



PSSOH

Open research data
practices for researchers

Case study in psychology:
Anonimization and sharing research data

Lazarević Ljiljana

Institute for Psychology and Laboratory for research of Individual differences
Faculty of Philosophy



The second PSSOH Conference,
University of Belgrade –
School of Electrical Engineering
Belgrade, November 2, 2019,
<http://pssoh.etf.bg.ac.rs/>

The value of large, prospective data sets



- Human Genome project (<https://www.ncbi.nlm.nih.gov/genome/guide/human/>)
- Sloan Digital Sky Survey (<http://www.sdss.org/>) – detailed 3D maps of the universe
- Human Connectome Project (<http://www.humanconnectomeproject.org/>),
- Allen Brain Atlas (<http://brain-map.org/>),
- Psychiatric Genomics Consortium (<https://www.med.unc.edu/pgc>)

Psychology in the world of open data

- Few would argue that full potential of the dataset can be captured in the manuscript written.
- Some research domains in psychology have a history of public longitudinal studies: ([National Longitudinal Study of Youth](#)) or are beginning to assemble large datasets, e.g., psychotherapy (Owen & Imel, 2016).
- Other [examples](#) are available from the APA website

Still...

„Routine data sharing, defined as the publication of the primary data and any supporting materials required to interpret the data acquired as part of a research study, is still in its infancy in psychology, as in many domains.“ (Martone et al., 2018, AmPsych).

Psychology in the world of open data



Publicly funded research, including the raw data, belongs to the public!

To the extent that researchers' evidence-based knowledge claims rely on data they themselves generated or collected, they should :

- provide access to those data
- or explain why they cannot.



Let's define open data (in psychology)

- Measurements, observations or facts taken or assembled *for* analysis as part of a study and upon which the results and conclusions of the study are based.
- Primary data - raw or minimally processed data that are collected for a study.
- Metadata - attributes of data or a data set.

Common arguments against data sharing

Arguments against data sharing are surprisingly similar across fields, including psychology (Alter & Vardigan, 2015; Eisenberg, 2015; Tenopir et al., 2011).

The key objections can be summarized as follows:

1. Fear for reputation: Someone will use my data against me by finding errors in my data or statistics.
2. Fear of scooping, aka, the “research parasite” (Longo & Drazen, 2016): Someone will do an analysis that I was planning to do, and then they will claim the scientific credit for my work;
3. Fear of harassment: Release of primary data on certain subjects may open the data provider to abuse or prosecution.
4. Too much effort: Who will pay for my time and other expenses for preparing a code book and preparing the data for storage and retrieval/re-use (Baldwin & Del Re, 2016)?
5. No one will understand them: My data are too complicated to understand and making them available may lead to bad science (Longo & Drazen, 2016).
6. No one needs to understand them: My data really don't have any use beyond this study and I've already extracted any meaning from them and published the results
7. The field will stagnate, because no one will collect new data, just re-analyze the old (Roche et al., 2014).

Barriers to data sharing/how to reach open data goal



- „I collected the data and data are exclusively mine (although I might never publish anything)“
- Informed consent and Loss of privacy
- Time and effort it takes to make data ready for sharing
- Lack of perceived validation and recognition for researchers and the research team for their efforts.
- Legal issues

My (precious) data – who has the right to use the data first?



- Researchers preserve the right to use data exclusively – most common practice unfortunately
- „The right of pre-emption“ – researchers who collected the data – 2-5 years embargo
- No pre-emption rights – *first come first served* if you are part of the panel researchers (e.g., GESIS panel)
- Datasets financed by the tax payers or international bodies - *first come first served* (e.g., PISA)



By Gage Skidmore, Gollum, Flickr:

<https://www.flickr.com/photos/gageskidmore/7584125246/in/photolist-cybCd7-VTk4pa-dzvM5X-rk6w2P-rxvgE-ZrDwg7-dwWEhU-4ednfD-dBs2g-4MyKBj-5r2pM-RsKmYT-a6oxxD-6kc7ib-fJB7gs-dqq2ud-pT2asB-5FrttL-68qoR4-56MMWc-6Hvvr3-a3kbbB-aBDZZv-RTvDFi-aAZEoh-mDGFS-baRl1p-9Uc>, CC BY-SA 2.0 Generic

- Secondary exploration – always notify primary authors
- Coauthorship of primary and secondary users – depends of the usage (copy)rights

Informed consent

- *International Ethical Guidelines for Biomedical Research Involving Human Subjects* - recognized as universally applicable.
- Informed consent processes involve three key features:
 - (a) disclosing to potential research subjects information needed to make an informed decision;
 - (b) facilitating the understanding of what has been disclosed; and
 - (c) promoting the voluntariness of the decision about whether or not to participate in the research – possibility to opt out
- Clarity about benefits to the community from whom data were collected:
 - How will the community benefit from the research?
 - Will specific interventions be implemented as a result of the research?
 - Clarity around the purpose and likely outcome of the data collection

Anonymization vs. Pseudonymization

- Anonymized data: “data rendered anonymous in such a way that the data subject is not or no longer identifiable.” (Recital 26, GRPR)
 - Data must be stripped of any identifiable information, making it impossible to derive insights on a discreet individual, even by the party that is responsible for the anonymization.
- Pseudonymization: “the processing of personal data in such a way that the data can no longer be attributed to a specific data subject without the use of additional information.” (Article 4(5) of the GDPR)
 - De-identified data are hold separately from the “additional information”
 - Data only becomes identifiable when both elements are held together.

Incentives for scientists – why should I spend my **valuable** time to share the data?

Making data publicly available is time-consuming



„Badges are simple, effective signals to promote open practices and improve preservation of data and materials by using independent repositories.“

Commentary

Psychological Science
2015, Vol. 26(5) 653–656
© The Author(s) 2015
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0956797614555946
pss.sagepub.com
SAGE

Assessing the Robustness of Power Posing: No Effect on Hormones and Risk Tolerance in a Large Sample of Men and Women

Eva Ranehill¹, Anna Dreber², Magnus Johannesson², Susanne Leiberg¹, Sunhae Sul³, and Roberto A. Weber¹

¹Department of Economics, University of Zurich; ²Department of Economics, Stockholm School of Economics; and ³Department of Psychological and Brain Sciences, Dartmouth College

Received 4/24/14; Revision accepted 8/19/14

META-RESEARCH ARTICLE

Badges to Acknowledge Open Practices: A Simple, Low-Cost, Effective Method for Increasing Transparency

Mallory C. Kidwell^{1*}, Ljiljana B. Lazarević², Erica Baranski³, Tom E. Hardwicke⁴, Sarah Plechowski⁵, Lina-Sophia Falkenberg⁵, Curtis Kennett⁶, Agnieszka Slowik⁷, Carina Sonneleitner⁷, Chelsey Hess-Holden⁸, Timothy M. Errington¹, Susann Fiedler⁸, Brian A. Nosek^{1,8}

1 Center for Open Science, Charlottesville, Virginia, United States of America, 2 University of Belgrade, Belgrade, Serbia, 3 University of California, Riverside, Riverside, California, United States of America, 4 University College London, London, United Kingdom, 5 MaxPlanck Institute for Research on Collective Goods, Bonn, Germany, 6 Mississippi State University, Starkville, Mississippi, United States of America, 7 University of Vienna, Vienna, Austria, 8 University of Virginia, Charlottesville, Virginia, United States of America

* Mallory@cos.io; Mallory.Kidwell@utah.edu

OPEN ACCESS

Citation: Kidwell MC, Lazarević LB, Baranski E, Hardwicke TE, Plechowski S, Falkenberg LS, et al. (2016) Badges to Acknowledge Open Practices: A Simple, Low-Cost, Effective Method for Increasing Transparency. *PLoS Biol* 14(5): e1002456. doi:10.1371/journal.pbio.1002456

Academic Editor: Malcolm R. Macleod, University of Edinburgh, UNITED KINGDOM

Published: May 12, 2016

Copyright: © 2016 Kidwell et al. 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 author and source are credited.

Data Availability Statement: All data and materials

Abstract

Beginning January 2014, *Psychological Science* gave authors the opportunity to signal open data and materials if they qualified for badges that accompanied published articles. Before badges, less than 3% of *Psychological Science* articles reported open data. After badges, 23% reported open data, with an accelerating trend; 39% reported open data in the first half of 2015, an increase of more than an order of magnitude from baseline. There was no change over time in the low rates of data sharing among comparison journals. Moreover, reporting openness does not guarantee openness. When badges were earned, reportedly available data were more likely to be actually available, correct, usable, and complete than when badges were not earned. Open materials also increased to a weaker degree, and there was more variability among comparison journals. Badges are simple, effective signals to promote open practices and improve preservation of data and materials by using independent repositories.

Badges to Acknowledge Open Practices

- Why is this important?
 - Current norms in publishing do not provide incentives for researchers to share data, materials or study designs.
 - Journals can provide this kind of incentive through acknowledging open practices.
- What if open practices are not possible or advisable? (For example, sharing some human participant data could violate confidentiality.)
 - When badge criteria cannot be met, a description in place of the badge can articulate why.
 - **Badges do not define quality of the study; badges certify that a particular practice was followed.**
 - Disclosure makes explicit the conditions under which the ethic of openness is superseded by other ethical concerns.



Legal issues about open data

- Is it safe to upload data to a repository (e.g. OSF, ArXiv)? What about privacy?
- OSF is not demanding hosting data on US servers
 - OSF integrates add-ons from local servers into the OSF.
- Uploaded data are responsibility of the researcher working on the project
- Different countries – different practices
 - UK – [Wellcome](#), RCUK (demand open data)
 - The Netherlands – Data Archiving and Networked Services - [DANS](#) hosts local data and connects with OSF
 - Germany – local repositories, hosting data on US based servers is not in line with all regulations
 - Serbia – still not regulated, we don't have some service providing data hosting on the state level

What are the characteristics of good publicly available dataset?



- FAIR data - data which meet standards of [findability](#), [accessibility](#), [interoperability](#), and [reusability](#).
- Anonymized (privacy protection)
- Communicability (other researcher can use it – common format, in English)
- Meta-data (e.g., DOIs available, authors, date of creation, keywords)
- Long-term availability - permanent links
- Independent institutions who host repositories

Some examples in psychology:

Open Science Framework : <https://osf.io>

Figshare: www.figshare.com

Re3 data: <http://www.re3data.org/>

Example 1



www.nature.com/scientificdata

SCIENTIFIC DATA

OPEN

DATA DESCRIPTOR

Data from the Human Penguin Project, a cross-national dataset testing social thermoregulation principles

Chuan-Peng Hu *et al.**

In the Human Penguin Project ($N=1755$), 15 research groups from 12 countries collected body temperature, demographic variables, social network indices, seven widely-used psychological scales and two newly developed questionnaires (*the Social Thermoregulation and Risk Avoidance Questionnaire* (STRAQ-1) and the *Kama Muta Frequency Scale* (KAMF)). They were collected to investigate the relationship between environmental factors (e.g., geographical, climate etc.) and human behaviors, which is a long-standing inquiry in the scientific community. More specifically, the present project was designed to test principles surrounding the idea of *social thermoregulation*, which posits that social networks help people to regulate their core body temperature. The results showed that all scales in the current project have sufficient to good psychometrical properties. Unlike previous crowdsourced projects, this dataset includes not only the cleaned raw data but also all the validation of questionnaires in 9 different languages, thus providing a valuable resource for psychological scientists who are interested in cross-national, environment-human interaction studies.

Received: 13 November 2018

Accepted: 25 February 2019

Published online: 17 April 2019

Let's explore it a bit: click [here](#)

Example 2



Journal **Journal of Personality Assessment** > Latest Articles

Enter keywords, authors, DOI, ORCID etc

99 Views
0 CrossRef citations to date
2 Altmetric

Articles

Relations Between Lexical and Biological Perspectives on Personality: New Evidence Based on HEXACO and Affective Neuroscience Theory

Goran Knežević ✉, Ljiljana B. Lazarević, Christian Montag & Ken Davis
Received 17 Jun 2018, Accepted 04 Nov 2018, Published online: 09 Apr 2019

Download citation | <https://doi.org/10.1080/00223891.2018.1553782> | Check for updates

Full Article | Figures & data | References | Citations | Metrics | Reprints & Permissions | **Get access**

Abstract

We provide evidence on the convergence of language-based questionnaire and biological perspectives on personality traits. The first study, conducted on Serbian students, provided evidence on the position of Panksepp's Affective Neuroscience Personality Scales (ANPS) in the personality space defined by HEXACO facets. The second, replicatory study was conducted on a sample of German young adults. Results show that the instruments based on these 2 approaches target highly similar personality phenomena, which is revealed in the high canonical correlations between them (the first 3 being above .70 in both samples). Despite the overlap, the scales measuring emotional systems do not map onto HEXACO factors one-to-one, and mostly have substantial loading on more than 1 HEXACO factor. The pattern of correlations between HEXACO and ANPS scales was highly similar in the 2 samples. The importance of the findings for the personality taxonomy and theory is discussed.

Let's explore it a bit: click [here](#)

Thank you for your attention!

All questions and comments are welcome!

Contact: ljliljana.lazarevic@f.bg.ac.rs

website: <https://lira.f.bg.ac.rs/sr/clanovi-lira/dr-ljliljana-b-lazarevic/>