

The How, When, Where, and Why of Open Data

2015-10-20

Ross Mounce: Natural History Museum, London
&

Dace Rozenberga: University of Coventry



orcid.org/0000-0002-3520-2046

orcid.org/0000-0002-2724-8706

Outline

- What is 'data'?
- The evolution of data availability
- Where are we now?
- Some goals & aspirations for the future
- Questions & Answers

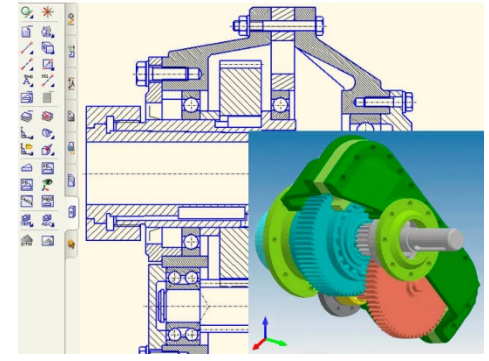
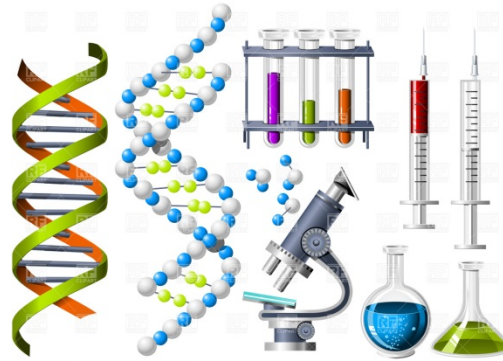
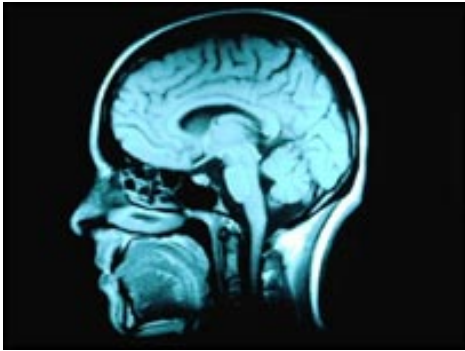
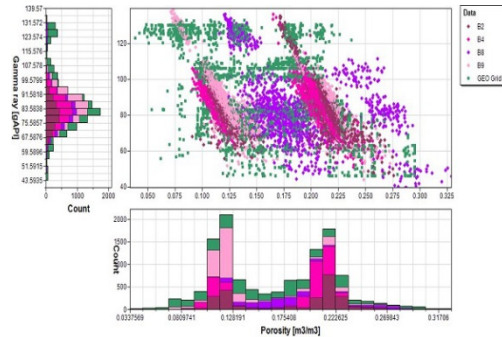


Figure 2: A screenshot of a data analysis software window showing a spreadsheet with columns labeled A through T. The data includes numerical values and percentages, likely representing experimental results or material properties.

Sample	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
General Total	2.8	1.5	1.8	1.2	1.0	1.1	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0	2.1	2.2	2.3	2.4	2.5	2.6
%	100.00%	53.57%	64.29%	42.86%	35.71%	39.29%	46.43%	50.00%	53.57%	57.14%	60.71%	64.29%	67.86%	71.43%	75.00%	78.57%	82.14%	85.71%	89.29%	92.86%



Definition of research data

- Research data is defined as **recorded factual material** commonly retained by and accepted in the scientific community as **necessary to validate research findings.** ([EPSRC](#))
- Research data: The **data, records, files or other evidence**, irrespective of their content or form (e.g. in print, digital, physical or other forms), that **comprise research observations, findings or outcomes**, including primary materials and analysed data. ([Monash University](#))

Definition of research data

- Research data is defined as **recorded factual material** commonly retained by and accepted in the scientific community as **necessary to validate research findings**. ([EPSRC](#))
- Research data: The **data, records, files or other evidence**, irrespective of their content or form (e.g. in print, digital, physical or other forms), that **comprise research observations, findings or outcomes**, including primary materials and analysed data. ([Monash University](#))

Disclaimer

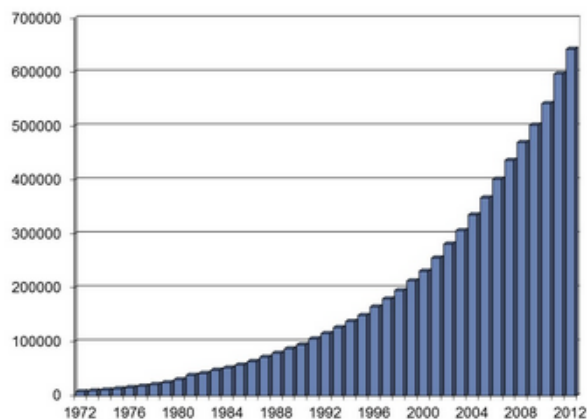
Summarising the state of open data is **HARD**



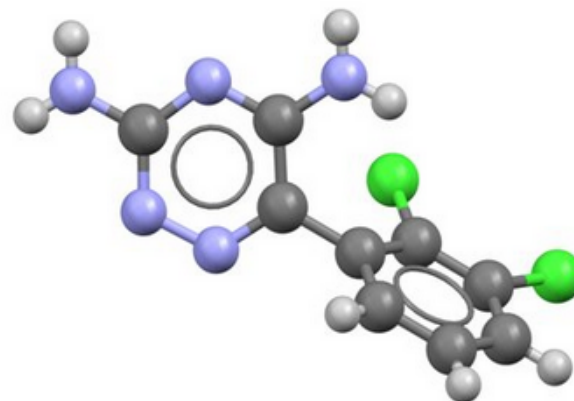
Centralised Data Centres

The Cambridge Crystallographic Data Centre, **est. 1965**

It maintains the Cambridge Structural Database **



Growth of the Cambridge Structural Database (CSD) since 1970.



Lamotrigine (EFEMUX01) - the 500,000th structure archived to the CSD

** Not open data *sensu stricto* ...but I'll leave that to Peter Murray-Rust to explain

Data Sharing (by snail mail)

e.g. “The full profile listings are on **floppy disks** which are available upon request”

Fernholz *et al* (1989) A survey of measurements and measuring techniques in rapidly distorted compressible turbulent boundary layers.



1989

Submission of data to GenBank

CHRISTIAN BURKS AND LAURIE J. TOMLINSON

Theoretical Biology and Biophysics Group T-10, MS K710, Los

In response to both the ever-increasing rate of de nucleotide sequences (1) and the growing tren journals to allow articles to appear that describe tl of determining a sequence without explicitly prese sequence (1), GenBank* (2-5) and a number of the that publish nucleotide sequence data are working to promote the direct, timely submission of nucleotide data to GenBank. The policy being establish PROCEEDINGS is described in the editorial on p. 407; will provide a brief summary, in the context of this

“The author will provide the accession number to the PROCEEDINGS [PNAS] office to be **included in a footnote to the published paper.**”

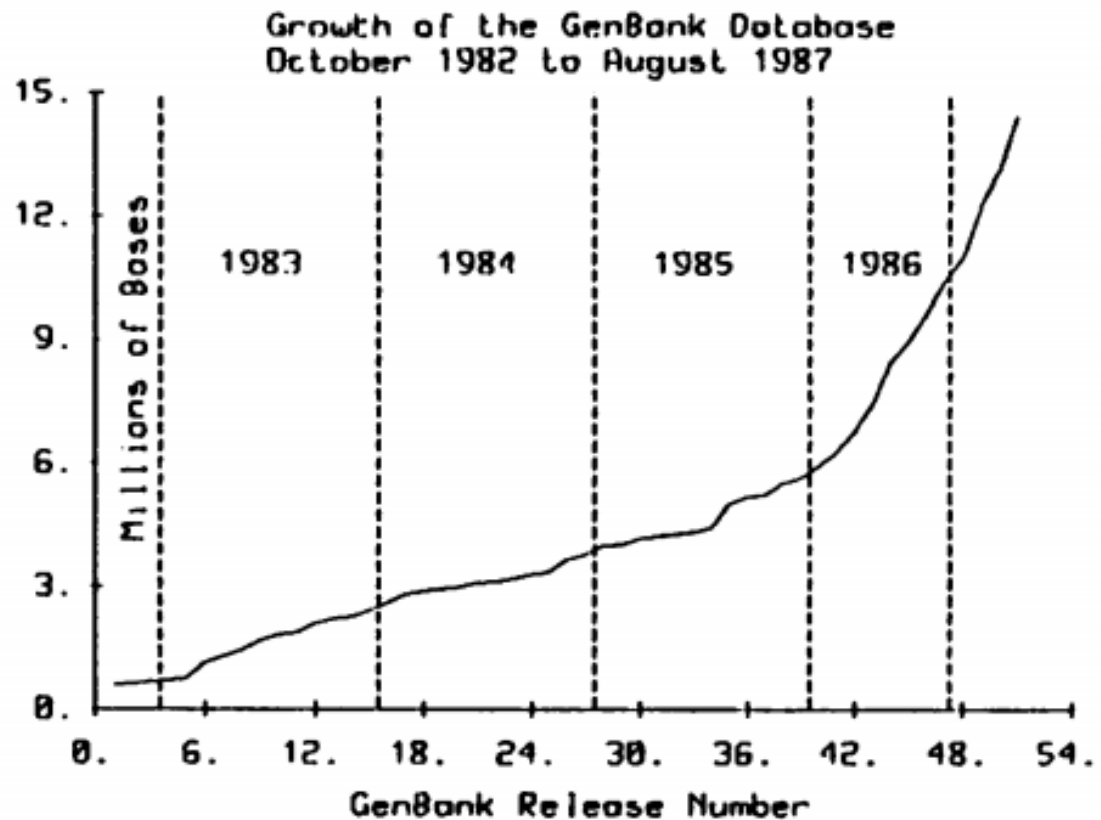


Figure 1.

Bilofsky & Burks (1988)
Nucleic Acids Research v16 n5

Reproducible research

Jon Claerbout,

Jon Buckheit & David Donoho, 1995

WaveLab and Reproducible Research

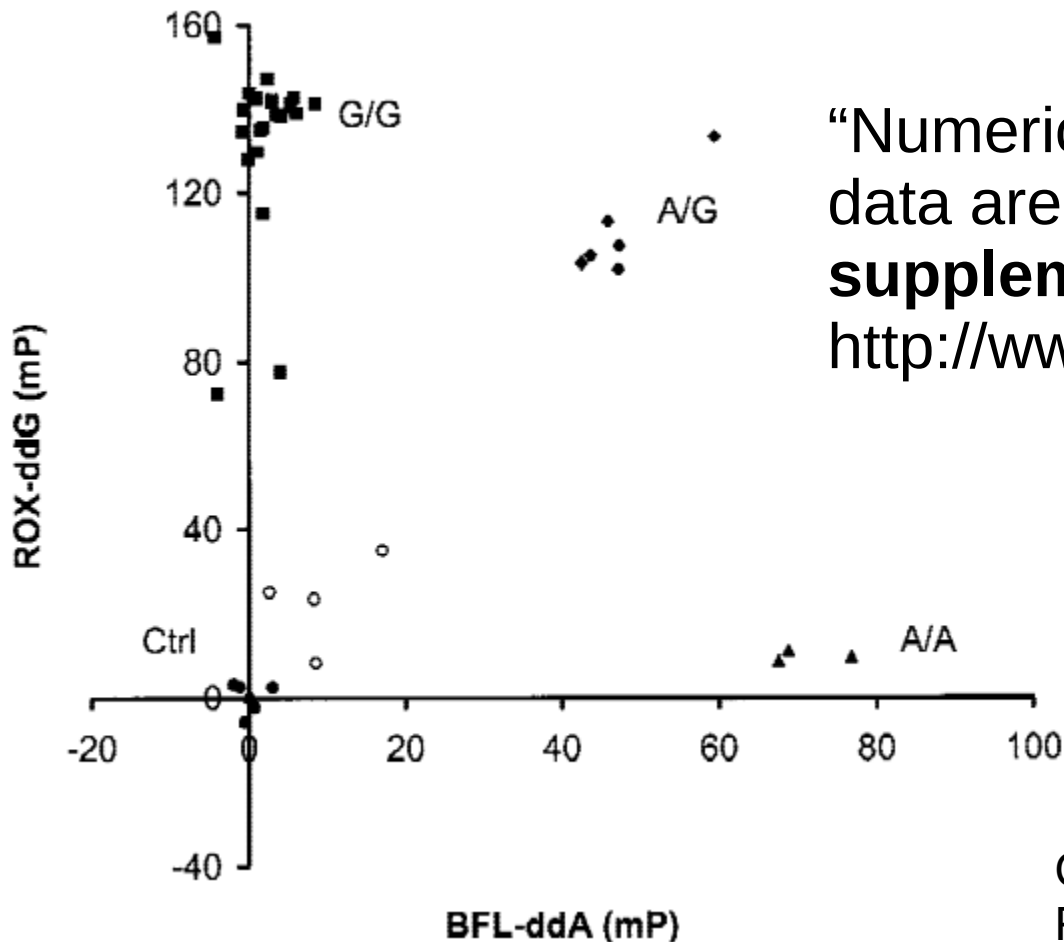
Jonathan B. Buckheit and David L. Donoho

Stanford University, Stanford CA 94305, USA

Abstract

WAVELAB is a library of MATLAB routines for wavelet analysis, wavelet-packet analysis, cosine-packet analysis and matching pursuit. The library is available free of charge over the Internet. Versions are provided for Macintosh, UNIX and Windows machines.

Supplementary Data (Online)



“Numerical values for the data are available as **online supplementary material** at <http://www.genome.org>.”

Chen et al (1999)
Fluorescence Polarization in
Homogeneous Nucleic Acid
Analysis. Genome Research



Not all databases succeed.
Build it, and they may not come...

<http://treebase.org/>



A Database of Phylogenetic Knowledge

The logo for TRY, consisting of the letters 'TRY' in a bold, green, sans-serif font with a white outline, set against a white rectangular background.

Plant Trait Database

Each custodian of data on plant traits will retain the right to be informed of any TRY activity that may involve his/her data, and will have the opportunity to negotiate whether his/her data can be used, and whether general guidelines of authorship need to be modified in that particular case

Custodians retain the rights to withdraw their data at any time.

Your data is NOT 'too big' to share

Data released on October 06, 2014

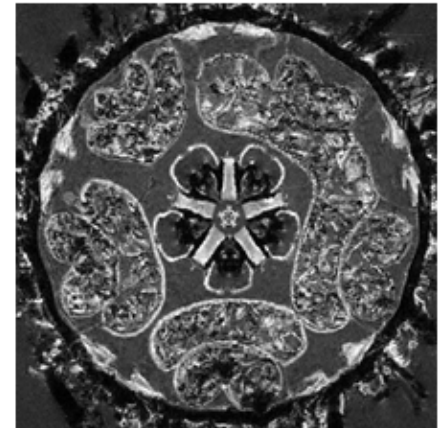
MRI scans of whole sea urchin specimens.

Ziegler, A; Faber, C; Mueller, S; Nagelmann, N; Schröder, L (2014): MRI scans of whole sea urchin specimens. GigaScience Database. <http://dx.doi.org/10.5524/100124> [RIS](#) [BibTeX](#) [TEXT](#)

Magnetic resonance imaging (MRI) is routinely used in human diagnostics, but can also be applied to study the internal anatomy of zoological specimens. Here, we present 141 MRI scans from 98 representative extant sea urchin species. The scanned specimens were whole sub-adult or adult individuals ranging in size from 5 to 43 mm. The specimens were almost entirely obtained from museum collections. Some of the samples were collected and fixed more than 135 years ago, while others were collected a few months prior to scanning. The detailed MRI acquisition and reconstruction parameters can be found in the metadata files deposited together with the raw image data. In addition, image stacks in tagged image file format (TIFF, .tif) were generated for each scan. Potential uses of the dataset include morphometric and volumetric analyses or comparative studies of internal organs.

For convenience users can choose to down load all 141 datasets in a single TAR archive file called [141_MRI_scans.tgz](#) (39GB) from the FTP server.

Imaging



39 Gigabytes (GB)
of MRI scans

(GIGA)ⁿ
SCIENCE

<http://gigadb.org/dataset/100124>

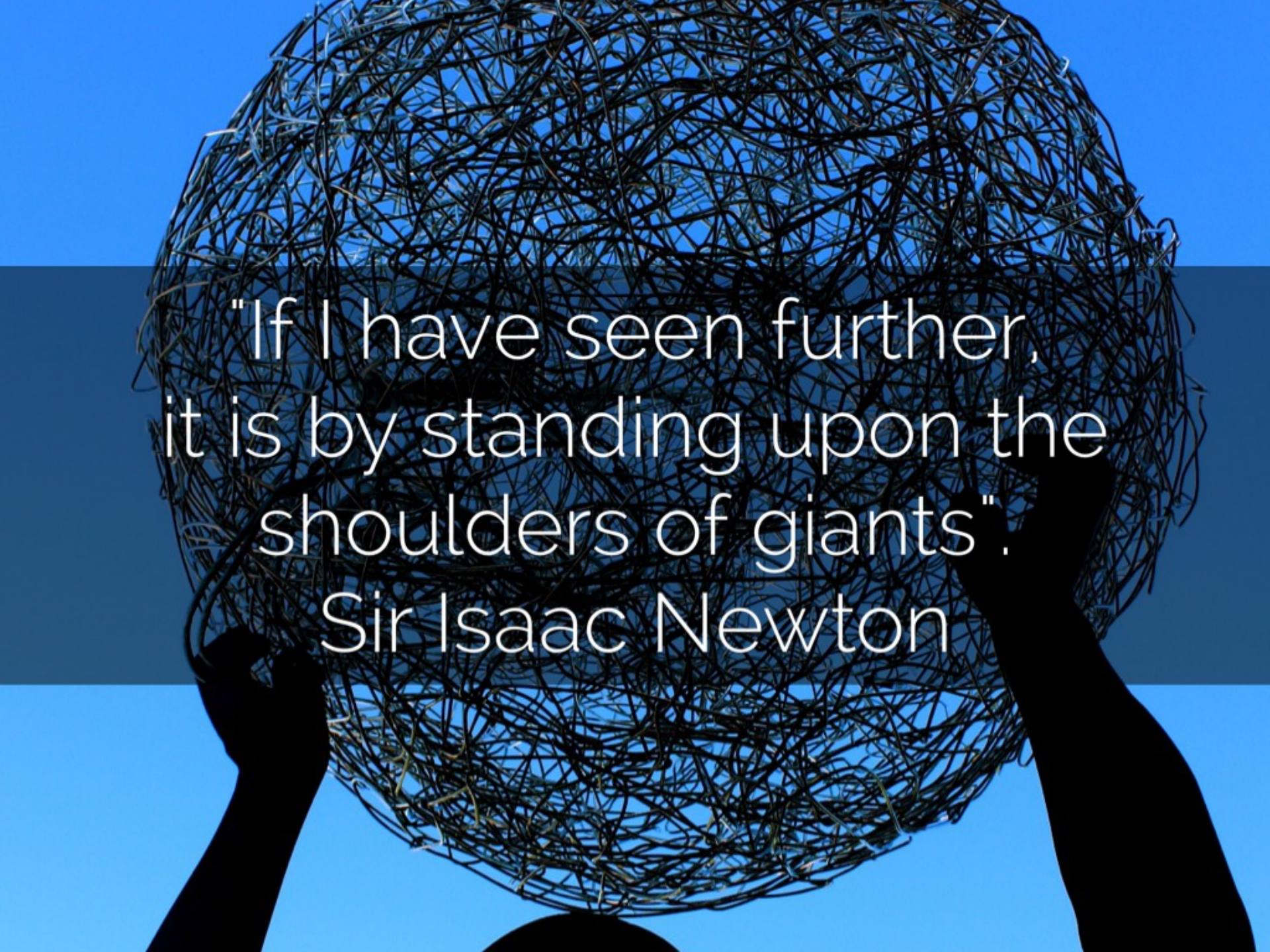
Why care about data?

Data is absolutely **fundamental** to most research

Science without data isn't science *

I 
DATA

*Entirely theoretical, data-free contributions to science are possible, but rare



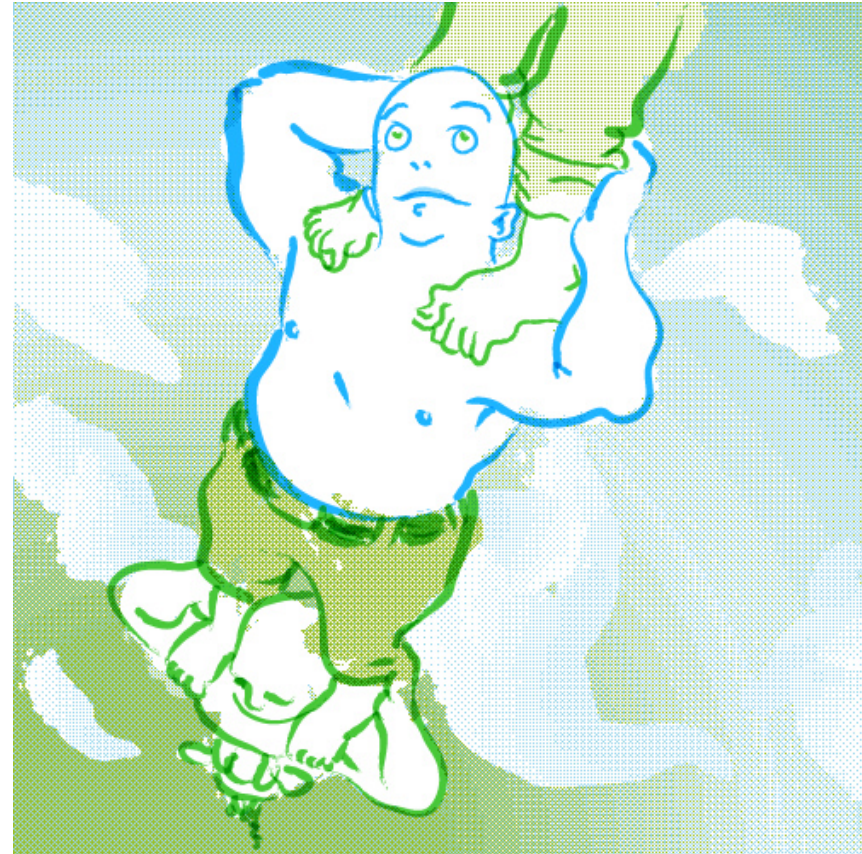
"If I have seen further,
it is by standing upon the
shoulders of giants".
Sir Isaac Newton

By sharing data we can see further

Data (& code) are the building blocks of science

Shared, re-used data allow us to more rigorously test hypotheses; “to see further”

...and to do it all more quickly and easily.



What exactly is *open* data?

Open means **anyone** can **freely access, use, modify,** and **share for any purpose** (subject, at most, to requirements that preserve provenance and openness)

From <http://opendefinition.org/>,
see <http://opendefinition.org/od/> for more detail



Legally, what is *open* data?

There are a great many open knowledge definition (OKD) conformant licences, including:



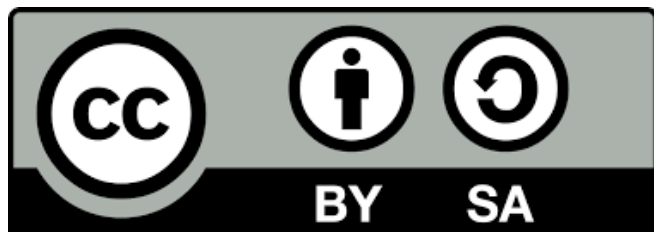
CC0

<http://creativecommons.org/publicdomain/zero/1.0/>



CC BY

<https://creativecommons.org/licenses/by/4.0/>



CC BY-SA

<https://creativecommons.org/licenses/by-sa/4.0/>



See here for the comprehensive list: <http://opendefinition.org/licenses/>

Not all Creative Commons licences are 'open'



NC -- You “may not use this work for commercial purposes”.
Work under this licence cannot be used for *any* purpose, therefore it is *not* open.
Can have significant, often unexpected negative impact on potential re-use.

ND -- “No Derivative Works”.
Work under this licence cannot be adapted if it is re-used. Not very helpful for research!

NC & ND – An extremely restrictive re-use licence, neither commercial purposes nor adaptations are allowed.

Non-open licencing causes real problems for research & education

The Creative Commons non-commercial (-NC) restriction is poorly defined in most jurisdictions, and even more poorly understood by many of its users.

“non-commercial” != “non-profit”

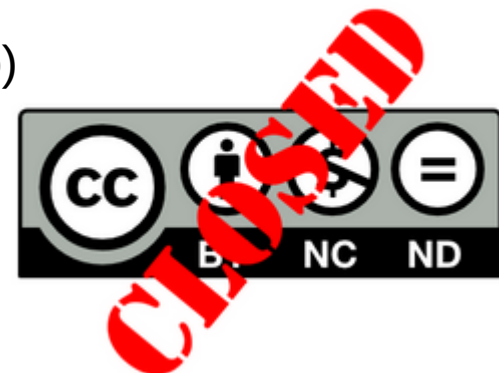
A) Non-commercial actually excludes many **teaching purposes**:
In the UK, university students typically pay expensive tuition fees to attend. Thus **university teaching is often a commercial activity**, -NC restricted materials cannot be used to teach students in these circumstances.

B) **Licence incompatibility** – NC licences are not compatible with licences used on major collaboration platforms like Wikipedia or Wikimedia Commons

C) **Non-commercial organizations** (e.g. Deutschlandradio) have been successfully sued for re-using CC BY-NC content without permission.

<http://zookeys.pensoft.net/articles.php?id=3036>

<https://www.techdirt.com/articles/20140326/11405526695/german-court-says-creative-commons-non-commercial-licenses-must-be-purely-personal-use.shtml>



Real problems of non-open data: GBIF & biodiversity data



Desmet, P. (2013) Showing you this map of aggregated bullfrog occurrences would be illegal <http://peterdesmet.com/posts/illegal-bullfrogs.html>

Open data in scholarship, and beyond

The open data movement is much broader than just academia/research

It's been successful & popular in areas like open *government* data:

For **transparency**, detecting & discouraging corruption

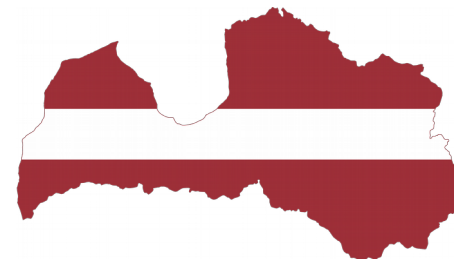
For **releasing social & commercial value** (governments collect already, why not make wider use of it, at little or no extra cost)

For **participatory governance** – citizens can be more informed, a “read/write” society

Each of these has clear parallels with open research data: transparency & fraud detection, extra value through research data re-use, participatory citizen science



Some text adapted from <http://opengovernmentdata.org/>



The State of Open Data in Latvia: 2014

Uldis BOJĀRS¹ and Renārs LIEPIŅŠ

Institute of Mathematics and Computer Science, University of Latvia,
Raina bulvaris 29, Riga, LV-1459, Latvia

`uldis.bojars@gmail.com, renars.liepins@lumii.lv`

Abstract. This paper examines the state of Open Data in Latvia at the middle of 2014. The study is divided into two parts: (i) a survey of open data situation and (ii) an overview of available open data sets. The first part examines the general open data climate in Latvia according to the guidelines of the OKFN Open Data Index making the results comparable to those of other

[arXiv:1406.5052v2](https://arxiv.org/abs/1406.5052v2)

	Transport Timetables	Government Budget	Government Spending	Election Results	Company Register	National Map	National Statistics	Legislation	Location datasets	Pollutant Emissions	Procurement tenders	Water Quality	Land Ownership	Health performance	Weather forecast	Total Score
32	2014		2014			2014				2014	(1)	(1)	(1)	(1)	(1)	570
33	2014	1	2014	1	2014	2014	1	2014	1		(1)	(1)	(1)	(1)	(1)	570
35	2014	1	2014	1	2014	2014	1	2014	1		(1)	(1)	(1)	(1)	(1)	565
36	2014	1	2014	1	2014	2014	1				(1)	(1)	(1)	(1)		545
37	2014	2014	2014	2014	2014	2014	2014	2014	2014	2014	+	+	+	+	+	540
38		2014	2014	1	2014	2014	1						(1)		(1)	535
39	1		2014	1	2014	2014	1		1			1	1	1	1	530
40	2014	2014	2014	2014	2014	2014	2014	2014	2014	2014		+	+	+	+	520
41	2013	2013	2013	2013	2013	2013	2013	2013	2013	2013	+	+	+	+	+	515

Open (government) Data
<http://global.census.okfn.org/>

The arts and culture sector must think about data ... but differently

From exhibition directories to museum floorplans, data can augment the audience experience and offers new approaches to solving problems



<http://www.theguardian.com/culture-professionals-network/culture-professionals-blog/2014/mar/28/arts-culture-sector-data-impact>

Open data in scholarship, and beyond

Similarly, and with some overlap to open research data, there's the open GLAM movement

(GLAM = Galleries, Libraries, Archives & Museums)

In this case, their data is typically collections metadata but also digital images of their collections



O P E N
G L A M

See <http://openglam.org/> for more

Explore and download the Natural History Museum's research and collections data.

3.5M
records

28
datasets


13
contributors

Search the Natural History Museum Specimen Collection

2,834,363 of the Museum's 80 million specimens are now available online.



 1,173,882
Zoology

 622,677
Botany

 391,811
Mineralogy

 364,109
Palaeontology

 281,884
Entomology

Technical aspects of open data

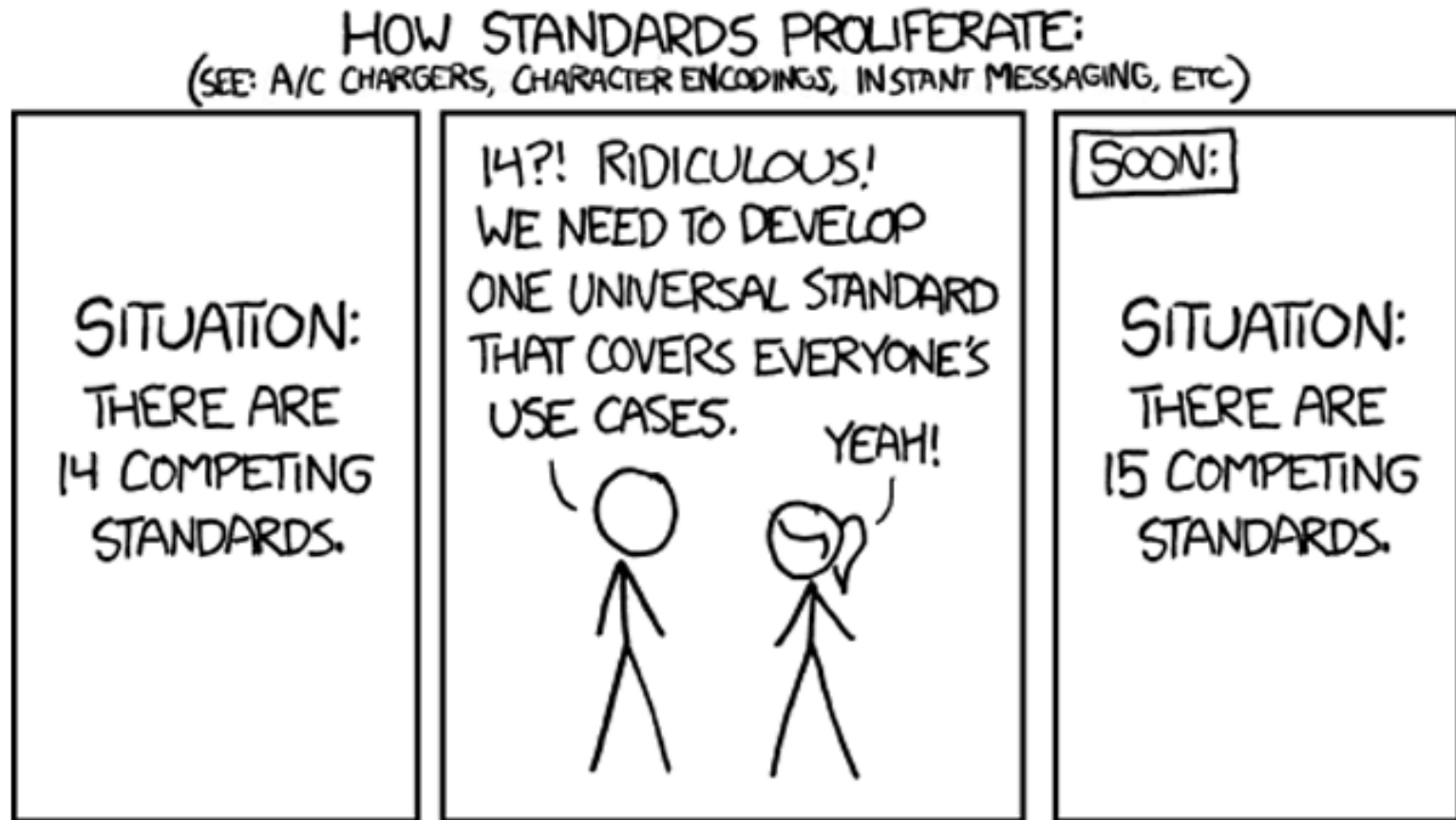
So, you understand the importance of licensing...
What next?

How best can we make our data openly available?

Where should I upload to?

What format(s) should I make the data available in?

Data Standards & Data File Formats



Adhere to existing standards!

Data Standards & Data File Formats

Take note of community standards:

e.g. the **Bermuda Principles** for sharing DNA seq. data

- Automatic release of sequence assemblies larger than 1 kb (preferably within 24 hours).
- Immediate publication of finished annotated sequences.
- Aim to make the entire sequence freely available in the public domain



Data Standards & Data File Formats

If there are no formally agreed community standards, canvas the community to create/formalise a standard

e.g. Best Practices for Data Sharing in Phylogenetic Research
(2014) [PLOS Currents Tree of Life](#)

e.g. The [1st Open Economics International Workshop](#)
(Cambridge, 2013) bringing together academic economists from around the world to discuss data sharing in economics research.



Open Economics

Data Standards & Data File Formats

If there are multiple, competing file formats:

Opt for file formats based on open standards

https://en.wikipedia.org/wiki/Open_standard

e.g.



Avoid proprietary formats

https://en.wikipedia.org/wiki/Proprietary_format

e.g.





please don't hurt the web
use open standards

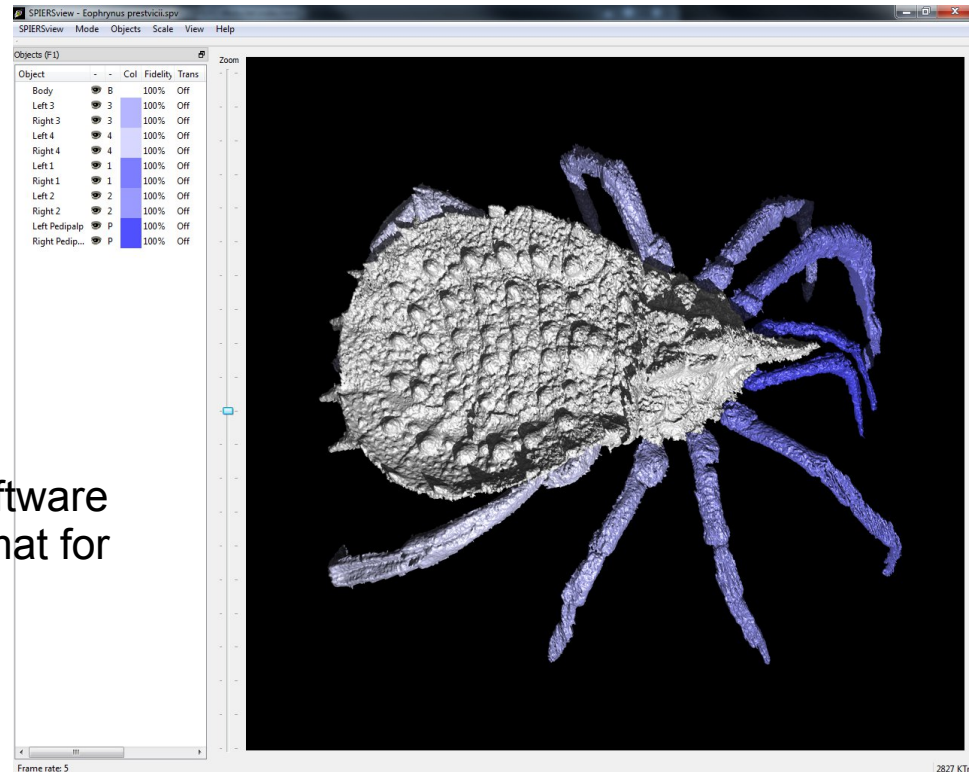
Data Standards & Data File Formats

A real example: recent creation of a new data standard for exchange of 3-dimensional reconstruction of objects from tomographic imaging data

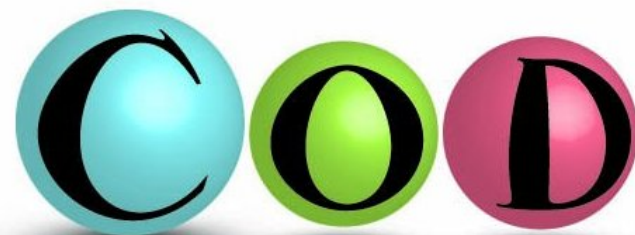
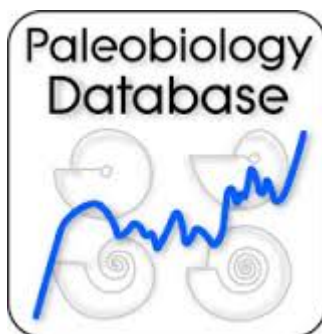
SPIERS software
+ VAXML data standard

Sutton et al (2012) SPIERS and VAXML: A software toolkit for tomographic visualisation and a format for virtual specimen interchange.

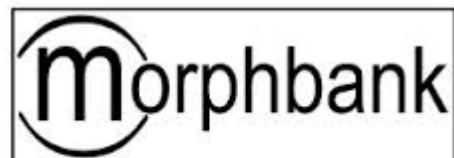
Palaeontologia Electronica



Where to upload open data?



<http://www.crystallography.net/>

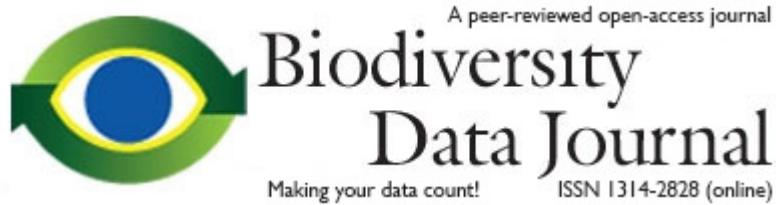


<http://morphbank.net>



Genbank,
SRA,
1000's more!

'Data paper' journals



Journal of
open archaeology data



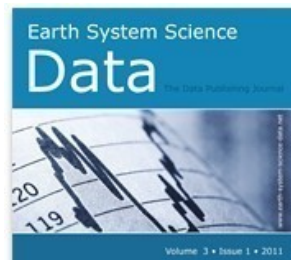
F1000Research
AN OPEN SCIENCE JOURNAL FOR LIFE SCIENTISTS



Journal of
open psychology data



data



open health data

<http://www.mdpi.com/journal/data/about>

Intelligent data papers allow databases to automatically pull-in your data



Biodiversity Data Journal 2: e1076 (26 Mar 2014)
doi: 10.3897/BDJ.2.e1076



Taxonomic paper

Dispatch from the field: ecology of ground-web-building spiders with description of a new species (**Araneae**, **Symphytognathidae**)

▼ Jeremy A. Miller, Menno Schilthuizen, Jennie Lilliendahl Burmester, Lot van der Graaf, Vincent Merckx, Merlijn Jocqué, Paul Joseph Antonius Kessler, Tom Maurice Fayle, Thijmen Breeschoten, Regi Broeren, Roderick Bouman, Wan-Ji Chua, Frida Feijen, Tanita Fermont, Kevin Groen, Marvin Groen, Nicolaas Johannes Cornelis Kil, Henrica Allegonda de Laat, Michelangelo Sergio Moerland, Carole Moncoquet, Elisa Panjang, Amelia Joyce Philip, Rebecca Roca-Eriksen, Bastiaan Rooduijn, Marit van Santen, Violet Swakman, Meaghan N. Evans, Luke J. Evans, Kieran Love, Sarah H Joscelyne, Anya Victoria Tober, Hannah F. Wilson, Laurentius N. Ambu, Benoit Goossens

Many publishers (e.g. Pensoft) intelligently markup data papers so that the data can be automatically ingested into appropriate db's on the day of publication!





Biodiversity Data Journal 3: e6351 (05 Oct 2015)
doi: 10.3897/BDJ.3.e6351



PDF

XML

Materials

[Download as CSV](#)

Holotype:

a. class: **Insecta**; order: **Diptera**; family: **Heleomyzidae**; genus: **Neossos**; specificEpithet: tombstonensis; scientificNameAuthorship: Solecki & Wheeler, 2015; continent: North America; country: Canada; stateProvince: Yukon Territory; verbatimLocality: Dempster Hwy nr North Fork Pass; verbatimElevation: 1200 m; decimalLatitude: 64.57942; decimalLongitude: -138.28212; geodeticDatum: WGS84; samplingProtocol: sweeping; year: 2011; month: 6; day: 24; habitat: wet tundra; fieldNumber: wet replicate 3; sex: male; recordedBy: NBP Field Party; datasetID: LEM0110624; institutionCode: LEMQ

Paratype:

a. class: **Insecta**; order: **Diptera**; family: **Heleomyzidae**; genus: **Neossos**; specificEpithet: tombstonensis; scientificNameAuthorship: Solecki & Wheeler, 2015; continent: North America; country: Canada; stateProvince: Yukon Territory; verbatimLocality: S Klondike Hwy, 18.2 km S Alaska

Share this article



[Contents](#)
[Article info](#)
[Citation](#)
[Metrics](#)
[Comment](#)

[Figs](#)
[Map](#)
[Taxa](#)
[Data](#)
[Refs](#)

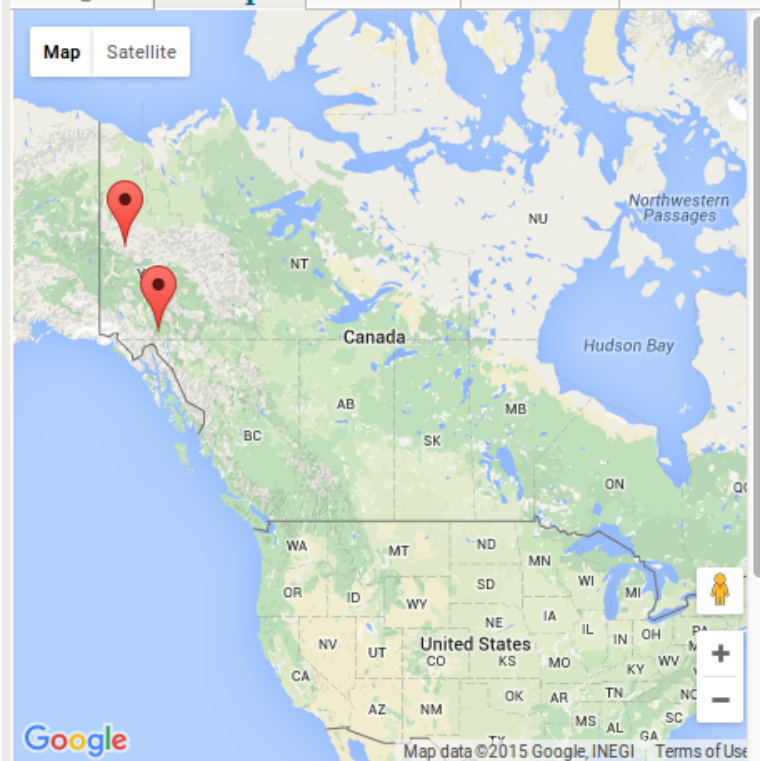


Figure 3. The different sub-strains show a large spectrum of values for the stripe deviation parameter.



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 boxplots: 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 the *F1000Research* website is interactive; readers can define the type of whiskers displayed as either Tukey whiskers (1.5 x IQR from 1st/3rd quartile; **A**) or the 10th–90th percentiles (**B**). The text color code used for the genotypes is analogous to that used in Figure 2. The red horizontal line corresponds to the median value for random walks: 44°. Sample size is 11–12 for each boxplot. No statistical analysis was performed.

Re-Plot Figure

Define Whisker

Tukey ▼

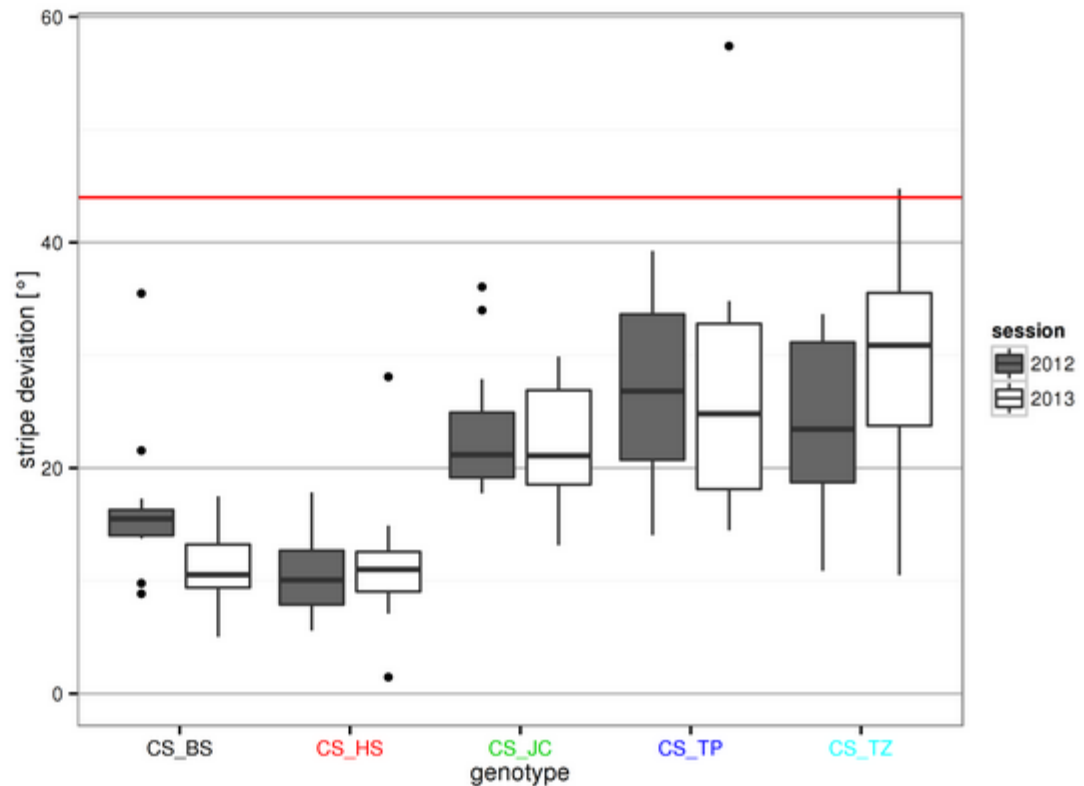
Whiskers extend to values
1.5 x IQR from 1st/3rd
quartile

Apply

Reset

HELP

Select option in the
dropdown box and click
Apply to re-plot figure.
Click Reset to return to
the default figure display.



Select date to see other versions

Select a date...
Select a date...
11 May 2015 CR_BE
22 Apr 2015 GR_GR
21 Apr 2015 BB_JB
30 Jul 2014 CS_TZ
30 Jul 2014 CS_TP
30 Jul 2014 CS_JC
30 Jul 2014 CS_HS
30 Jul 2014 CS_BS

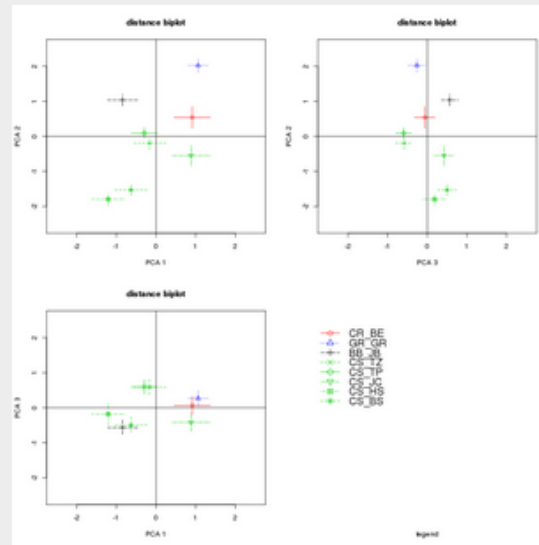


Figure 4. Updating principal component analysis of Canton S strains.

Results from the PCA obtained using the same analysis as for [Figure 2](#), but with data uploaded from different laboratories. The version of this figure on the *F1000Research* site is 'living'; it will automatically re-plot as and when new data for other Canton S strains are submitted, and users can visualize previous versions of this figure. The conclusions of this article only relate to the data available at the time of publication. The prefixes in the key are the initials of the data contributor (except CS_ strains, which were tested by Julien Colomb); full names and affiliations can be found in the figure legend of the article on the *F1000Research* site. The suffixes denote the initials of the principal investigators from where each sub-strain was sourced. The BB_JB (Jose Botella) strain was ordered from the Bloomington stock center (stock #1) approx. seven years ago. BB_JB falls within the range of variability seen so

...

CR_BE: Added on 11 May 2015 by Christa Rhiner, Eduardo Moreno and Andrés Gutiérrez García at Institute of Cell Biology, Universität Bern, Switzerland. DOI: 10.5256/f1000research.4263.d47579 | [Download data](#) | [Cite data](#)

GR_GR: Added on 22 Apr 2015 by Gregg Roman, Stefani Garcia and Miguel de la Flor at Department of Biology and Biochemistry, University of Houston, TX, USA. DOI:

[Download All Data For This Figure](#)

[Close](#)

Colomb J and Brembs B. (2015) *F1000Research* doi: [10.12688/f1000research.4263.2](https://doi.org/10.12688/f1000research.4263.2)

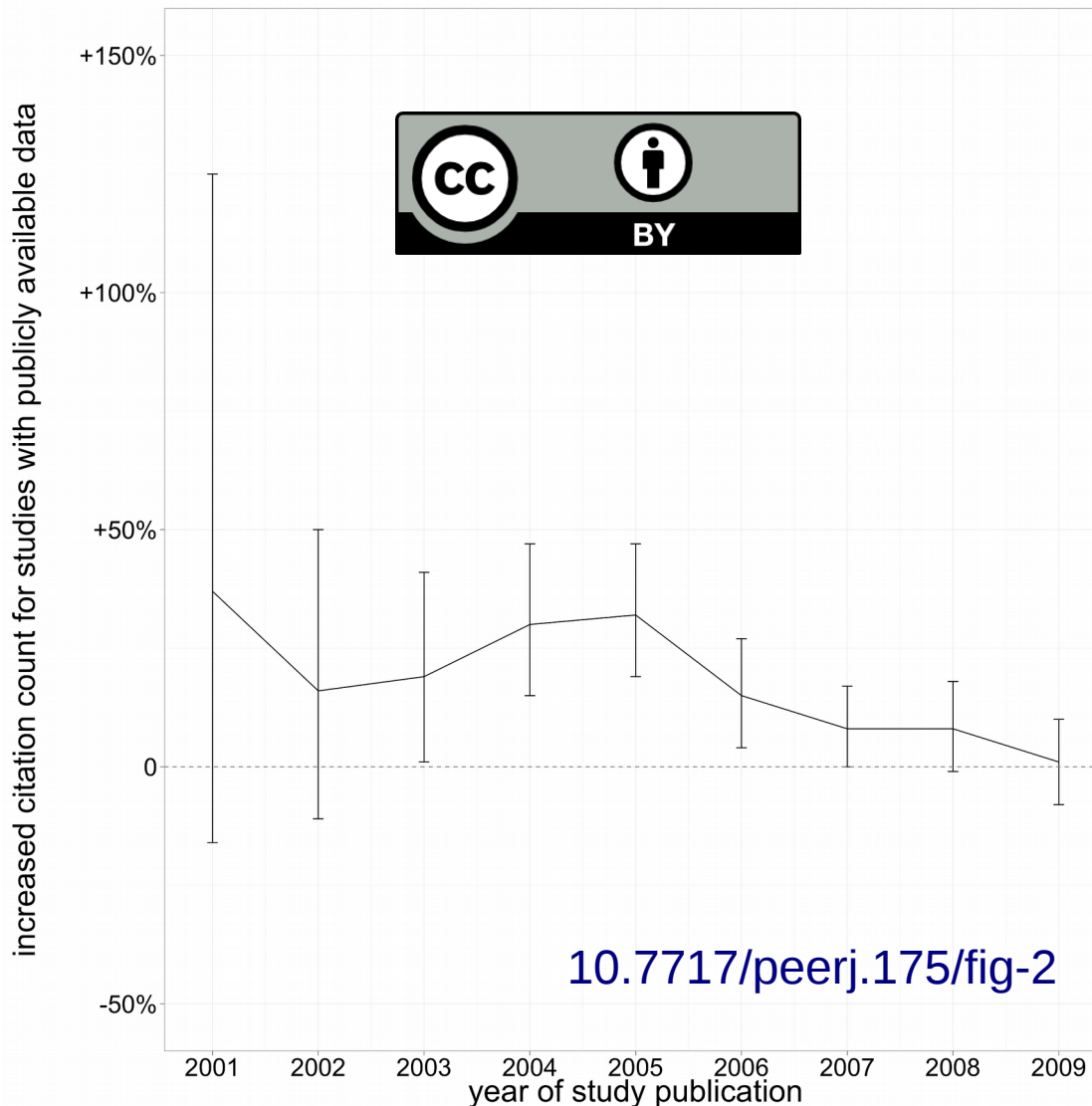
Access to data in social sciences

- UK Data Service:
 - [Overview](#)
 - [Teaching resource](#)
 - Interview data – [Psychological interviews](#)
 - Non-interview data – [Diaries](#), [Visual materials](#)
 - Example of [metadata record](#)
 - Other [open data providers](#)
- [European Union Open Data Portal](#)

Access to data (sources) in arts & humanities

- Many sources online, for example:
 - Visual data - Wikimedia Commons ([Yorck project](#))
 - Dance data - [The Archive of Siobhan Davies Dance](#)
- [Registry of research data repositories](#)
- Article in [The Guardian](#)
- Jisc Digital Media tools - <http://www.jiscdigitalmedia.ac.uk/>

Data sharing benefits authors & re-users



Piwowar HA, Vision TJ. (2013) Data reuse and the open data citation advantage. **PeerJ** 1:e175

“...open data citation benefit for this sample to be 9%”

relative to papers providing no public data, for gene expression microarray data

See also previous work by Piwowar:
[10.1371/journal.pone.0000308](https://doi.org/10.1371/journal.pone.0000308)

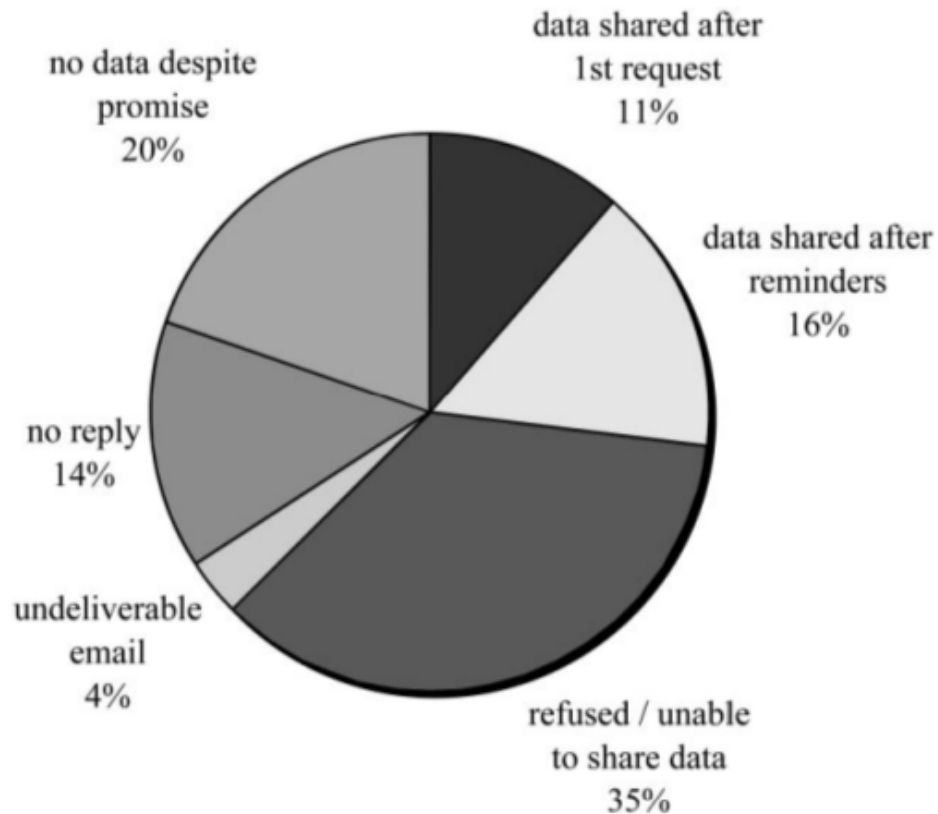
Those who share data, do better science

Wicherts, J. M., Bakker, M. & Molenaar, D. (2011)
Willingness to share research data is related to the strength of the evidence and the quality of reporting of statistical results. PLoS ONE 6, e26828+ URL
<http://dx.doi.org/10.1371/journal.pone.0026828>

The authors examined psychological papers for the quality of statistical reporting & asked the authors of those papers for the full data underlying the reported results. Generally, those who shared, had more statistically robust, reproducible results.

“Email the author for data” - doesn't work

Figure 1.
Percentages of Empirical Articles' Corresponding Authors in Different Response Categories



A well-known problem, which I myself have also faced many times!!!

Many legacy journals unfortunately still pretend that “email the author” is still acceptable.

Wicherts JM, Borsboom D, Kats J, Molenaar D (2006) **The poor availability of psychological research data for reanalysis.** *American Psychologist* 61: 726–728 [link](#)

Best practice open data is time consuming
(but still worth the extra effort!)

Emilio M. Bruna recently provided an estimate of the amount of time it took him to prepare & upload open data related to publication to figshare & dryad.

11 & \$90
Hours (for Dryad)

Providing open-source code was the most time consuming part (25.5 hours), and Open Access publication the most expensive (\$600).

<http://brunalab.org/blog/2014/09/04/the-opportunity-cost-of-my-openscience-was-35-hours-690/>

Not all data should be open!

Obviously, there are some types of data which should NOT be made mandatorily *open* e.g. sensitive medical data

However, with informed consent, if patients really want to, they should be allowed to publish their own medical data

Ethical / legal considerations

- Duty to protect participants, guarantee their confidentiality and inform about data use
- Informed consent forms on data collection AND how data will be stored and shared
- Consent forms should not restrict data sharing and reuse unnecessarily
- Don't collect sensitive data unnecessarily

Data anonymisation for sharing

- Direct and indirect identifiers
- Quantitative data
 - Remove identifiers; Aggregate or reduce the precision of variables; Generalise the meaning of a detailed text variable ; Anonymise relation, geo-referenced data
- Qualitative data
 - Time consuming, even harder for audio-visual data
 - Use pseudonyms, replacement terms, vaguer descriptors or systems of coding
- [UK Data Service guide](#) and [ICO guide on Anonymisation](#)

But with informed consent, sharing sensitive health data can have good outcomes



<http://blog.ted.com/2013/06/14/why-i-opensourced-cures-for-my-cancer-salvatore-iaconesi-at-tedglobal-2013/>

Other exceptions to the open default

Sensitive species conservation data
e.g. exact geocoordinates of home range

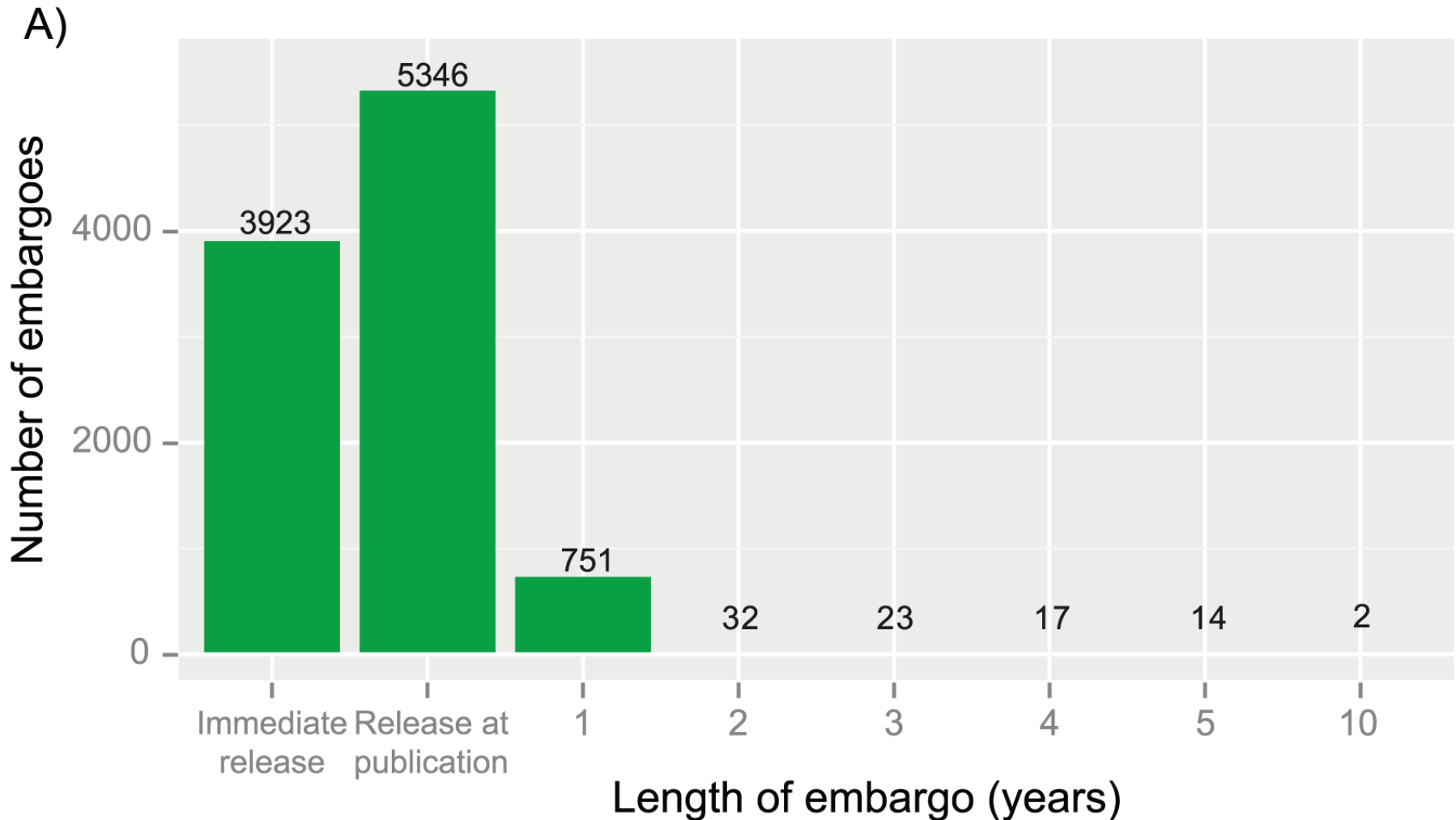


Certain species of wild orchids, cacti & carnivorous plants are highly endangered by illegal harvesting.

Publishing the exact geolocation data of the remaining populations of commercially-desirable, endangered species is really dumb thing to do.

Such data is typically held privately in databases (not publicly available).

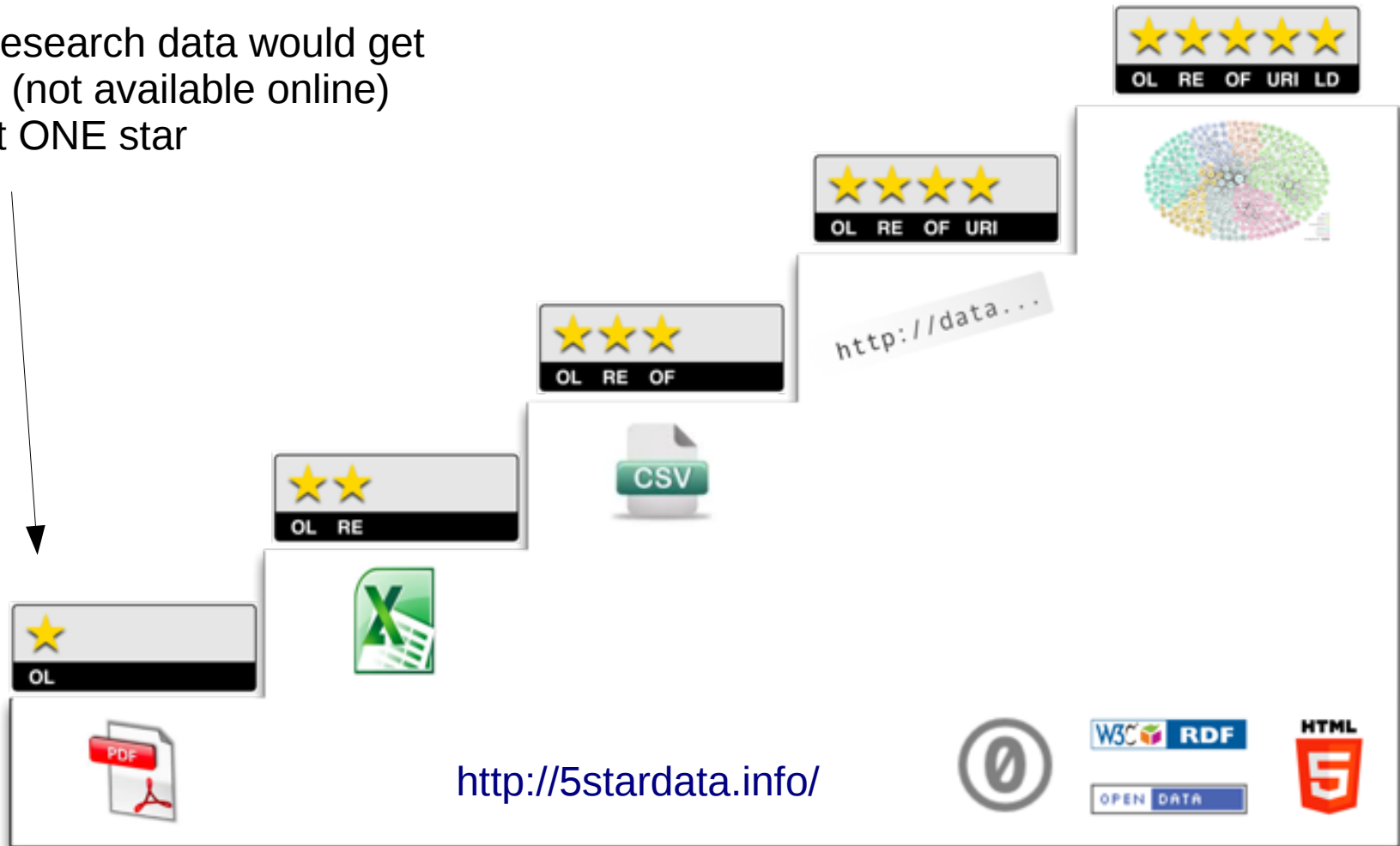
When to open data?



Source: [10.1371/journal.pbio.1001779](https://doi.org/10.1371/journal.pbio.1001779)

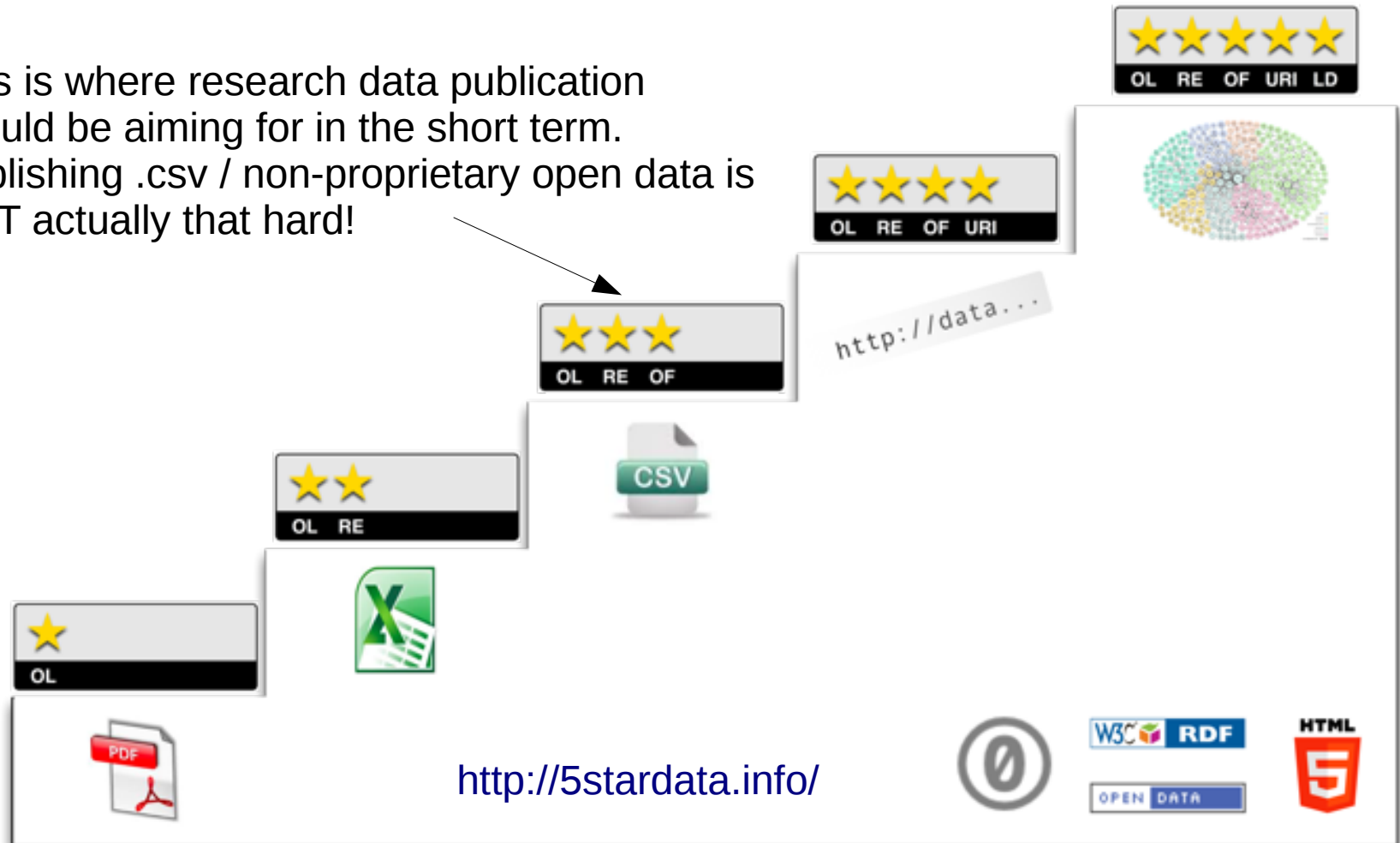
The 5 stars of open data

Most research data would get
ZERO (not available online)
Or just ONE star



3-star open research data is achievable

This is where research data publication should be aiming for in the short term. Publishing .csv / non-proprietary open data is NOT actually that hard!



Open Access & Open Data have similar goals

80%
of research is
publicly
funded¹

e.g. maximising the return on investment, provided to research by taxpayers and charities

Further Reading



Ubiquity Press
(Open Access, CC BY)
DOI: [10.5334/ban](https://doi.org/10.5334/ban)

1. Editor's Introduction - Samuel A. Moore
2. Open **Content Mining** - Peter Murray-Rust, Jennifer C. Molloy, Diane Cabell
3. The **Need to Humanize Open Science** - Eric C. Kansa
4. Data Sharing in a Humanitarian Organization: The Experience of **Médecins Sans Frontières** - Unni Karunakara
5. Why Open **Drug Discovery** Needs Four Simple Rules for Licensing Data and Models - Antony J. Williams, John Wilbanks, Sean Ekins
6. Open Data in the **Earth and Climate Sciences** - Sarah Callaghan
7. Open Minded **Psychology** - Wouter van den Bos, Mirjam Jenny, Dirk Wulff
8. Open Data in **Health Sciences** - Tom Pollard
9. Open Research Data in **Economics** - Velichka Dimitrova
10. Open Data and **Palaeontology** - Ross Mounce

Further Reading

- The Open Data Handbook - <http://opendatahandbook.org/>
- 5 star Open Data - <http://5stardata.info/>
- Science as an open enterprise (2012) [A Royal Society report](#)
- Caetano, D. S. & Aisenberg, A. 2014 Forgotten treasures: the fate of data in animal behaviour studies [Animal Behaviour](#)

Data sharing in phylogenetics

- Magee et al 2014 The Dawn of Open Access to Phylogenetic Data [PLOS ONE](#)
- Drew et al 2013 Lost Branches on the Tree of Life. [PLOS Biology](#)
- Stoltzfus et al 2012 Sharing and re-use of phylogenetic trees (and associated data) to facilitate synthesis. [BMC Research Notes](#)

On licencing & legal issues with re-use

- Hagedorn et al 2011 Creative commons licenses and the non-commercial condition: Implications for the re-use of biodiversity information. [ZooKeys](#)
- Mounce 2012. Life as a palaeontologist: Academia, the internet and creative commons. [Palaeontology Online](#)
- Klimpel, P. 2012 Consequences, Risks, and side-effects of the license module Non-Commercial – NC [[PDF](#)]

Further Reading

- Murray-Rust, P. Open data in science. *Serials Review* 34, 52-64 (2008). URL <http://dx.doi.org/10.1016/j.serrev.2008.01.001>
- Leonelli, S., Smirnoff, N., Moore, J., Cook, C. & Bastow, R. Making open data work for plant scientists. *Journal of Experimental Botany* 64, 4109-4117 (2013). URL <http://dx.doi.org/10.1093/jxb/ert273>
- Hrynaszkiewicz, I. & Cockerill, M. Open by default: a proposed copyright license and waiver agreement for open access research and data in peer-reviewed journals. *BMC Research Notes* 5, 494+ (2012). URL <http://dx.doi.org/10.1186/1756-0500-5-494>
- Boulton, G., Rawlins, M., Vallance, P. & Walport, M. Science as a public enterprise: the case for open data. *The Lancet* 377, 1633-1635 (2011). URL [http://dx.doi.org/10.1016/s0140-6736\(11\)60647-8](http://dx.doi.org/10.1016/s0140-6736(11)60647-8)
- Parr, C. S. Open sourcing ecological data. *BioScience* 57, 309-310 (2007). URL <http://dx.doi.org/10.1641/b570402>
- Poisot, T., Mounce, R. & Gravel, D. Moving toward a sustainable ecological science: don't let data go to waste! *Ideas in Ecology and Evolution* 6 (2013). URL <http://dx.doi.org/10.4033/iee.2013.6b.14.f>

Further reading on data in arts, humanities & social sciences

- UK Data Archive on data management <http://www.data-archive.ac.uk/create-manage>
- OA journals
 - Digital Humanities Quarterly <http://www.digitalhumanities.org/dhq/>
 - Journal of Digital Humanities <http://journalofdigitalhumanities.org/>
- Digital Humanities Data Curation <http://guide.dhcuration.org/>
- Kaptur project (research data in visual arts) <http://www.vads.ac.uk/kaptur/>
- Working with digital media – Jisc Digital Media <http://www.jiscdigitalmedia.ac.uk/>



Research Ideas and Outcomes riojournal.com

Thank you!

Happy to answer all questions:

ross.mounce@nhm.ac.uk
@rmounce

<http://rossmounce.co.uk/>