the WINNOWER

SCIENCE AND SOCIETY

Reproducibility in modern day open science is a keystone beyond intentions

KENNETH VERHEGGEN [1]

1. Ghent University

CORRESPONDENCE:
kenneth.verheggen@gmail.com

DATE RECEIVED:
May 26, 2016

KEYWORDS:
#LJAFreproducibility, data

An ancient martial arts dogma states that one should not envy those who practice 10.000 punches once, but respect those that practice one punch 10.000 times. It is in a way not different in science because as data gets generated at incredible rates, it is often overlooked as to how valuable a repeated observation of an outcome for an experiment is. This phenomenon is not even bound to a certain field of expertise. Indeed there is no possible research that has not dealt with that one magical word that is often swept under the rug in an attempt to hide it and make the overall picture look better. Dear reader, please allow me to once more seek for acknowledgement of a keystone in science and what should become a mantra to anyone that wants to make a change through science : global reproducibility. Assuming that one has mastered the first level of local reproducibility, there is a lot more to be covered. The term itself has an inherent connotation of transparency and this is what is unfortunately still lacking in modern day science. Research in general has to evolve alongside the means and data to become “open science”. Resting aside the near mathematical formulation and experimental setup, I would like to devote this essay on what I believe is of equal if not larger importance.

Reproducibility is often regarded as a purely statistical concept, which is without a shadow of a doubt valuable and ensures that results are consistent and interpretable by anyone in the same objective fashion, even when the results are not in line with what was expected. Historically, this concept has grown to be an integrated part in any lab and has produced many highly valued quality checks and standard operating procedures. This is what I would like to call a local reproducibility, a way to quantify the ability of an experiment or a methodology to be repeated and will deliver the same outcome within a certain margin under equal circumstances by an individual. The clearly defined importance of such local reproducibility has yet to move from a moderately kept secret during research to an open presentation and inviting for not only repeated studies, but also for follow-up investigations. It has to be the intention to be able to pursuit reproducibility outside of the lab, perhaps even outside the topic itself in which the experiment has been performed, hopefully culminating in the actual reuse of results. Aside from repeating an experiment, it is possible to draw new, possibly unrelated, conclusions from existing data, however in order to repurpose these data it needs to be reproducible in its own right. Then and only then can the full potential of data be exploited and can novel science be uncovered. At this point, the reproducibility will have moved on to become global.

The follow-up on existing result-sets are no longer to be deemed “meta-science”, especially considering that there is a lot of valuable information in any experiment that is beyond the original conductor’s intention. Ensuring that these data can be inspected, verified and eventually reused by

anyone that seeks knowledge obtainable through the conduct experiment will lead to true, meaningful reproducibility. There is a genuine benefit to releasing data to the outside world, but off course there are limits to which data can be exposed and which facts and figures should remain protected behind the scenes. It is clear that in a world dominated by industry driven research, it is near impossible to convince a board of directors or a head of research to make all results worth millions if not billions available to John or Jane Doe and to allow him or her to exploit the potential within. Patents are not up for grabs and it is a utopian mind-set to assume science is always for reserved for the greater good. Aside from the economical aspect, it might prove quite difficult to avoid certain ethical topics. For example, releasing private data (such as patient derived data) is and has been up for debate since the earliest forms of science known from history. On one hand it could help large scale research and bring forth huge leaps in innovation and evolution of demographic studies. These studies now often occur behind closed doors and are traditionally kept secret in house, heavily protected by a plethora of juridical guards. But does this mean that data is reproducible?

Despite these ominous ideas, I want to highlight the other options to improve reproducibility of data beyond the actual numbers. There are three keystones that can lead to having a reproducible experiment that will in the end even. At first there is the traditional good laboratory practices and good data practices, they lead to the first tier of reproducibility as is described in many textbooks. The second one revolves around persistency, which is often overlooked post publication. In modern day research it is near mandatory to have data stored in a persistent way, preferably using some kind of versioning. In fact, many journals already instated a policy that mandates the availability of the discussed data upon publication of research. This is true even for intermediate data, which can definitely prove useful, especially during optimisation studies. It is not a farfetched idea that there is someone out there (or will be in the future) that might be working on a similar problem and could benefit from repeating the experiments, perhaps even in silico. Therefore, keeping in mind the potential repurposing of data, online platforms and repositories such as GitHub are to become the bread and butter of scientist in the modern age.

Having data available and resistant to time is but one aspect that should be managed. To achieve an ascended form of reproducibility, one needs to ensure that as much detail as possible is exposed. Having a log of parameters that could influence the experimental outcome is vital to the repurposing of data. Not surprisingly, this is also what hampers meta-science the most. Indeed, the lack of metadata is often prohibiting a successful reproduction of research, which inevitably will lead to incorrect conclusions. To meaningful, reproducible science, nothing is worse than the latter as it undeniably can influence future research and cost funding agencies handfuls of money, which means other research become financially unappealing. Ensuring access to both the experimental data and the related metadata data would also help for reviewers during the peer-review process of manuscripts, as it will definitely aid in understanding the applied science. There to, a tight coupling between results and data is desirable.

So in conclusion, reproducibility will forever be linked to pure performance characteristics. But there is another level that plays a key role, namely the availability to repurpose the data and metadata with awareness of the original experimental setup in terms of a meticulous enumeration of experiment bound parameters. Perhaps it is time after decades of desperately clinging on to data and keeping it stored in a locked box in a private office cabinet, fearing that one might steal it along with the blood, sweat and tears that have been sacrificed, that science moves on to an open world format where the reuse of data is most welcome and over time becomes almost trivial. As a cognitive, scientific community, we would all benefit from this evolution. Novel discoveries and relevant knowledge is hidden beyond what researchers envision when they have set up and conducted an experiment, with the caveat of corporate or clinically managed sensitive data remaining under an economical or ethical veil. Perhaps there is a need to redefine what reproducibility truly is. Starting today, let it be an assessment of how well an experiment can be repeated, not only by researchers themselves but also by the scientific community as a whole, even when the intentions of the latter is beyond the envisioned goal.