Otvaranje podataka u psihologiji: Korisni saveti za istraživače

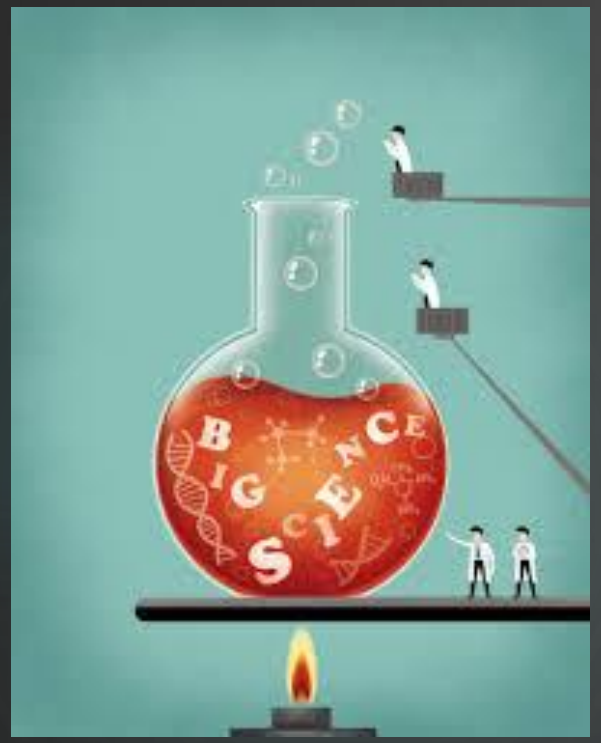
Open data in psychology: useful tips for researchers

LAZAREVIĆ LJILJANA

FACULTY OF PHILOSOPHY, UNIVERSITY OF BELGRADE

Dani otvorene nauke, Beograd, Srbija
Novembar 2020

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)

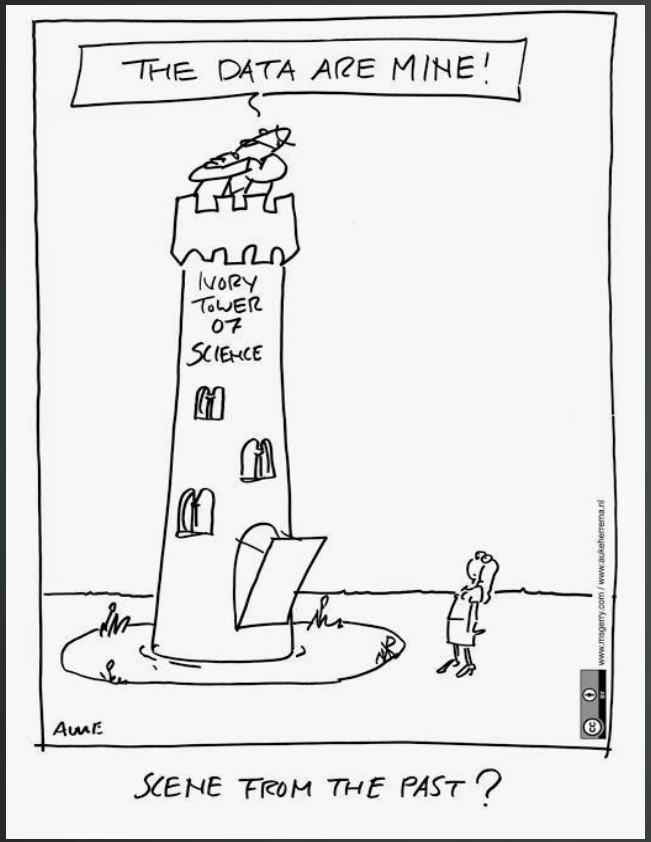
Psychology in the world of open data

- 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

Research transparency

Open data: issues to be solved

Research transparency ↔ privacy rights. Privacy rights have to be respected, and in case of doubt they win over openness. But if data can be properly **anonymized**, there's no problem in sharing.

Data reuse ↔ right of first usage. Optimal reuse of data versus the right of first usage of the original authors. Recommendations allow to extend the right of first usage by an **embargo of 5 more years**. A guideline also defines **how data reuse should be handled**, as well as if **co-authorship** should be offered to the data providers and in which cases this is not necessary.

Verification ↔ fair treatment of original authors. Whenever a reanalysis of a data set is going to be published, **the original authors have to be informed** about this. They cannot prevent the reanalysis, but they have the chance to react to it.

Source: German Psychological Society Guidelines

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.

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 – possiblity 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

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

- Researchers preserve the right to use data exclusively – most common practice unfortunatelly
- „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)

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

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

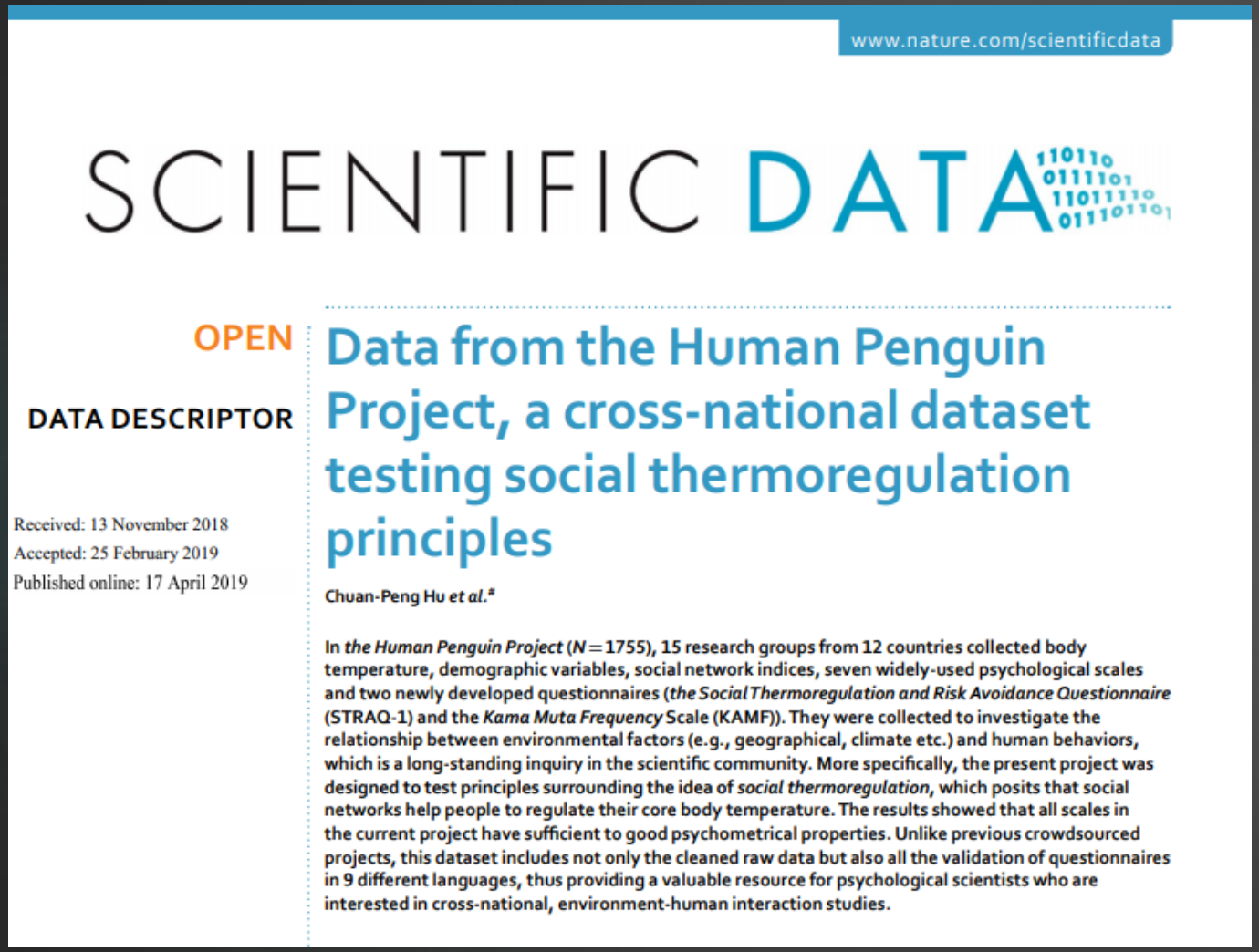
What are the characteristics of good publicly avaliable dataset?

- FAIR data - data which meet standards of findability, accessibility, interoperability, and reusability.

- Anonymized (privacy protection)
- Communicablity (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

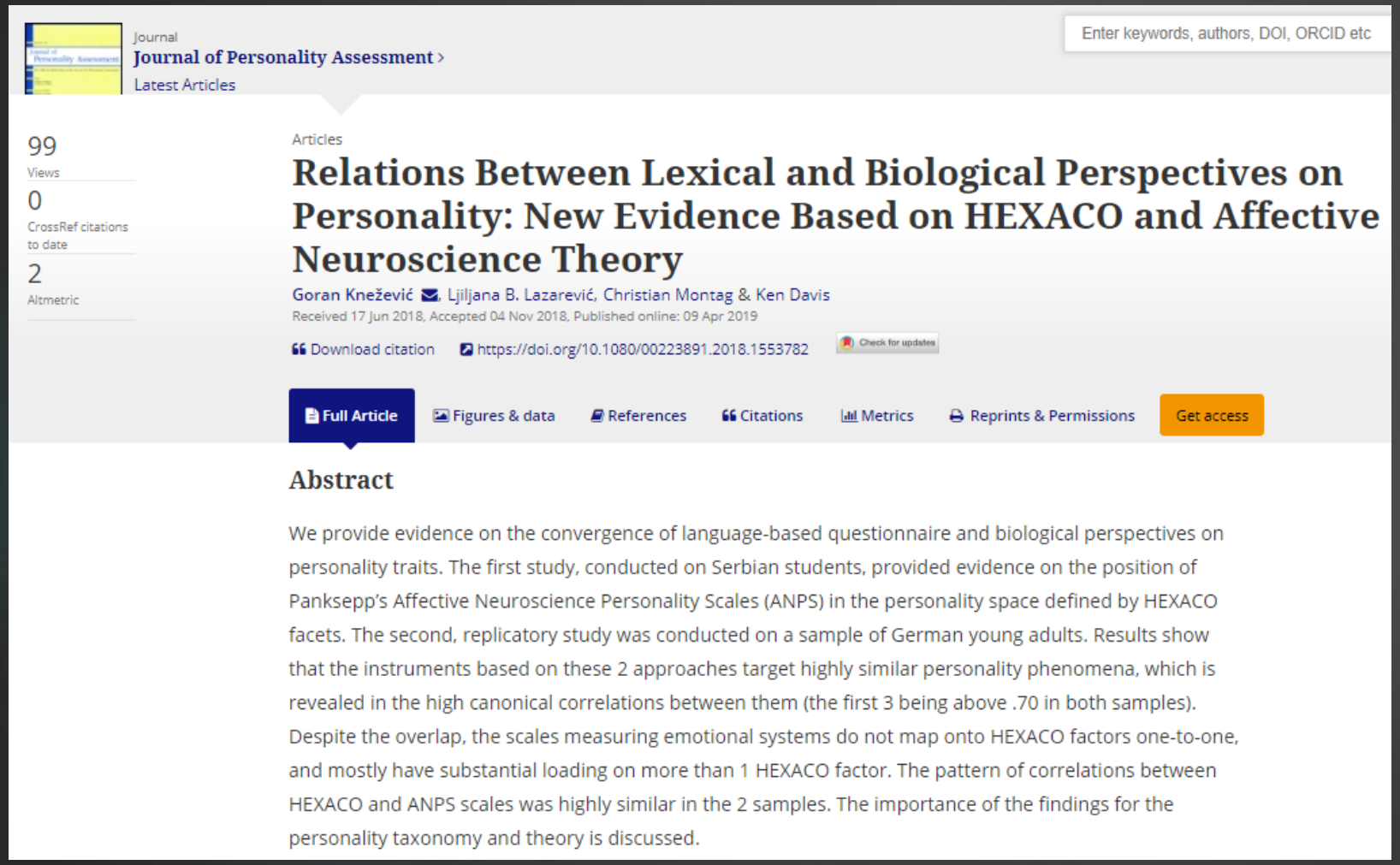
Received: 13 November 2018
Accepted: 25 February 2019
Published online: 17 April 2019

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.

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

Final words

Open science describes the transformations in the way research is being performed: researchers collaborate and knowledge is shared so that everybody can contribute to scientific advancements through a more effective use of research results.

Open science represents a systemic change in the modus operandi of science: open science shifts research from the "publish or perish" mantra to a knowledge-sharing ideal.

However, it shouldn't be portrayed as an utopist movement that doesn't provide clear benefits for the actor involved.

Benefits of open science

Sharing resources from publicly funded research (opposed to reinventing the wheel every time) – economically wiser.

Facilitating access to research data encourages its re-use outside academia – to the interested public, but also by businesses.

Faster exchange of information serves innovation and growth.

Better communication with the public leads to more responsiveness to public needs.

Thank you for your attention!

ALL QUESTIONS AND COMMENTS ARE WELCOME!

CONTACT: LJILJANA.LAZAREVIC@F.BG.AC.RS

WEBSITE: HTTPS://LIRA.F.BG.AC.RS/SR/CLANOVI-LIRA/DR-LJILJANA-B-LAZAREVIC/

"Ova prezentacija je rezultat rada na projektu „Boosting EOSC readiness: Creating a scalable model for capacity building in RDM", koji finansira Evropska unija u okviru projekta H2020-EU.1.4.1.1. EOSC Secretariat br. 831644."

"This presentation results from the project „Boosting EOSC readiness: Creating a scalable model for capacity building in RDM", financed by the European Union, H2020-EU.1.4.1.1. EOSC Secretariat no. 831644."