When decade-old functionality would be progress - the desolate state of our scholarly infrastructure

Björn Brembs
Universität Regensburg
http://brembs.net
Life change
yesterday

today
Institutions produce publications, data and software
CRISIS I

Dysfunctional scholarly literature
• Limited access
• No global search
• No functional hyperlinks
• No flexible data visualization
• No submission standards
• (Almost) no statistics
• No text/data-mining
• No effective way to sort, filter and discover
• No scientific impact analysis
• No networking feature
• etc.

…it’s like the web in 1995!
CRISIS II

Scientific data in peril
Small Data – Long Tail

Data on Disks and in Drawers

(2) Risk that supplements to articles turn into Data Dumping places

(1) Top of the pyramid is stable but small

(3) Too many disciplines lack a community endorsed data archive

(4) Estimates are that at least 75% of research data is never made openly available

PubMed has been designated to be maintained with minimal staff during the lapse in government funding. The information on this website will be kept as up to date as possible, and the agency will attempt to respond to urgent operational inquiries during this period.

Updates regarding government operating status and resumption of normal operations can be found at http://www.usa.gov.
Databases fight funding cuts

Online tools are becoming ever more important to biology, but financial support is unstable.

Monya Baker

05 September 2012
SWISS-PROT should have been 10 years old in July 1996, but it may disappear on June 30, 1996

Due to funding problems, SWISS-PROT as well as PROSITE, and the ENZYME nomenclature databases will disappear on June 30, 1996 if no solution is found before that date. The ExPASy WWW server and all services associated with it will also shut down. The distribution of the SWISS-2DPAGE database will also be discontinued. Other external databases, WWW services and software packages that depend on SWISS-PROT,
Dear TAIR user community,

To help us through the current funding crisis, we recently established a new TAIR corporate sponsorship program. We feel that this approach is preferable to implementing a subscription requirement for the private sector because it will allow us to keep TAIR open and free of login requirements, facilitating the free exploration of data by all scientists. Two companies (Dow AgroSciences and Syngenta) and one research organization (Gregor Mendel Institute) have already become TAIR sponsors. More information can be found on our sponsorship page.

TAIR Funding Crisis (10/16/2009)

Dear TAIR user community,
CRISIS III

Non-existent software archives
Replicable Research in Computational Science

Roger D. Peng

To whom correspondence should be addressed. E-mail: rpeng@ihpsh.edu

ABSTRACT

Computational science has led to exciting new developments, but the nature of the work has exposed limitations in our ability to evaluate published findings. Reproducibility has the potential to serve as a minimum standard for judging scientific claims when full independent replication of a study is not possible.
The case for open computer programs

Darrel C. Ince, Leslie Hatton & John Graham-Cumming

Affiliations | Contributions | Corresponding author

*Nature* 482, 485–488 (23 February 2012) | doi:10.1038/nature10836
Received 09 May 2011 | Accepted 05 January 2012 | Published online 22 February 2012

Scientific communication relies on evidence that cannot be entirely included in publications, but the rise of computational science has added a new layer of inaccessibility. Although it is now accepted that data should be made available on request, the current regulations regarding the availability of software are inconsistent. We argue that, with some exceptions, anything less than the release of source programs is intolerable for results that depend on computation. The vagaries of hardware, software and natural language will always ensure that exact reproducibility remains uncertain, but withholding code increases the chances that efforts to reproduce results will fail.
Today’s Digital Dystopia

- Institutional email
- Institutional webspace
- Institutional blog
- Library access card
- Open access repository
- No archiving of texts
- No archiving of code
- No archiving of data
Consequences?

Graph showing the number of retractions per 10k publications over the years 2000 to 2010. The graph indicates an increasing trend in retractions over time.
Only read publications from high-ranking journals
Only publish in high-ranking journals
METRICS

Is journal rank like astrology?
The Impact Factor

Introduced in 1950’s by Eugene Garfield: ISI

\[ IF(\text{year } 3) = \frac{C_{12}}{A_1 + A_2} \]
The Impact Factor

Introduced in 1950’s by Eugene Garfield: ISI

\[
 IF(\text{year 3}) \quad = \quad \frac{100}{40+60} = 1
\]
The Impact Factor

Journal X IF 2013 =

All citations from TR indexed journals in 2013 to papers in journal X

Number of citable articles published in journal X in 20011/12

€30,000-130,000/year subscription rates
Covers ~11,500 journals (Scopus covers ~16,500)
Main Problems with the IF

- Negotiable
- Irreproducible
- Mathematically unsound
• PLoS Medicine, IF 2-11 (8.4)  
  http://www.plosmedicine.org/article/info:doi/10.1371%2Fjournal.pmed.0030291)

• Current Biology IF from 7 to 11 in 2003  
  – Bought by Cell Press (Elsevier) in 2001...
Journal: CURRENT BIOLOGY

<table>
<thead>
<tr>
<th>Mark</th>
<th>Journal Title</th>
<th>ISSN</th>
<th>Total Cites</th>
<th>Impact Factor</th>
<th>Immediacy Index</th>
<th>Citable Items</th>
<th>Cited Half-life</th>
<th>Citing Half-life</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CURR BIOL</td>
<td>0960-9822</td>
<td>20020</td>
<td>7.007</td>
<td>2.713</td>
<td>341</td>
<td>3.5</td>
<td>3.7</td>
</tr>
</tbody>
</table>

Journal Information

- Full Journal Title: CURRENT BIOLOGY
- JCR Abbrev. Title: CURR BIOL
- ISSN: 0960-9822
- Issues/Year: 24
- Language: ENGLISH
- Journal Country/Territory: UNITED STATES
- Publisher: CELL PRESS
- Publisher Address: 1100 MASSACHUSETTS AVE, CAMBRIDGE, MA 02138
- Subject Categories: BIOCHEMISTRY & MOLECULAR BIOLOGY

Journal Rank in Categories

Journal Impact Factor

- Cites in 2002 to items published in: 2001 = 3314
- Number of items published in: 2001 = 528
- 2000 = 3917
- 2000 = 504
- Sum: 7231
- Sum: 1032
- Calculation: Cites to recent items = 7231 = 7.007
- Number of recent items = 1032
### Journal: CURRENT BIOLOGY

<table>
<thead>
<tr>
<th>Mark</th>
<th>Journal Title</th>
<th>ISSN</th>
<th>Total Cites</th>
<th>Impact Factor</th>
<th>Immediacy Index</th>
<th>Citable Items</th>
<th>Cited Half-life</th>
<th>Citing Half-life</th>
</tr>
</thead>
<tbody>
<tr>
<td>☐</td>
<td>CURR BIOL</td>
<td>0960-9822</td>
<td>20020</td>
<td><strong>7.007</strong></td>
<td><strong>2.713</strong></td>
<td>341</td>
<td>3.5</td>
<td>3.7</td>
</tr>
</tbody>
</table>

#### Journal Impact Factor

Cites in 2002 to items published in: 2001 = 3314  Number of items published in: 2001 = 528

2000 = 3917  2000 = 504

Sum: 7231  Sum: 1032

Calculation: Cites to recent items 7231 = **7.007**

Number of recent items 1032
### Journal: CURRENT BIOLOGY

#### 2002

<table>
<thead>
<tr>
<th>Mark</th>
<th>Journal Title</th>
<th>ISSN</th>
<th>Total Cites</th>
<th>Impact Factor</th>
<th>Immediacy Index</th>
<th>Citable Items</th>
<th>Cited Half-life</th>
<th>Citing Half-life</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CURR BIOL</td>
<td>0960-9822</td>
<td>20020</td>
<td>7.007</td>
<td>2.713</td>
<td>341</td>
<td>3.5</td>
<td>3.7</td>
</tr>
</tbody>
</table>

#### Journal Impact Factor

Calculations:
- Cites in 2002 to items published in 2001: 3314
- Number of items published in 2001: 528
- Cites to recent items: 7231
- Sum: 7231

#### 2003

<table>
<thead>
<tr>
<th>Mark</th>
<th>Journal Title</th>
<th>ISSN</th>
<th>Total Cites</th>
<th>Impact Factor</th>
<th>Immediacy Index</th>
<th>Citable Items</th>
<th>Cited Half-life</th>
<th>Citing Half-life</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CURR BIOL</td>
<td>0960-9822</td>
<td>22589</td>
<td>11.910</td>
<td>2.683</td>
<td>331</td>
<td>3.8</td>
<td>4.0</td>
</tr>
</tbody>
</table>

#### Journal Impact Factor

Calculations:
- Cites in 2003 to items published in 2002: 3628
- Number of items published in 2002: 334
- Cites to recent items: 7551
- Sum: 7551

### Current Biology - 2002 Impact Factor

- **Total Cites**: 2020
- **Citable Items**: 341
- **Cited Half-life**: 3.5
- **Citing Half-life**: 3.7
- **Impact Factor**: 7.007
- **Immediacy Index**: 2.713

### Calculation:
- Cites to recent items: 7231
- Number of recent items: 1032

### Journal Impact Factor - 2003

- **Total Cites**: 22589
- **Citable Items**: 331
- **Cited Half-life**: 3.8
- **Citing Half-life**: 4.0
- **Impact Factor**: 11.910
- **Immediacy Index**: 2.683

### Calculation:
- Cites to recent items: 7551
- Number of recent items: 634
<table>
<thead>
<tr>
<th>Year</th>
<th>Impact Factor</th>
<th>Cites in Year</th>
<th>Items Published</th>
<th>Calculation</th>
</tr>
</thead>
<tbody>
<tr>
<td>2002</td>
<td>7.007</td>
<td>2002 - 2001</td>
<td>528</td>
<td>Cites to recent items: 7231 Sum: 1032</td>
</tr>
</tbody>
</table>

**Journal: CURRENT BIOLOGY**

**ISSN:** 0960-9822

**Total Cites:**

**Number of recent items:**
- 2002: 7231
- 2003: 7551
• Rockefeller University Press bought their data from Thomson Reuters
• Up to 19% deviation from published records
• Second dataset still not correct


“My question is: Are we making an impact?”
Not Mathematically Sound

- Left-skewed distributions
- Weak correlation of individual article citation rate with journal IF

Seglen PO (1997): Why the impact factor of journals should not be used for evaluating research. BMJ 1997;314(7079):497 [http://www.bmj.com/cgi/content/full/314/7079/497](http://www.bmj.com/cgi/content/full/314/7079/497)
Journal Rank and Methodology

Journal Rank and Fraud/Error

A. Fraud or Suspected Fraud

B. Error

C. Plagiarism or Duplicate Publication

D. Mean IF by Cause

Fang et al. (2012): Misconduct accounts for the majority of retracted scientific publications. PNAS 109 no. 42 17028-17033
Journal Rank and Retractions

Journal rank is a figment of our imagination.
INCENTIVES

“High-Impact” journals attract the most unreliable research
HOW SCIENCE GOES WRONG

Britain’s angry white men
How to do a nuclear deal with Iran
Investment tips from Nobel economists
Junk bonds are back
The meaning of Sachin Tendulkar

OCTOBER 19TH - 25TH 2013

Economist.com Worldwide cover
“Do you trust scientists?”
“Who can you trust these days?”
The costs of legacy publishing

The costs of legacy publishing

Status Quo
Status Quo
SO MUCH FOR THAT

The disaster that is our digital infrastructure
WHAT NOW?

Science, tear down this paywall!
UTOPIA 8.535 km
Potential for innovation

1. International Coordination
2. Hire software developers
3. Cancel all subscriptions
Superior Alternative

• Sustainable, global search and access for all literature, code and data
• Intelligent sort, filter and discover functionalities
• Scientific, evidence-based reputation system
• Authoring tools for collaborative writing and single-click submission
• Orders of magnitude cheaper: US$90/paper (e.g. SciELO) vs. US$5,000/paper (subscription)
One person is not an institutional infrastructure
Software to control the experiment and save the data
Software to analyze and visualize the data
<table>
<thead>
<tr>
<th>Name</th>
<th>Modified</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>other(blueprints)</td>
<td>2012-09-17</td>
<td></td>
</tr>
<tr>
<td>Analysis</td>
<td>2012-09-17</td>
<td></td>
</tr>
<tr>
<td>Tracker</td>
<td>2011-11-30</td>
<td></td>
</tr>
<tr>
<td>data examples</td>
<td>2011-04-11</td>
<td></td>
</tr>
</tbody>
</table>

Totals: 23 Items
778.9 MB
Publication and Citation of Scientific Software with Persistent Identifiers

Software has become an integral part of science, yet software is not properly integrated into the scientific discourse. We will look at the requirements for software publication, code archives, and persistent identification of software.
Segmentation of dynamic PET images with kinetic spectral clustering

Mouyssset, S ; Zbib, H ; Stute, S ; Girault, JM; et al

Segmentation is often required for the analysis of dynamic positron emission tomography (PET) images. However, noise and low spatial resolution make it a difficult task and several supervised and unsupervised methods have been proposed in the literature ...

Uploaded by yannic on 05 March 2014.
Open Source Tracking and Analysis of Adult Drosophila Locomotion in Buridan's Paradigm with and without Visual Targets

Julien Colomb, Lutz Reiter, Jędrzej Błaszkiewicz, Jan Wessnitzer, Björn Brembs

1 FB Biologie, Chemie, Pharmazie, Institut für Biologie-Neurobiologie, Freie Universität Berlin, Berlin, Germany, 2 Institute for Perception, Action and Behaviour, School of Informatics, University of Edinburgh, Edinburgh, United Kingdom, 3 Department of Genetics, Universität Leipzig, Leipzig, Germany

Abstract

Background

Insects have been among the most widely used model systems for studying the control of locomotion by nervous systems. In Drosophila, we implemented a simple test for locomotion: in Buridan’s paradigm, flies walk back and forth between two inaccessible visual targets [11]. Until today, the lack of easily accessible tools for tracking the fly position and analyzing its trajectory has probably contributed to the slow acceptance of Buridan’s paradigm.

Methodology/Principal Findings

To add a note, highlight some text. Hide notes

Make a general comment

Jump to

Abstract
Introduction
Materials and Methods
Results
Discussion
Supporting Information
Acknowledgments
Author Contributions
References

Download: PDF | Citation | XML
Print article
Eurexpress New & Improved!

Metrics

Total Article Views: 1,144
Published on August 9, 2012

Citations
Social Networks

No citations found
Facebook (12)
Twitter (3)
Mendeley (8)

Related Content

Related Subject Categories
Biology and life sciences, Computer and information sciences, Engineering and technology, Physical sciences

Related Articles on the Web
Google Scholar
PubMed

Share this Article

Email this article
Features

1GB of private space
Upload all formats

Unlimited public space
Quick & simple upload

Publish negative data
Cloud based

Upload to your heart's content

API
Desktop uploader

Collaborative spaces
Installing

Stable version available

Development Version  API Keys

```r
install.packages("figshare")
# If you would like to install a development version:
library(devtools):
install_github("figshare", "ropensci")
```

Quick start guide

```r
library("figshare")
```

For a full list of functions and a web manual, visit the package repository on GitHub.

Tutorials

Tutorials coming shortly

Use cases & resources

If you have ideas for use-cases or have written about this package anywhere, please drop us a line.
Add new, or update an existing article

```r
# brems lab account
options(FigshareKey = "A"

options(FigsharePrivateKey = "b"

options(FigsharePrivatekey = "c"

options(FigshareToken = "d"

options(FigsharePrivatetoken = "e"

### end figshare info

require(rfigshare)

fs_auth()

## need to create the article and get its id here: do it only once, then write the id and comment this part:

if (is.na(id_test)) {
  # article_title = "d42Gal4 UPKCI experiment: with and without UASGFP and chaGal80"
  # article_description = "this is a temporary article"
  # article_type = "figure"  #, "dataset" #, "media", "poster", "paper", "fileset"
  # article_tags = c("self-learning", "genetics")
  # article_categories = "Neuroscience"
  # article_files = "T: dataforfigshare.png"
  # article_visibility = "draft" # "private", "public"
  # article_authors = c("julien colomb")
  # article_links = "http://lab.brems.net"

  id <- fs_new_article(title = article_title, description = article_description,
    type = article_type, tags = article_tags, categories = article_categories,
    files = article_files, visibility = article_visibility, #authors = article_authors,
    links = article_links)

  # add bjorn as author (the "0" leads to error on figshare at this time):
  rfigshare:::fs_add_author(article_id = id_test, author_id = 96464)

  id

} else {
  newfile = "T: dataforfigshare.png"
  fs_upload(id_test, file = newfile)

}  ```
Run your script and...

Same type of experiments → same script
Default: → same categories
→ same tags
→ same authors
→ same links
→ same description
→ One complete article, in one click.

Update the figure:
Higher sample size directly published while analysed, your boss may see the results before you do! (or you may see the results of your student before they do)

Possibility to make it public and citable in one click or directly in the R code.
Citable?

DataCite

http://dx.doi.org/10.6084/m9.figshare.97792
PKC53e putative mutant in self learning

Tags
- self-learning
- genetics
- figsharetests
- FBal0197029

Description
This is a demonstration for figshare. This post was 100% produced in R and uploaded to figshare afterwards. This is the first step toward an automatic upload of our data on Figshare...

Export
- Export to Ref. Manager
- Export to Endnote
- Export to Mendeley
Sub-strains of Drosophila Canton-S differ markedly in their locomotor behavior [vi; ref status: indexed, http://f1000r.es/3is]

Julien Colomb1, Bjorn Brembs2

Grant information: The author(s) declared that no grants were involved in supporting this work.

Abstract
We collected five sub-strains of the standard laboratory wild type Drosophila melanogaster Canton Special (CS) and analyzed their walking behavior in Burdun's paradigm using the CtrAio software. According to twelve different aspects of their behavior, the sub-strains fall into three groups. The group separation appeared not to be correlated with the origin of the stocks. We conclude that founder effects but not laboratory selection likely influenced the gene pool of the sub-strains. The flies' stripe fixation was the parameter that varied most. Our results suggest that differences in the genome of laboratory stocks can render comparisons between nominally identical wild-type stocks meaningless. A single source for control strains may settle this problem.

Corresponding author: Bjorn Brembs

Copyright: © 2014 Colomb J and Brembs B. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. Data associated with the article are available under the terms of the Creative Commons Zero "no rights reserved" data waiver (CC0 1.0 Public domain dedication).

Competing interests: No competing interests were disclosed.
Figure 3. The different sub-strains show a large spectrum of values for the stripe deviation parameter.

For every movement of the fly, the angle between its direction and the direction toward the stripes was calculated. The median of these angles was calculated for each fly, representing a quantification of stripe fixation by the fly. The value of each sub-strain in each session is depicted in biplots; for each group, we represent the median, 25–75% quantiles and the total spread of the values (excluding outliers) as line, box and whiskers, respectively. The version of this figure on their Research website is interactive, allowing one to define the type of whiskers displayed as either the 10th–90th percentiles (A) or ‘Tukey’ whiskers (0.6 x IQR from 10th quantile).

The subset of genes is analogous to that used in Figure 2. The red horizontal line corresponds to the median value for random walks. 

Sample size is 11–12 for each borplot. No statistical analysis was performed.
Figure 3. The different sub-strains show a large spectrum of values for the stripe deviation parameter.

For every movement of the fly, the angle between its direction and the direction toward the stripes was calculated. The median of these angles was calculated for each fly, representing a quantification of stripe fixation by the fly. The values of each sub-strain in each session were aggregated for each genotype, we represent the median, 25–75% quantiles and the total spread of the values (excluding outliers) as a line, box and whiskers, respectively. The version of this figure on the FluoGraph research website is interactive; readers can define the type of whiskers displayed as either the 10th–90th percentiles (A) or Tukey whiskers (1.5 x IQR from 1st/3rd quartile). (B). The red horizontal line corresponds to the median values for random walks: 44°. Sample size is 11–12 for each bar plot. No statistical analysis was performed.
Figure 3. The different sub-strains show a large spectrum of values for the stripe deviation parameter.

For every movement of the fly, the angle between its direction and the direction toward the stripe was calculated. The median of those angles was calculated for each fly, representing a quantification of stripe fixation by the fly. The value of each sub-strain in each session is depicted in boxplots: for each group, we represent the median, 25–75th percentiles and the total spread of the values (excluding outliers) as line, box and whiskers, respectively. The version of this figure on the Fr000Research website is interactive: readers can define the type of whiskers displayed as either the 10th–90th percentiles (A) or Tukey whiskers (1.5 x IQR from 10th/90th quartile). The box color code used in the panels is analogous to that used in Figure 2. The red horizontal line corresponds to the median value to random walks, 44°. Sample size is N=12 for each boxplot. No statistical analysis was performed.