

# ***From Open Source to Open Science***

***J. Daniel Gezelter  
University of Notre Dame  
(also: [openscience.org](http://openscience.org))***

# ***Singh's Law***

Every discussion in science ends up in a discussion on tenure and grants.

--Deepak Singh

[mndoci.com/2010/05/06/singhs-law/](http://mndoci.com/2010/05/06/singhs-law/)

# What is Open Science?

- Open Source
- Open Notebook
- Open Data
- Open Metadata
- Open Peer Review
- Open Access
- Science 2.0

# What is Open Science?

- Open Source
- Open Notebook

*Transparency in experimental methodology, observation, and collection of data.*

- Open Data
- Open Metadata
- Open Peer Review
- Open Access
- Science 2.0

# What is Open Science?

- Open Source
- Open Notebook

*Transparency in experimental methodology, observation, and collection of data.*

- Open Data
- Open Metadata

*Public availability and re-use of scientific data.*

- Open Peer Review
- Open Access
- Science 2.0

# What is Open Science?

- Open Source
- Open Notebook

*Transparency in experimental methodology, observation, and collection of data.*

- Open Data
- Open Metadata

*Public availability and re-use of scientific data.*

- Open Peer Review
- Open Access

*Public accessibility and transparency of scientific communication.*

- Science 2.0

# What is Open Science?

- Open Source
- Open Notebook

*Transparency in experimental methodology, observation, and collection of data.*

- Open Data
- Open Metadata

*Public availability and re-use of scientific data.*

- Open Peer Review
- Open Access

*Public accessibility and transparency of scientific communication.*

- Science 2.0

*Using web-based tools to facilitate scientific collaboration.*

# Reproducibility

***“The statements constituting a scientific explanation must be capable of test by reference to publicly ascertainable evidence.”***

- C. Hempel, *Philosophy of Natural Science* **49** (1966).

Experiments must be reproducible in both a *conceptual* and an *operational* sense. This is one of the cornerstones of the scientific method.

Is it healthy for scientific papers to be supported by computations that *cannot* be verified except by a few employees at a commercial software developer? *Is this even science?*

See: [www.bannedbygaussian.org](http://www.bannedbygaussian.org)

# Reproducibility & Open Source Software

- Modern science relies to a very large degree on computer simulations, computational models, and computational analysis of large data sets.
- These methods for doing science all have underlying theoretical assumptions that are *reproducible in principle*. For very simple systems, this is nearly the same as *reproducible in practice*.
- As systems become more complex, calculations that are reproducible in principle are no longer reproducible in practice without *public access to the code, data, and meta-data*.
- *It is therefore imperative for skeptical scientific inquiry that software for simulating complex systems be available in source-code form.*

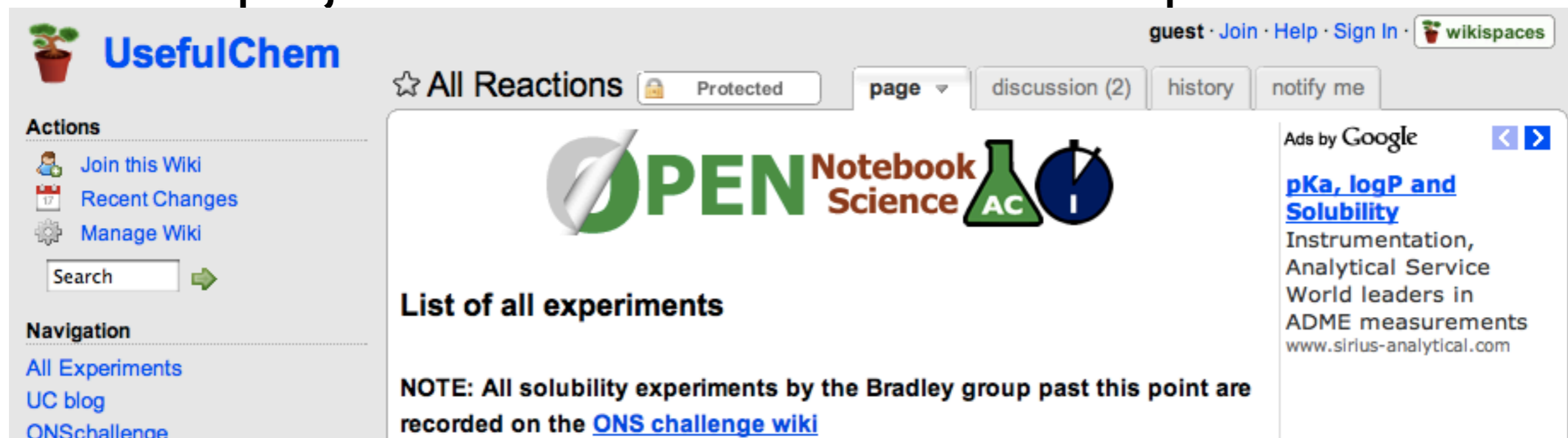
# Two Open Source Science Codes

	Jmol	OpenMD
Started:	1998	2004
Purpose:	Molecular Visualization	Molecular Dynamics Simulations
Languages:	Java	C++, F95, Python
Developers:	29	9 (mostly graduate students)
Lead Developers:	5	1
Code base:	365,465 lines	91,373 lines
Person-Years:	95	23
Estimated Development Costs:	\$5,234,000	\$545,000
Explicitly-funded Costs:	\$0	\$0
Downloads:	Over 350,000 at SourceForge alone, (possibly millions more)	29,000
External Citations:	13	3
Citations to lead developers:	3	3

**The entries in red point out some *major* problems.**

# Open Notebook

Open Notebook science makes available the entire record of a research project as it is recorded. Some examples:



The screenshot shows a wiki page on UsefulChem. The page title is "All Reactions" and it is marked as "Protected". The main content area features the "OPEN Notebook Science" logo, which includes a green circle with a white pencil tip, the word "OPEN" in green, "Notebook Science" in brown, and icons of a green flask labeled "AC" and a blue flask labeled "I". Below the logo is the heading "List of all experiments" and a note: "NOTE: All solubility experiments by the Bradley group past this point are recorded on the [ONS challenge wiki](#)". The left sidebar contains "Actions" (Join this Wiki, Recent Changes, Manage Wiki) and "Navigation" (All Experiments, UC blog, ONSchallenge). The top right shows user options (guest, Join, Help, Sign In) and a "wikispaces" logo. An advertisement for Sirius Analytical is visible on the right.



The screenshot shows a lab log entry on a website. The header includes a "Login" link and "Dashboard | Help" links. The Science & Technology Facilities Council (ISIS) logo is prominently displayed. The main title is "Cameron's LaBLog" with the subtitle "The online open laboratory notebook of Cameron Neylon". The entry is titled "Freeze dried Sortase" and dated "21st April 2010 @ 09:37". The "Material" section lists "Powder" and describes it as "The freeze dried Sortase from [Concentration and exchange of sortase](#)". The "Total weight of material" is listed as "141 mg". A search bar with a "Find" button and an "Archives" section are also visible on the right side.

# Open Notebook

Publicly searchable lab protocols:

The screenshot shows the OpenWetWare website interface. At the top right, there is a "Log in" link. Below it are navigation tabs for "page", "talk", "view source", and "history". The main header features a DNA double helix logo and the text "OPENWETWARE". A central banner reads: "OpenWetWare is an effort to promote the sharing of information, know-how, and wisdom among researchers and groups who are working in biology & biological engineering. **Learn more about us.** If you would like edit access, would be interested in helping out, or want your lab website hosted on OpenWetWare, please join us."

Below the banner is a horizontal menu with four items: "Labs & Groups" (From around the world), "Courses" (Host & view classes), "Protocols" (Share techniques & more), and "Blogs" (Read OWW blogs).

The left sidebar contains several sections: "navigation" with links to Main Page, Recent changes, Help, Contact OWW, and Add a Lab Notebook; "research" with links to Materials, Protocols, and Resources; "search" with a search box and buttons for Search, Go, and a help icon; and "toolbox" with links for What links here, Related changes, Upload file, Special pages, and Printable version, along with sub-links for Permanent link, Cite this page, and Subscribe to.

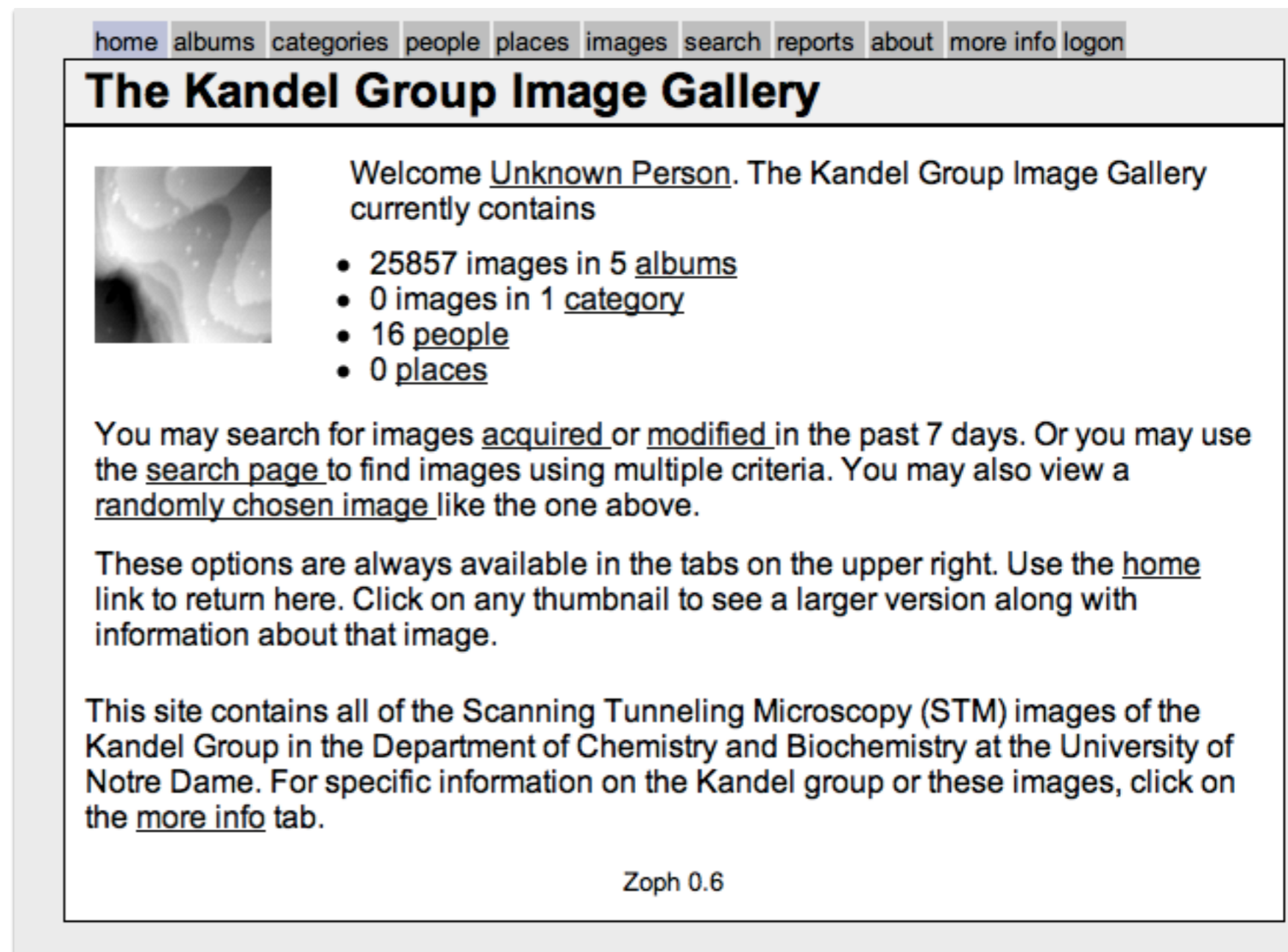
The main content area is divided into two columns. The left column features a "Welcome 2010 iGEM Teams" banner with the iGEM logo. Below it is a section for "OpenWetWare Lab Notebooks" with an image of a notebook and a pencil. A yellow "New! One-click setup" badge is present. The text lists new features: "Dynamic calendars" (Create or view entries with a click), "Local search" (Search within your lab notebook), and "Improved navigation" (Jump between entries with ease). Below this is a "Welcome new OWW users!" section listing names: Meysam Bastami, Amro Mentash, Zhiqiang Lu, Lillian Seu, James F Southern, Fabio FR Vicente, Lucas Hartsough, Jacob Trueb, and Dr.

The right column features an "OWW Community Blog" section with an RSS icon. It includes a sub-section "Referencing a DOI Within OpenWetWare" with the text: "To lookup an article using a document object identifier, there's a cheap and cheerful way to do it based upon the work we did earlier to add access to pubget." Below this is another sub-section "OpenWetWare: Where DOI Begin(s)?" with the text: "MediaWiki is the software that OpenWetWare.org is built on. We customize it by applying our own styling to the page, add our own member management software to it, and either write our own extensions to it or download and install others." A third sub-section "Reading Articles Referenced Within OpenWetWare" includes the text: "To paraphrase Sigmund Freud, 'Sometimes a reference is just a reference.' But not all the time. By the time we're reading an

# Open Data

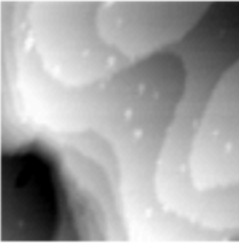
The idea that primary scientific data should be available to *anyone* without restrictions from copyright, patents, or other mechanisms of control.

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*.



home albums categories people places images search reports about more info logon

## The Kandel Group Image Gallery



Welcome Unknown Person. The Kandel Group Image Gallery currently contains

- 25857 images in 5 albums
- 0 images in 1 category
- 16 people
- 0 places

You may search for images acquired or modified in the past 7 days. Or you may use the search page to find images using multiple criteria. You may also view a randomly chosen image like the one above.

These options are always available in the tabs on the upper right. Use the home link to return here. Click on any thumbnail to see a larger version along with information about that image.

This site contains all of the Scanning Tunneling Microscopy (STM) images of the Kandel Group in the Department of Chemistry and Biochemistry at the University of Notre Dame. For specific information on the Kandel group or these images, click on the more info tab.


Zoph 0.6

# Open Access


Publishing scientific findings in such a way that the findings of a study are accessible to all potential users without any barriers.



**DIRECTORY OF  
OPEN ACCESS  
JOURNALS**



**SPARC  
EUROPE  
AWARD  
2009**



**For Outstanding  
Achievements  
in Scholarly  
Communications**  
**SPARC**  
The Scholarly Publishing and Academic Resources Coalition

**Find Journals**  
**New titles**  
**Find articles**  
**Suggest a journal**

**About**  
**FAQ**  
**News**  
**Links**  
**Sponsors**  
**Long term archiving**  
**Membership**  
**Feedback**

**For journal owners**  
**For authors**

Today's visitors  
Total 4943

Welcome to the Directory of Open Access Journals. This service covers free, full text, quality controlled scientific and scholarly journals. We aim to cover all subjects and languages. There are now **5033** journals in the directory. Currently **2070** journals are searchable at article level. As of today **396349** articles are included in the DOAJ service.

**Support the development and operation of DOAJ. Sign up for membership - go to the [membership page](#).**  
**We are very thankful for the support from those of you who have already decided to become DOAJ members. See the [list of members](#)**

**Browse by title**

[A](#) [B](#) [C](#) [D](#) [E](#) [F](#) [G](#) [H](#) [I](#) [J](#) [K](#) [L](#) [M](#) [N](#) [O](#) [P](#) [Q](#) [R](#) [S](#) [T](#) [U](#) [V](#) [W](#) [X](#) [Y](#) [Z](#)

**Browse by subject**

**Expand Subject Tree**

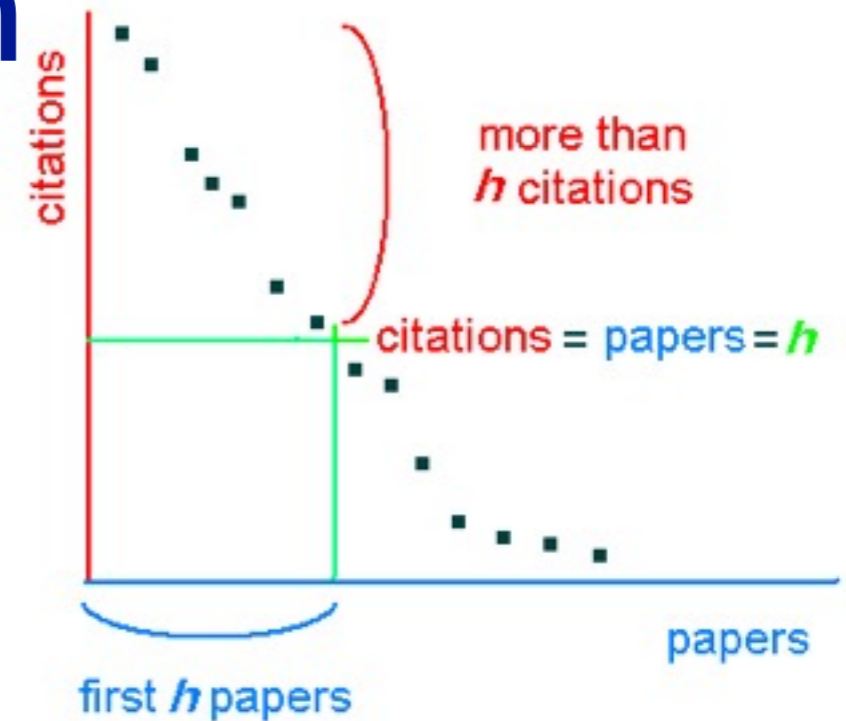
- [Agriculture and Food Sciences](#)
- [Arts and Architecture](#)
- [Biology and Life Sciences](#)
- [Business and Economics](#)
- [Languages and Literatures](#)
- [Law and Political Science](#)
- [Mathematics and Statistics](#)
- [Philosophy and Religion](#)

# ***Things we'll have to figure out before Open Science is a workable model:***

- Recognition & Attribution
- Copyright & Licensing
- Sustainability

# Recognition & Attribution

- Scientists stay alive professionally by publishing.
  - Paper count
  - Citation count
  - $h$ -index
- What happens if time and effort exerted in pursuit of open science projects reduces the ability to publish?
- *Why aren't open projects treated like publications?* How do you cite an online STM image database? A blog? An open source project? How does that citation get tied to a particular researcher?



## **Attribution metrics *should* (but *don't*) take into account:**

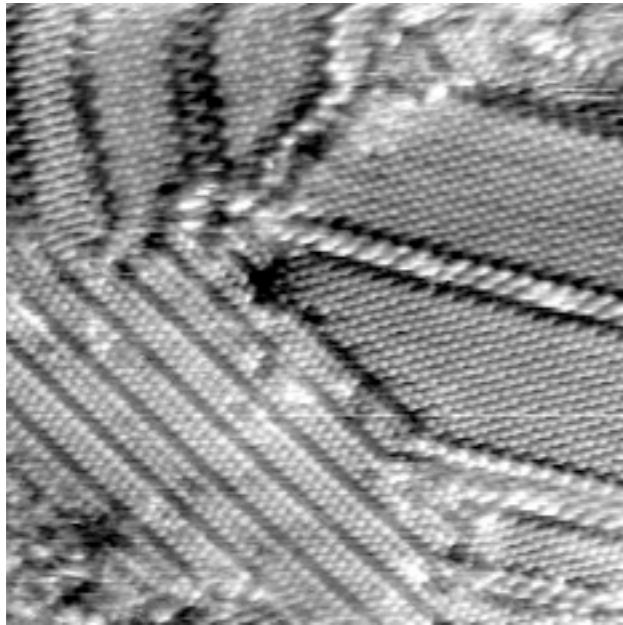
1. Amount of scientific effort,
2. Complexity of the work,
3. Importance of the work to the scientific community,
4. Externalities of the work beyond the scientific community.

# Recognition & Attribution

*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.*

The screenshot shows the OpenID website homepage. At the top, there is a navigation bar with links for 'Get an OpenID', 'Add OpenID to your site', 'Developers', 'Foundation', 'Community', and 'Government'. The OpenID logo is in the top right corner. Below the navigation bar is a large blue banner with the text: 'Can't remember your passwords? Tired of filling out registration forms? OpenID is a **safe, faster, and easier** way to log in to web site'. Below the banner are three columns of content. The first column is titled 'feedback' and contains the text: 'You can [get an OpenID](#) from several popular and well-known providers or host your own. We'll help you make a smart decision.' Below this text is a button labeled 'Get an OpenID'. The second column is titled 'EVERYONE' and contains the text: 'Chances are you already have an OpenID but don't even know it! No problem — we'll walk you through [using your OpenID](#) for the first time.' Below this text is a button labeled 'Start using your OpenID'. The third column is titled 'SITE OW' and contains the text: 'Run a web site and want to make it easier to sign up or sign in? Find out how to [add OpenID to your site](#) in a few steps by using free, open source library.' Below this text is a button labeled 'Add OpenID to your site'. At the bottom of the page, there is a yellow banner with the text: 'From the blog: [Authenware Joins OpenID Foundation](#)'.

# Copyright & Licensing



```
Algorithm 0.0.1: REDUCE(projection, x, y, f)  
  
for i ← 1 to y/f  
  for j ← 1 to x/f  
    sum ← 0  
    for m ← 1 to f  
      for n ← 1 to f  
        sum = sum + projection[i * f + m][j * f + n]  
      reducedProjection[i][j] = sum / (f * f)  
    return (reducedProjection)
```

- If you publish an algorithm in a society journal (even in a condensed or pseudo-code form), is it OK to release that code under the GPL?
- Do online lab image archives violate journal copyrights?
- Is it OK to reprint an excerpt of your paper in your blog?

# Sustainability

From the Strategic Content Alliance (SCA) report: “*Sustainability and Revenue Models for Online Academic Resources*” by Kevin Guthrie, Rebecca Griffiths, Nancy Maron:

*We 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.*

Why is sustainability so difficult? PIs in academia have experience: 1) doing research, 2) writing papers, 3) teaching, and 4) securing funding. This expertise is quite different from what is required of a leader of “service enterprises”.

Their suggested models for Sustainability:

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

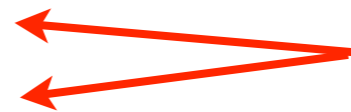
[www.jisc.ac.uk/contentalliance](http://www.jisc.ac.uk/contentalliance)

# Sustainability

Can these mechanisms work to support OpenSource science software?  
Certainly not all of them:

## 1. Direct beneficiaries pay

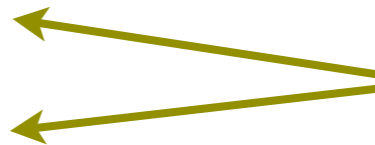
- a. *Subscription or one-time payment*
- b. *Pay per use*
- c. *Contributor pays*



*Direct barriers to verifiability*

## 2. Indirect beneficiaries pay

- a. *Host institution's support*
- b. *Corporate sponsorships*
- c. *Advertising*
- d. *Philanthropic funding*
- e. *License content*



*Current models. Not working well!*



*Potential conflicts of interest*



*Sparse, and cannot be relied upon*



*Who would license it?*

Our attempts to define sustainability models for scientific software projects:

1. Sell something physical (i.e. a spectrometer)
2. Sell services (i.e. support for a complicated program)
3. Sell advertising
4. Dual-license (i.e. academic vs. commercial software licensing)
5. Differentiate between single-run and high-throughput versions
6. Philanthropic funding
7. Anything Else?

# Sustainability

**Most of the grant-funding agencies are mission-driven.**

NIH, DOE, DARPA all fund specific scientific projects. There is little room for projects which may enrich the overall scientific enterprise but which generate little data themselves. One agency is an exception: NSF

**What are the broader impacts of the proposed activity?**

How well does the activity advance discovery and understanding while promoting teaching, training, and learning? How well does the proposed activity broaden the participation of underrepresented groups (e.g., gender, ethnicity, disability, geographic, etc.)? To what extent will it enhance the infrastructure for research and education, such as facilities, instrumentation, networks, and partnerships? Will the results be disseminated broadly to enhance scientific and technological understanding? What may be the benefits of the proposed activity to society?

# Outlook

**It is vital that we do these things, but I'm not particularly sanguine about the outlook for Open Science. I do have some advice:**

- Have multiple project “leads”.
- Think about how other researchers can cite your work.
- Think about how your institution will view your work.
- Use an appropriate widely-recognized license.
- Don't expect your project to be sustainable in the current funding climate.
- Any solution to these problems is going to have to work *within* the established behaviors of the various communities.

# Acknowledgments



**The Alfred P. Sloan Foundation**  
Startup funding for the OpenScience Project



**PhDs.org**  
Geoff Davis' site on issues related to graduate school  
Geoff was an early collaborator on OpenScience.