



Research Data Management

Wouter Haak
VP Research Data Management

w.haak@Elsevier.com



May 2016



A view on Research Data:

Series

Increasing value and reducing waste: addressing inaccessible research

Dr An-Wen Chan, DPhil^a,  , Prof Fujian Song, PhD^b, Andrew Vickers, PhD^c, Tom Jefferson, MD^d, Prof Kay Dickersin, PhD^e, Prof Peter C Gøtzsche, DrMedSci^f, Prof Harlan M Krumholz, MD^{g, h, i}, Davina Gherzi, PhD^j, H Bart van der Worp, PhD^k

Summary

The methods and results of health research are documented in study protocols, full study reports (detailing all analyses), journal reports, and participant-level datasets. **However, protocols, full study reports, and participant-level datasets are rarely available, and journal reports are available for only half of all studies and are plagued by selective reporting of methods and results.** Furthermore, information provided in study protocols and reports varies in quality and is often incomplete. **When full information about studies is inaccessible, billions of dollars in investment are wasted, bias is introduced, and research and care of patients are detrimentally affected.** To help to improve this situation at a systemic level, three main actions are warranted. First, academic institutions and funders should reward investigators who fully disseminate their research protocols, reports, and participant-level datasets. Second, standards for the content of protocols and full study reports and for data sharing practices should be rigorously developed and adopted for all types of health research. Finally, journals, funders, sponsors, research ethics committees, regulators, and legislators should endorse and enforce policies supporting study registration and wide availability of journal reports, full study reports, and participant-level datasets.



A view on Research Data:

Series

Increasing value and reducing waste: addressing inaccessible research

Dr An-Wen Chan, DPhil^{a, b, c}, Prof Fujian Song, PhD^d, Andrew Vickers, PhD^e, Tom Jefferson, MD^f, Prof Kay Dickersin, PhD^g, Prof Peter C Getzsche, DrMedSci^h, Prof Harlan M Krumholz, MD^{h, i, j}, Davina Ghersi, PhD^k, H Bart van der Worp, PhD^k

Summary

The methods and results of health research are documented in study protocols, full study reports (detailing all analyses), journal reports, and participant-level datasets. However, protocols, full study reports, and participant-level datasets are rarely available, and journal reports are available for only half of all studies and are plagued by selective reporting of methods and results. Furthermore, information provided in study protocols and reports varies in quality and is often incomplete. When full information about studies is inaccessible, billions of dollars in investment are wasted, bias is introduced, and research and care of patients are detrimentally affected. To help to improve this situation at a

systemic level, three main actions are warranted. First, academic institutions and funders should reward investigators who fully document their research in protocols, reports, and participant-level datasets. Second, standards for the content of protocols, full study reports and for data sharing practices should be rigorously developed and adopted for all types of research. Finally, funders, regulators, and legislators should endorse and enforce policies supporting study registration and data sharing.

When full information about studies is inaccessible, billions of dollars in investment are wasted, bias is introduced, and research and care of patients are detrimentally affected



A view on Research Data:

Series

Increasing value and reducing waste: addressing inaccessible research

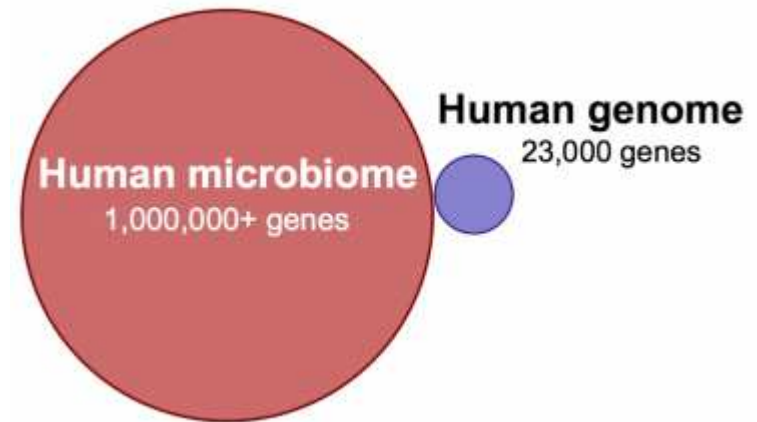
Dr An-Wen Chan, DPhil^a, Prof Fujian Song, PhD^b, Andrew Vickers, PhD^c, Tom Jefferson, MD^d, Prof Kay Dickersin, PhD^e, Prof Peter C Gøtzsche, DrMedSci^f, Prof Harlan M Krumholz, MD^g, Davina Ghersi, PhD^h, H Bart van der Worp, PhDⁱ

Summary

The methods and results of health research are documented in study protocols, full study reports (detailing all analyses), journal reports, and participant-level datasets. **However, protocols, full study reports, and participant-level datasets are rarely available, and journal reports are available for only half of all studies and are plagued by selective reporting of methods and results.** Furthermore, information provided in study protocols and reports varies in quality and is often incomplete. When full information about studies is **inaccessible, research, and care of patients are detrimentally affected.** To help to improve this situation, a systems review of the current status of the availability of study protocols and full study reports, and participant-level datasets. **Second, standards for the content of protocols and full study reports should be developed and adopted for all types of health research.** Finally, journals, funders, sponsors, research ethics committees, regulatory agencies, and funders should enforce policies supporting study registration and wide availability of journal reports, full study reports, and participant-level datasets.

What are we really after: Bio & LS

With enough observations, trends and anomalies can be detected:



- “Here we present resources from a population of 242 healthy adults sampled at 15 or 18 body sites up to three times, which have generated 5,177 microbial taxonomic profiles from 16S ribosomal RNA genes and over 3.5 terabases of metagenomic sequence so far.”

The Human Microbiome Project Consortium, Structure, function and diversity of the healthy human microbiome, *Nature* 486, 207–214 (14 June 2012) doi:10.1038/nature11234

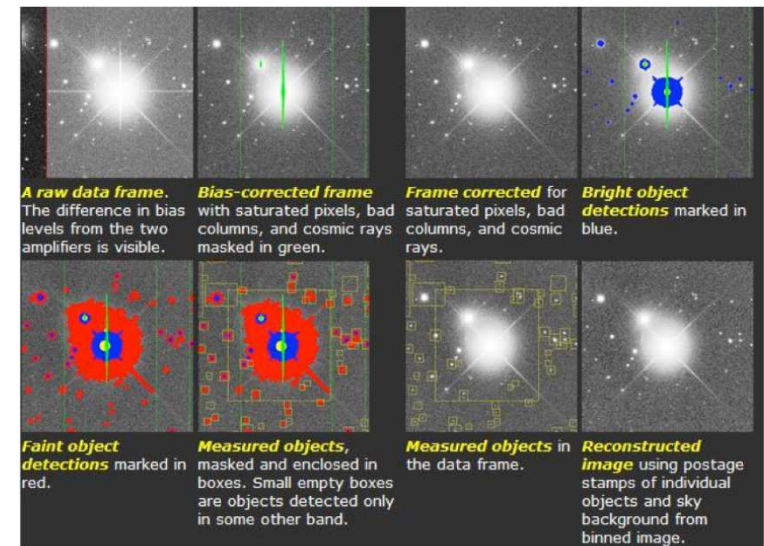
- “The large sample size — 4,298 North Americans of European descent and 2,217 African Americans — has enabled the researchers to mine down into the human genome.”

Nidhi Subbaraman, *Nature News*, 28 November 2012, High-resolution sequencing study emphasizes importance of rare variants in disease.

What are we really after: astronomy

Extracts from “the top 10 benefits of data sharing in astronomy”, from Sloan Digital Sky Survey:

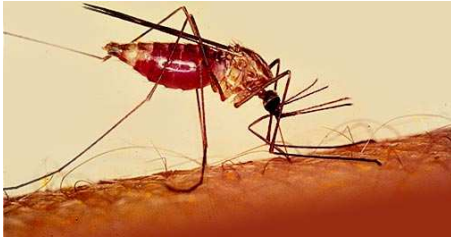
- **Early data releases greatly improve the final product**, e.g. more people “looking” at the data increases the chance of finding subtle problems, especially important for space missions with finite lifetime, e.g. the ESA’s Gaia mission
- **More science is extracted from the same dataset**, e.g. diversity of ideas: many of the most visible SDSS results were unanticipated in the original project proposal
- **Sometimes the only way to secure scarce resources**, “easy things” (e.g. those that can be put together by a small number of groups/institutions) have been done in the last century; the “road ahead” requires more substantial merging of research resources, like HST Deep Field, UKIDSS, LSST
- **Results in more citations and prestige to the team who produced data**; practically all postdocs from the first phase of SDSS hold faculty-level positions today



http://www.astro.washington.edu/users/ivezic/Outreach/Talks/NAS2011_Ivezic.pdf

Željko Ivezić, Department of Astronomy, University of Washington - The Sloan Digital Sky Survey Telescope - Apache Point Observatory, NM

With contributions from: Andy Connolly, Bob Hanisch, David Hogg, Mario Jurić, Andy Lawrence, Robert Lupton, Mathias Steinmetz, Michael Strauss, Alex Szalay, Tony Tyson, Roy Williams



What are we really after: malaria



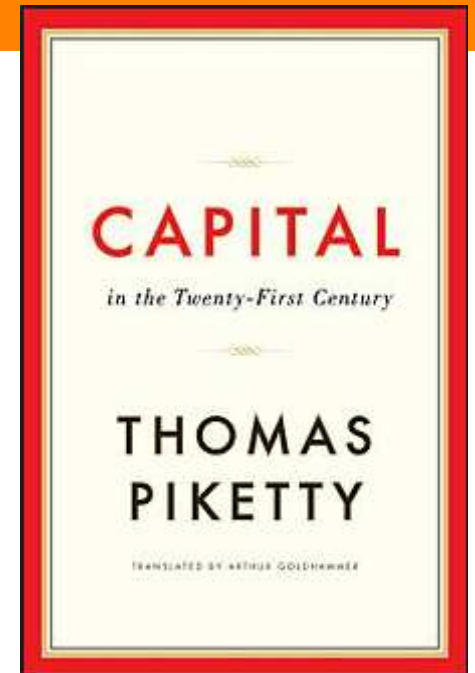
WWARN, the first malaria data sharing network, has used pooled analysis of shared data to provide evidence to **help improve dosing regimens** of malaria treatments

- research partners from over [260 institutions globally](#) have worked with WWARN, and over 120,000 individual patient records have been contributed to the WWARN Data Centre. That equates to around 80% of all the available artemisinin combination therapy trial data.
- Based on the results from the [Dihydroartemisinin-Piperaquine \(DP\) Dose Impact Study Group](#) and pharmacometric modelling of piperaquine, the World Health Organization has revised the recommended dose of DP, a commonly used antimalarial for young children. These [revised dose regimens \(link is external\)](#) are predicted to provide similar piperaquine concentrations across all age groups.
- Similarly, a meta-analysis undertaken by the Artesunate Amodiaquine (ASAQ) Dose Impact Study Group, based on 9,106 patients, found that although the overall efficacy of ASAQ is adequate in most settings, efficacy varies with the formulation and is affected by a range of risk factors including age. The [Artemether lumefantrine \(AL\) Dose Impact Study Group](#) found that cure rates were lowest in young children from Asia, especially those with high parasitemia, and young underweight children from Africa.

What are we really after: social sciences

Capital in the Twenty-First Century is a 2013 book by French economist Thomas Piketty.

- It focuses on wealth and income inequality in Europe and the United States since the 18th century
- Central thesis is that when the rate of return on capital (r) is greater than the rate of economic growth (g) over the long term, the result is concentration of wealth, and this unequal distribution of wealth causes social and economic instability
- All raw data, normalized data, the analysis, and methods have all been made publicly available on a dedicated website
<https://www.quandl.com/data/PIKETTY>



“Here are enormous quantities of information distilled from tax rolls, inheritance records, and various other public data sources, laid out in charts that should be readily accessible to the layest of lay readers. Not all of the information in these sections is novel or startling. Having it together in one place, however, is valuable, and even most of the book’s fiercest critics respect this achievement.” [1]

It also shows data sharing can lead to issues [2]:

- Chris Giles, economics editor of the Financial Times (FT), identified what he claims are "unexplained errors" in Piketty's data, in particular regarding wealth inequality increases since the 1970s. “contain a series of errors that skew his findings”
- Subsequently, Piketty wrote a response defending his findings; the accusation and responses received wide press coverage
- E.g. Scott Winship, a sociologist at the MIPR, claims the allegations are not "significant for the fundamental question of whether Piketty's thesis is right or not"

[Piketty's Capital: An Economist's Inequality Ideas Are All the Rage](https://en.wikipedia.org/wiki/Capital_in_the_Twenty-First_Century) by [Megan McArdle](#), *Bloomberg Businessweek*, May 29, 2014
https://en.wikipedia.org/wiki/Capital_in_the_Twenty-First_Century

Sharing data – more citations

OPEN ACCESS Freely available online



Sharing Detailed Research Data Is Associated with Increased Citation Rate



Heather A. Piwowar*, Roger S. Day, Douglas B. Fridsma

Department of Biomedical Informatics, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania, United States of America

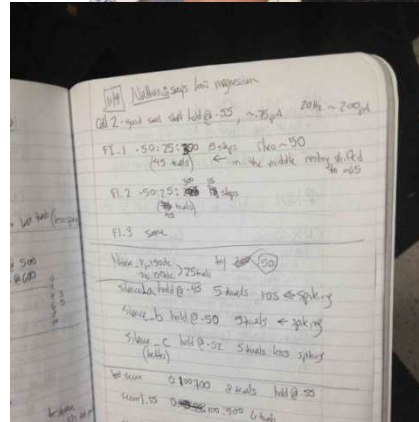
Background. Sharing research data provides benefit to the general scientific community, but the benefit is less obvious for the investigator who makes his or her data available. **Principal Findings.** We examined the citation history of 85 cancer microarray clinical trial publications with respect to the availability of their data. The 48% of trials with publicly available microarray data received 85% of the aggregate citations. Publicly available data was significantly ($p = 0.006$) associated with a 69% increase in citations, independently of journal impact factor, date of publication, and author country of origin using linear regression. **Significance.** This correlation between publicly available data and increased literature impact may further motivate investigators to share their detailed research data.

Data in the lab

“Using antibodies and squishy bits, grad students experiment and enter details into their lab notebook”

The PI then tries to make sense of their slides, and writes a paper.

End of story.




Journal of Neuroscience Methods
 Journal homepage: www.elsevier.com/locate/jneumeth

A simple method of *in vitro* electroporation allows visualization, recording, and calcium imaging of local neuronal circuits[☆]

Kenneth R. Hovis^{a,b}, Krishnan Padmanabhan^{a,b}, Nathaniel N. Urban^{a,b,c,*}

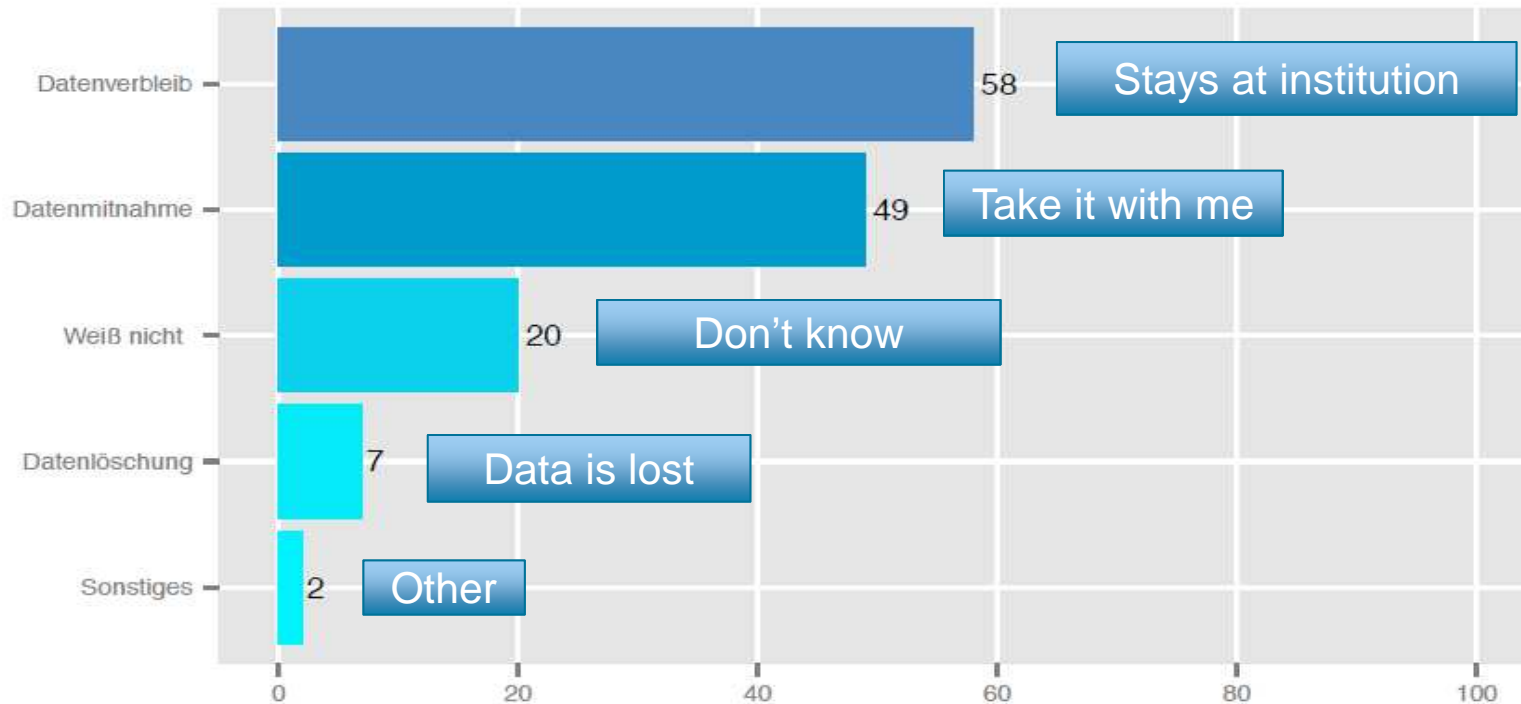
^a Department of Biological Sciences, Carnegie Mellon University, Pittsburgh, PA 15213, United States
^b Center for the Neural Basis of Cognition, Carnegie Mellon University, Pittsburgh, PA 15213, United States
^c Center for Neuroscience, University of Pittsburgh, Pittsburgh, PA 15260, United States

ARTICLE INFO
 Article history:
 Received 3 March 2010
 Received in revised form 17 May 2010
 Accepted 24 May 2010
 Keywords:

ABSTRACT
 Since Cajal's early drawings, the characterization of neuronal architecture has been paramount in understanding neuronal function. With the development of electrophysiological techniques that provide unprecedented access to the physiology of these cells, experimental questions of neuronal function also become more tractable. Fluorescent tracers that can label the anatomy of individual or populations of neurons have opened the door to linking anatomy with physiology. Experimentally however, current techniques for bulk labeling of cells *in vitro* often affect neuronal function creating a barrier for ex-

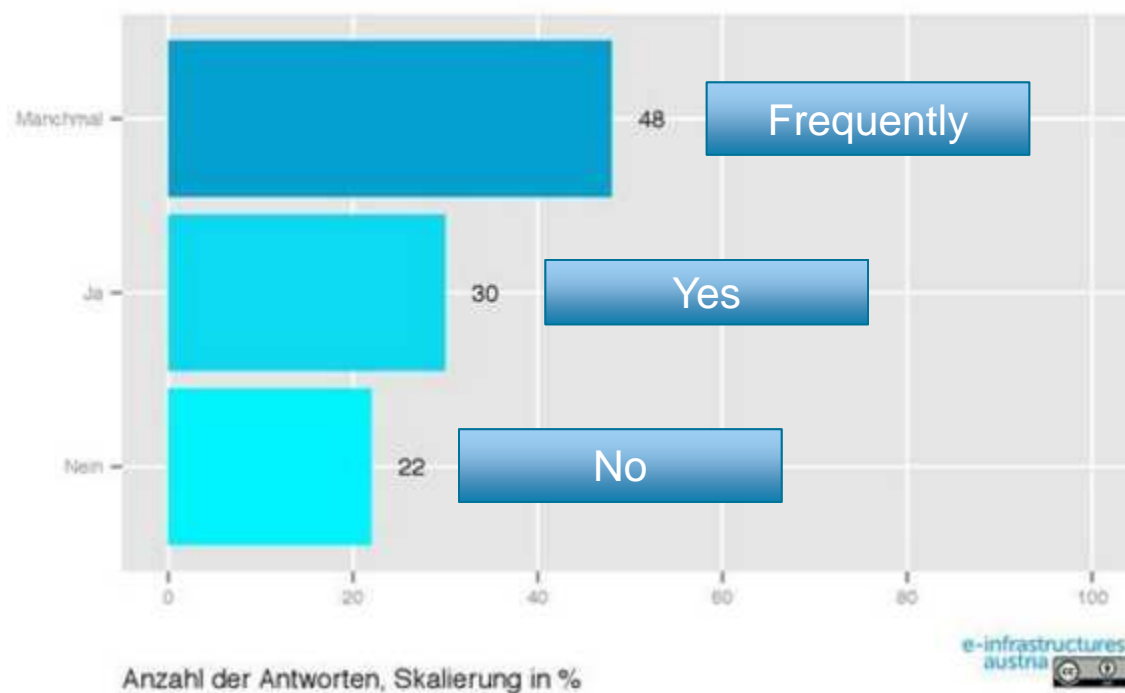


When You Leave Your Institution, What Happens To Your Data?



Anzahl der Antworten, Skalierung in %

Is Your Research Data Useful To Others?



„Forschende und ihre Daten. Ergebnisse einer österreichweiten Befragung (eBook)“

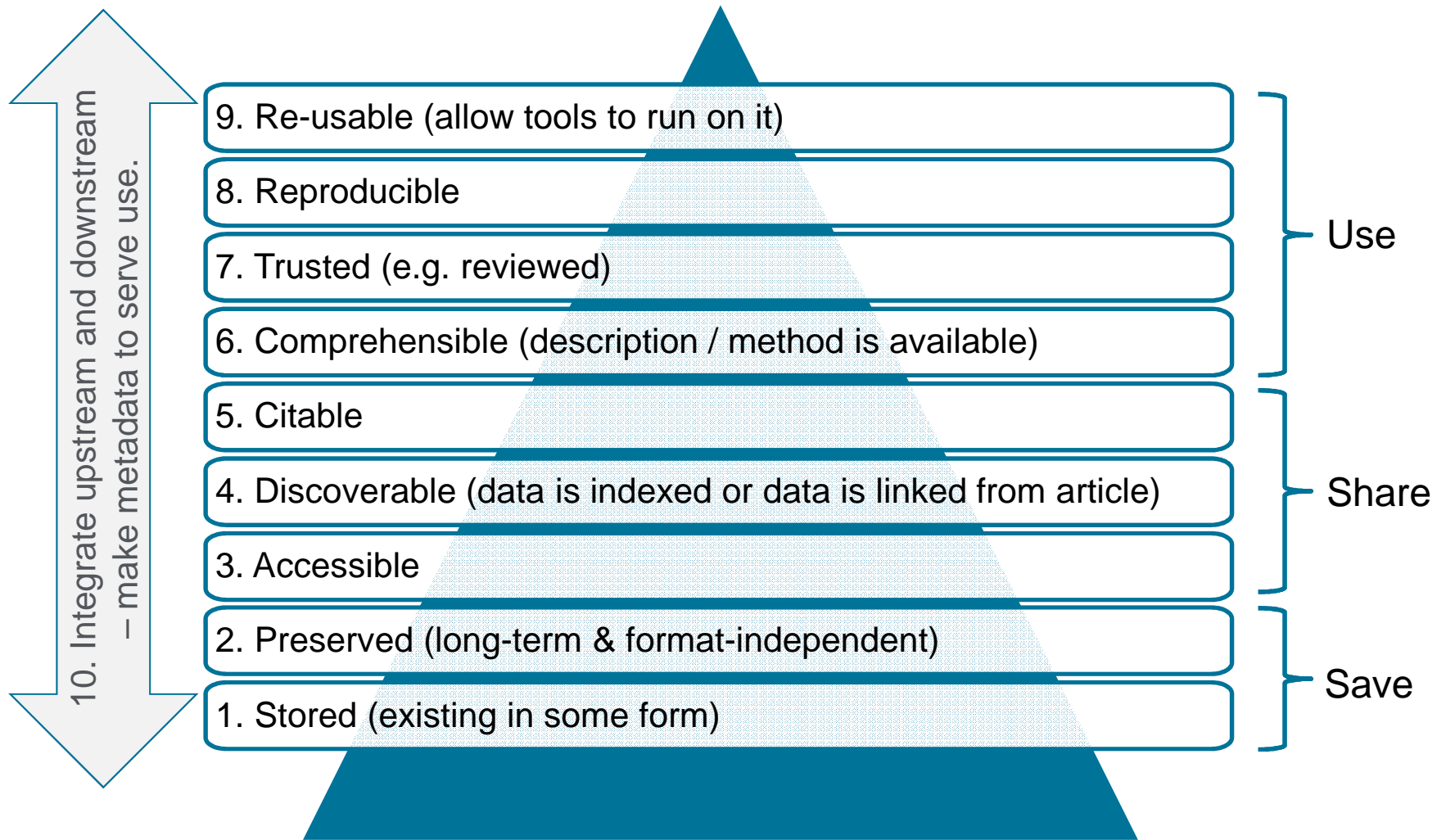
E-infrastructures Austria

Bauer, B. (Bruno) et al

Oct 2015

https://phaidra.univie.ac.at/detail_object/o:407736

Elsevier's approach; climbing the data pyramid



1. Research Data linking – linking articles to external datasets

ScienceDirect Journals Books Wouler Haak Help

Download PDF Export Search ScienceDirect Advanced search

Article outline Show full outline

Highlights Abstract Graphical abstract Keywords

1. Introduction 2. Results and discussion 3. Conclusions 4. Experimental Acknowledgments Appendix A. Supplementary data References

Figures and tables

Table 1

European Journal of Medicinal Chemistry
Volume 96, 26 May 2015, Pages 281–295

Original article
Synthesis, crystal structure and effect of indeno[1,2-b]indole derivatives on prostate cancer *in vitro*. Potential effect against MMP-9

Gricela Lobo^a, Melina Monasterios^a, Juan Rodríguez^a, Neira Gamboa^a, Mario V. Capparelli^b, Javier Martínez-Cuevas^a, Michael Lein^{a, c}, Klaus Jung^{a, e}, Claudia Abramjuk^{a, f}, Jaime Charris^a

doi:10.1016/j.ejmech.2015.04.023 Get rights and content

Data for this Article

CCDC Cambridge Crystallographic Data Centre
Crystallographic data

Recommended articles

Design, synthesis and biological evaluation of quin...
2015, European Journal of Medicinal Chemistry more

Discovery of the 2-phenyl-4,5,6,7-Tetrahydro-1H-in...
2015, European Journal of Medicinal Chemistry more

Antibacterial active compounds from *Hypericum as...*
2015, European Journal of Medicinal Chemistry more

Citing articles (0)

Related book content

Highlights

- Compounds were easily synthesized and with highly regioselectivity.
- Crystals consist of equimolar mixtures of the RR and SS diastereomers.
- All tested compounds proved to be moderately active, except one.

Cambridge Crystallographic Data Centre

CSD entry: KUVVOC Search

Your query was: Doi: 10.1016/j.ejmech.2015.04.023 and returned 2 records

Results

CCDC #	Refcode
<input checked="" type="checkbox"/>	1022819 KUVVOC
<input checked="" type="checkbox"/>	1022641 KUVVIW

Download

KUVVOC : 7,7-Dimethyl-5-(2,4-dimethylphenyl)-(4*BRS*,9*BRS*)-dihydroxy-4*b*,5,6,7,8,9*b*-hexahydro-indeno[1,2-*b*]indole-9,10-dione
Space Group: P2₁/n, Cell: a 9.6597(4)Å b 18.0974(6)Å c 12.2889(5)Å α 90.00° β 93.7970(10)° γ 90.00°

3D viewer

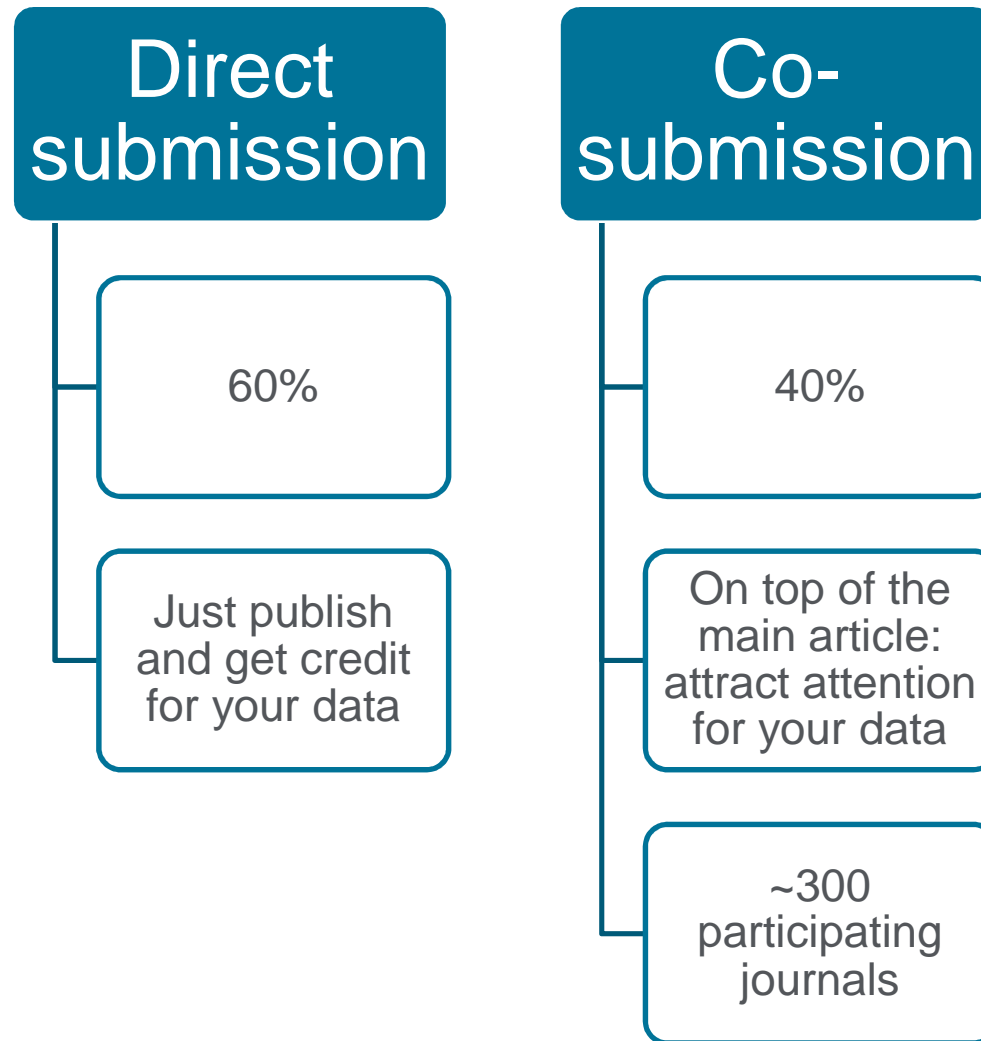
Chemical diagram

Style: Ball and Stick Labels: No Labels Packing: None Measure: None

View group symbols key

<http://www.sciencedirect.com/science/article/pii/S022352341500272X>

2. Gold OA reviewed **data journals**, software journals, materials & methods journals (collectively called “Research Elements”)



e.g.:



SoftwareX



Data in Brief



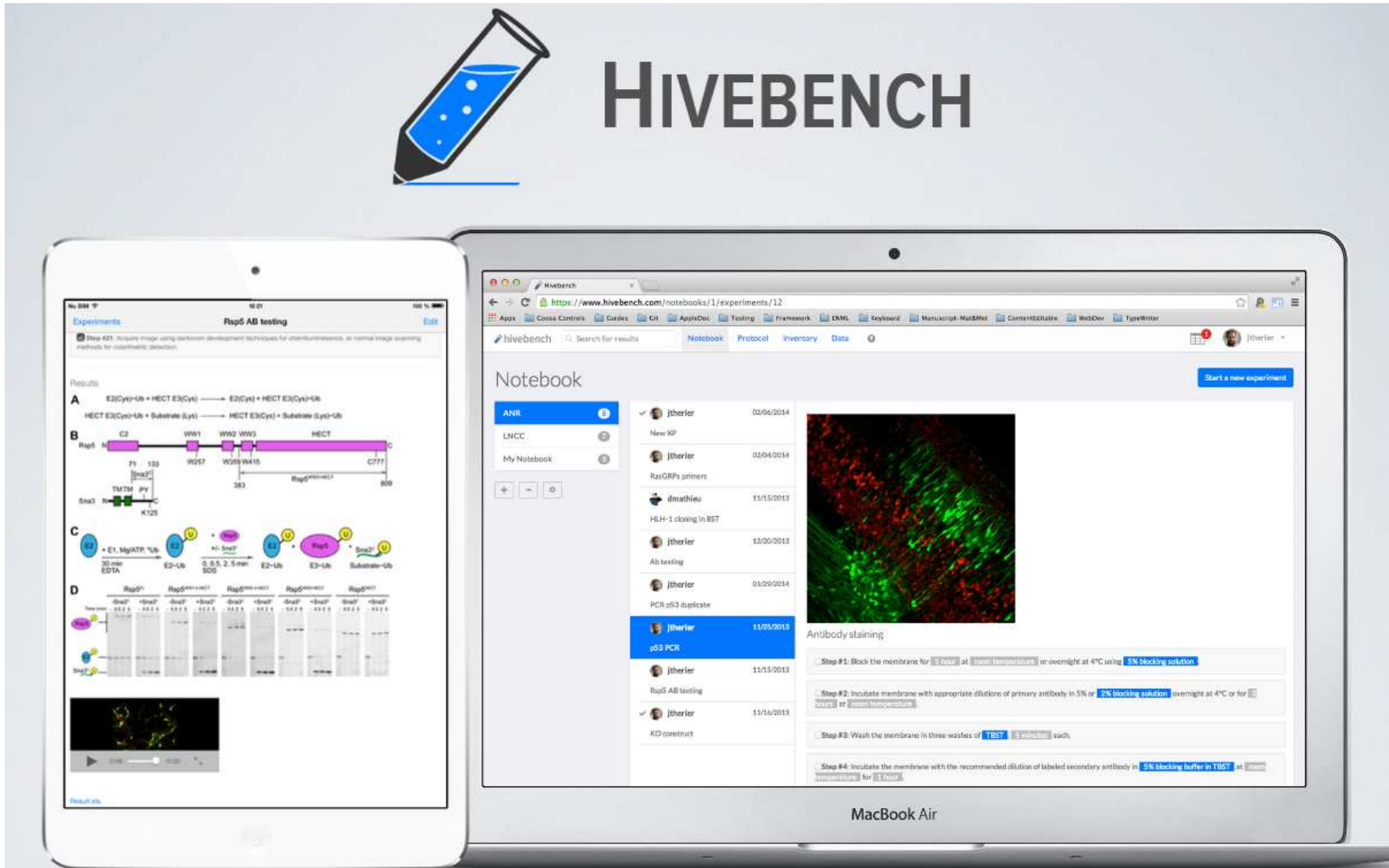
MethodsX

3. Development Partnership (France) – Lab Data Tool: structure in the lab

www.hivebench.com



HIVEBENCH



4. Manage, Store: Mendeley Data launched Dec 2015

The screenshot shows the Mendeley Data interface for a dataset. At the top, there is a navigation bar with 'MENDELEY DATA Beta', 'Browse', 'My datasets', 'New dataset', 'Log in', and 'Create account'. The main content area is divided into several sections:

- Title:** Reproducible experiments on dynamic resource allocation in cloud data centers
- DOI:** 10.17632/xz6gv65m6d.6
- Contributor(s):** Andreas Wolke
- Description of this data:** In Wolke et al. we compare the efficiency of different resource allocation strategies experimentally. We focused on dynamic environments where virtual machines need to be allocated and deallocated to servers over time. In this companion paper, we describe the simulation framework and how to run simulations to replicate experiments or run new experiments within the framework.
- Experiment data files:** A list of files for download, including 'Results.zip' (63 KB), 'github.paper.IS2015-master.zip' (8 MB), 'github.workload-master.zip' (222 MB), 'Dockerfile' (1 KB), 'IS2015.tar.gz' (1.3 GB), and 'reprozip.rpz' (160 MB).
- Version 6 | Published: 13 Dec 2015:** A box indicating the current version and publication date.
- Linked to published publication:** A box stating 'This data is associated with the following peer reviewed publication: Reproducible experiments on dynamic resource allocation in cloud data centers'.
- Latest version:** A table showing the latest version (Version 6) published on 2015-12-13.
- Previous versions:** A table listing previous versions from Version 5 to Version 1, with their respective publication dates.
- Version comparison:** A section for comparing different versions, currently showing 'Version 5'.

Researcher in control (embargo / visible)

Linked to published papers – or not

Data citation (DataCite)

Versioning and provenance

<https://data.mendeley.com/datasets/xz6gv65m6d/6>



5. Prototype: Research Data Search

DataSearch

Q

Data-set sources ▾
Data-set types ▾
☰ view results as a list ▾

Found 34406 results

ScienceDirect

Table 2 - Partitioning of rare earth and high field strength elements between titanite and

P.H. Olin & J.A. Wolff

Partition coefficients, ionic radii and model values.

ScienceDirect

Fig. 5 - Partitioning of rare earth and high field strength elements between titanite and pl

P.H. Olin & J.A. Wolff

Mineral/melt DY/DHo vs. temperature for clinopyroxene and titanite. Clinopyroxene-temperature data from Andujar et al. (2008) and Starkel (2008); titanite-temperature data from Andujar et al. (2008), Tiepolo et al. (2002) and Prowatke and Klemme (2005). Neither of the latter two studies presents Ho data; DHo was

Table 3 from GEOCHEMICAL EFFECTS OF DYNAMIC MELTING BENEATH RIDGES:

G W DEVEY, D GARBE-SCHÖNBERG, P STOFFERS, G CHAUVEL & D F MERTZ

RARE EARTH ELEMENT CONCENTRATIONS IN THE KOLBEINSEY RIDGE SAMPLES

ScienceDirect

Fig. 2 - Partitioning of rare earth and high field strength elements between titanite and pl

P.H. Olin & J.A. Wolff

A. Lattice strain model plotted as Log (partition coefficient) versus ionic radius for idealized elements of the same valence state and coordination number, where D0 is the partition coefficient for the optimal ionic radius r0, and EM is the Young's modulus for the site. B. Average of best-fit DREE values for high-Zr titanite, plotted

Data-set source: PetDb

Data-set type: Table

Preview:

Sample	SampleID	IGSN	Method	Material	Co
	BE-N	N/A	ICPMS	N/A	64.6
	BIR-1	N/A	ICPMS	N/A	54.4
	BR	N/A	ICPMS	N/A	51.8

<http://demo-rdm-datasearch-1436039625.eu-west-1.elb.amazonaws.com/indexed#/>
 (prototype under username / password. Upon request)
 search for “rare geochemical ionic liquid” or “mantle calcium variation”

Data submissions

The screenshot shows the 'Upload your submission files' section of the Elsevier submission portal. On the left, there is a sidebar with options: 'Enter manuscript information', 'Upload files' (highlighted), 'Provide additional information', and 'Review & submit'. Below the sidebar is a 'Guide for authors' link. The main content area is titled 'Upload your submission files' and contains a list of instructions:

- Select a File Type for each submission file.
- Mandatory File Types are indicated in the dropdown list.
- The total size of your submission files may not exceed 700MB.
- The Manuscript File size may not exceed 150MB.
- Update the File Order if necessary, then click Save to preserve the new order.

 Below the list is a button 'Select and upload files' with a link 'Upload from arXiv'. The next section is 'Share your research data (optional)', which is highlighted by an orange arrow from the text box on the right. This section explains that authors can make their research data available with their article to help researchers evaluate findings and increase trust. It lists two options:

- Link research data:** If your research data is already hosted in a repository, you can link it to your article here. [Learn more](#)
- Upload research data:** Upload your research data to the data repository, Mendeley Data, where it will be published and citable, and linked from your article. [Learn more](#)

 A note below the second option states: 'This button opens a new window and will not interfere with your submission, as uploading will continue in the background.' At the bottom of the page, there are navigation buttons: 'Previous', 'Save', and 'Save & Continue'.

A new “Share your research data” section is added.

Here the author can either:

1. Link existing dataset

2. Upload (and link) a new dataset to Mendeley Data

Data submissions

Journal / EVISE header

Main menu

Enter manuscript information
Upload your submission files

Upload files
Provide additional information




Review & submit

Upload your research data files

Title of the dataset:

Data for: Lorem ipsum dolor sit amet (your article title)

Contributors: [Jenneke Fokker](#) Dani Kuipers Harriëtte van Delft [+ Add](#)

 IMG_8572.jpg	×
 IMG_8575.jpg	×
 shepperton.png	×


+ [Click or Drop](#) your files here to upload

Description

Dataset publishing
Your dataset will be published on Mendeley Data with a persistent identifier (DOI), and your article will automatically link to it. [Read more](#)

Licensing
Your dataset will be published under a CC-BY 4.0 licence. [Read more](#)

You can set an embargo date

Publication of the dataset
 6 month from now
 12 month from now
 Specific date: 

Continue
Cancel

The ability to create datasets is included within the submission process, seamlessly

Co-submission of Research Elements

Share your research data (optional)

Articles associated with available datasets have a citation advantage. Make your research data accessible by uploading it, linking to it and/or submitting an additional article about your data. [Read more about data sharing options](#)

[Journal X policy section]

Add link(s) to an external repository to make your data more discoverable

If your data is already hosted in a repository you can link it to your article here. [Learn more about database linking](#)

Upload your research data to link and visualize it within your article

You can post your data - including raw and processed data, directly in the Journal X repository in [Mendeley Data](#), making it more discoverable, accessible and citable.

- Your data will be then linked with your article and
- Supported data formats will be visualized within your article automatically. [Read more](#)

All data files can be posted; for visualization, this journal encourages: [3D Molecular Models](#) [3D Neuroimaging Data](#)

I wish to explain why I am not linking to or uploading research data

Submit an additional brief, peer-reviewed article

Describe your data, methods or software in addition to your main article.

If accepted and published, it will be linked automatically to your main article. Select the article type and journal below to submit your article with an easy-to-use template.

Submit data article for publication in Data in Brief

[Learn more about Data in Brief](#)
[See an example article](#)

Submit method article for publication in MethodsX

[Learn more about MethodsX](#)
[See an example article](#)

Save and continue

This may not be the final structure - currently being user tested.

The 10 components for effective research data

and how Elsevier can help

