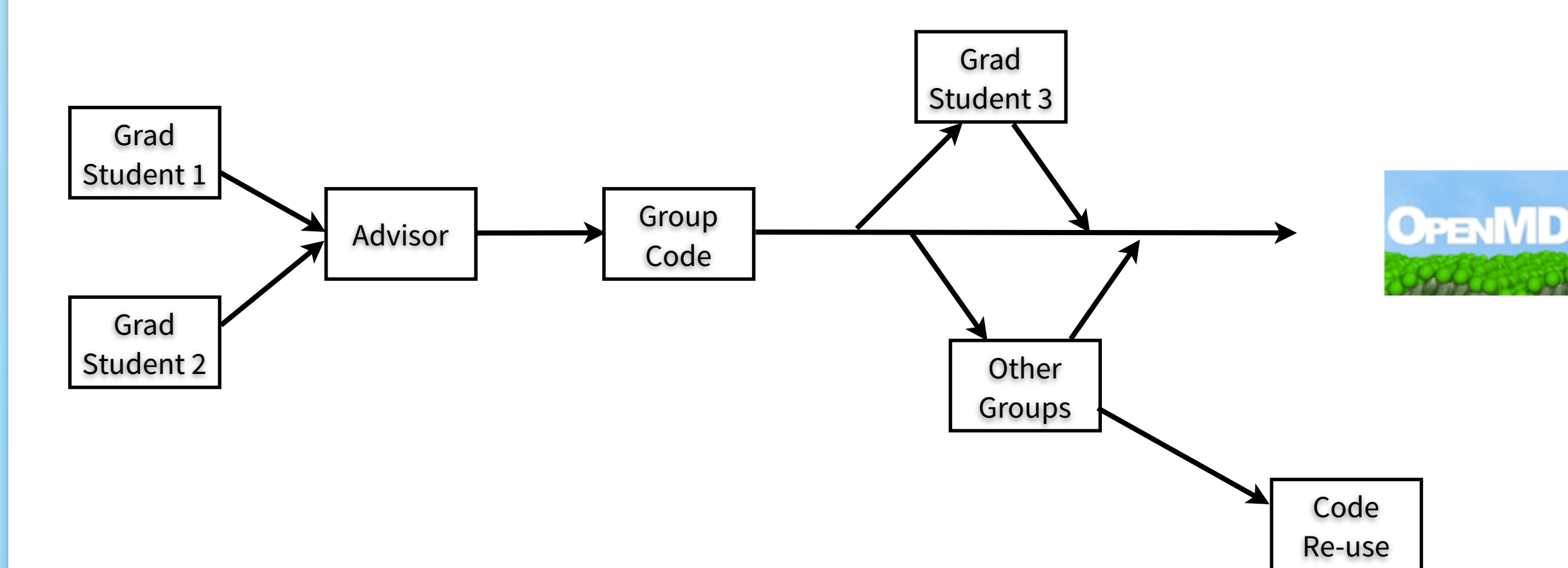


Two successful OpenScience Projects

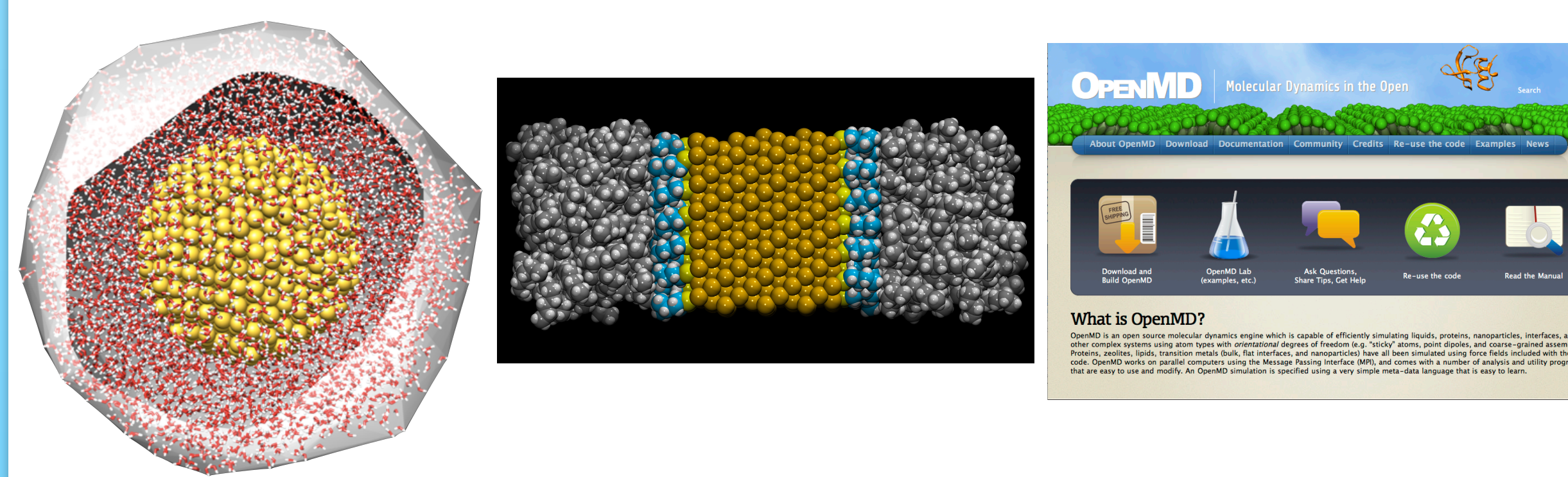
		
Started:	1998	2004
Purpose:	Molecular Visualization	Molecular Dynamics
Languages:	Java	C++, Python
Developers:	36	17 (graduate students)
Lead Developers:	7	1
Code base:	487,836 lines	84,801 lines
Person-Years:	129	21
Estimated Development Costs:	\$7,074,587	\$577,051
Explicit Funding:	\$0	\$0
Downloads:	742,381 at SourceForge, (possibly millions more as an embedded applet)	4,861
External Citations:	201	15
Citations to lead developers:	~30	15

Data from ohloh.net, sourceforge.net, and webofknowledge.com

An OpenScience case study: OpenMD



- Merged student codes that carried out similar tasks.
- Development was done within one research group and was piggy-backed on other funded projects.
- A journal article outlines the code's capabilities, and attribution is requested in the license.
- Application development *preserved group memory*.
- External users of the code tend to *re-use* algorithms rather than run the application.



Science 2.0 - Social media in science

Science 2.0 is a way of describing the increasing collaboration between scientists that is being brought about by social media and the internet. A growing number of scientists are finding ways to communicate their work using wikis, blogs, Twitter, and other social media.

Challenges facing Open Science

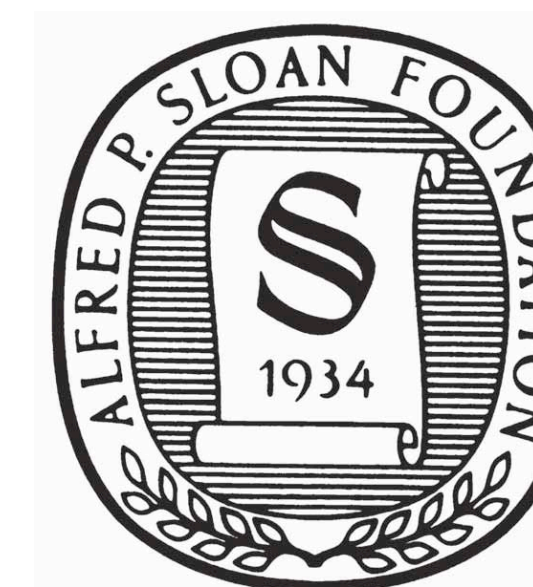
The incentive network that scientists are working under currently favors “closed” science. To make Open Science work, we need to address two big issues:

- **Recognition & Attribution** - Scientific productivity and importance are often measured by:
 - papers in traditional journals with high impact factors
 - citation count

Both of these measures help determine funding and promotions at most institutions, and working on Open Science is either neutral or damaging by these measures. Time spent cleaning up code for release, or setting up a public database is time spent away from writing a proposal or paper. Also, scientists rarely cite “tools” which are the products of Open Science. If we want Open Science to flourish, we should raise our expectations. Research shouldn't be considered *complete* until the data and meta-data is put up on the web for other people to use and until the code is documented and released. Citation of the products of open science (i.e. code and data) should become the norm, not the exception.

- **Sustainability** - Funding for science has traditionally been based on discovery. Open Science requires a bit more attention to the *products* of the research (code, data sets, paper repositories) to complement the initial discoveries. This is going to require discussion about the public value of the products of scientific research.

Acknowledgments



We gratefully acknowledge seed funding for the Open Science Project from the Alfred P. Sloan Foundation.



Development of OpenMD was indirectly supported under NSF grant CHE-0848243.



Current OpenMD team (2013)
OpenScience logo courtesy Kristina Furse Davis