IROSiE

Training Materials for Responsible Open Science

Case study 1

Retraction of the paper due to the noncompliance with the journal's data policy

SOURCE: "PLOS ONE retracts perfume study when data don't pass the sniff test." https://retractionwatch.com/2019/09/26/plos-one-retracts-perfume-study-when-data-dont-pass-the-sniff-test/

In July of 2019 PLOS ONe published a paper titled "Social success of perfumes" by Vaiva Vasiliauskaite and Tim S. Evans of the Theoretical Physics Group and Centre for Complexity Science at Imperial College London. In the abstract of the paper the authors wrote: "We study data on perfumes and their odour descriptors—notes—to understand how note compositions, called accords, influence successful fragrance formulas. We obtain accords which tend to be present in perfumes that receive significantly more customer ratings. Our findings show that the most popular notes and the most over-represented accords are different to those that have the strongest effect to the perfume ratings. We also used network centrality to understand which notes have the highest potential to enhance note compositions. We find that large degree notes, such as musk and vanilla as well as generically-named notes, e.g. floral notes, are amongst the notes that enhance accords the most. This work presents a framework which would be a timely tool for perfumers to explore a multidimensional space of scent compositions."

However, the paper was soon retracted by the editors of PLOS ONE due to concerns about the reproducibility of the study and noncompliance with the journal's Data Availability policy. The editors explained their concerns, writing: "After this article was published, questions were raised about the dataset used in the study. In following up on these questions it came to light that the dataset was obtained from a third-party commercial entity whose identity cannot be shared due to a nondisclosure agreement, and that the authors cannot share the raw data or provide clarifications about how the data were collected or processed. The authors posted anonymized summary data on Figshare, as noted in the article's Data Availability Statement. However, the reported Methods are not sufficient to enable other researchers to reproduce the study and the data provided do not meet PLOS ONE's requirements as outlined in our Data Availability policy. The authors noted that they cannot reproduce the analyses using another public dataset as no comparable dataset is currently available."

The authors wrote to Retractionwatch that: "the data is owned by a third party and we had to agree to very tight restrictions in order to use the data. For instance we no longer have access to the original data. So we were very aware of the restrictions when writing the paper. As we want to be as open as possible, we made as much of the data available as we could and this has always been accessible in the repository listed in the references. This was explained to the referees and to the journal before publication. The journal reviewed the situation again after publication and at that point decided the paper did not comply with their open data policy."

Questions for discussion:

1. Who is right in this debate? Are commercial interests and protection of intellectual property legitimate arguments not to share raw data?
2. Why might scientists have reservations about sharing their data?

This project has received funding from the European Union's Horizon 2020 research and innovation programme under GA No 101006430

IRoSiE

Training Materials for Responsible Open Science

Case study 2

Should scientists share the data in climate science?

SOURCE: McAllister, J. W. (2012). Climate science controversies and the demand for access to empirical data. *Philosophy of Science*, *79*(5), 871-880. https://doi.org/10.1086/667871

"Some critics of climate science had long believed that the raw data would reveal little evidence of anthropogenic climate change. They had increased their efforts in previous years to gain the release of data from the CRU [*Climatic Research Unit at the University of East Anglia*], in many cases appealing to the Freedom of Information Act 2000 in the United Kingdom. Climate scientists had come to regard these applications as a campaign to waste their time and to gain material by which to undermine climate research unfairly (Heffernan 2009; Mann 2012, 200–201). Once the correspondence was in the public domain, the critics cited some messages as showing that the climate scientists had conspired to frustrate requests for access to data in order to prevent external scrutiny of their work. [..]

The reports of the subsequent five inquiries dwelled at length on the issue of access to empirical data. Some noted a reluctance of the climate scientists to place data in the public domain and commented on the desirability of sharing scientific data with other scientists and the general public. For example, the report of the House of Commons Science and Technology Committee in the United Kingdom quoted the reply that Phil Jones of the CRU had sent to Warwick Hughes, who had asked for the raw data held by the CRU: "Even if WMO [World Meteorological Organization] agrees, I will still not pass on the data. We have 25 or so years invested in the work. Why should I make the data available to you, when your aim is to try and find something wrong with it." The report commented, "On the face of it, this looks like an unreasonable response to a reasonable request. As Lord Lawson put it: 'Ask any decent scientist and they will say the keystone for integrity in scientific research is full and transparent disclosure of data and methods'" (GCSTC 2010, 12).

The report then summarized further arguments offered by Jones and other witnesses, which suggested that it was neither necessary nor possible for them to release all data: portions of the data were already available from other sources such as the Global Historical Climatology Network in the United States, the CRU was prevented from publishing some of the data by commercial agreements, most scientists preferred to work with adjusted data rather than with raw data anyway, and the CRU had no special duty to provide raw data to the general public. The committee appeared to accept some of these points and sympathized with the frustration that it suggested Jones must have felt in handling requests for data that he believed were motivated by a desire to undermine his work. Nonetheless, the report concluded, "In our view, CRU should have been more open with its raw data and followed the more open approach of NASA to making data available."

Questions for discussion:

1. Who is right in this debate? Is contested and politicized nature of some research fields a legitimate argument not to share raw data?
2. Why might scientists have reservations about sharing their data?

This project has received funding from the European Union's Horizon 2020 research and innovation programme under GA No 101006430

Case study 3

When should scientists share the data?

SOURCE: Barron, D. (2018). How freely should scientists share their data. *Scientific American Blog Network. https://blogs. scientificamerican. com/observations/how-freely-should-scientists-share-their-data*

Jack Gallant is a cognitive neuroscientist at the University of California, Berkeley who works on brain decoding technology. In 2016 he showed, what listening to the Moth podcast does to our brains. Before that he showed that basing only on brain activity it is possible to reconstruct images of movies people are watching. His analysis of Moth podcast was published in *Nature*, he has been interviewed by Freakonomics and NPR. Gallant's work has made him a prominent neuroscientist who runs a successful lab.

In 2018 on July 4 Gallant was promoting open science on his Twitter account. He argued that giving away free code is pointless if it only works within an expensive software system. The next day a theoretical physicist Manilo De Domenico tweeted in reply to Gallant: "Nice advice. But what about data? We keep trying to ask access to data in your Nature 2016, but we received not a single reply, yet". To which Gallant replied that "The original authors are still writing further primary research papers on these data so they haven't been released yet but we expect to be able to do that very soon." Another twitter user - Andre Brown pointed out that "We still want exclusivity to publish more papers' isn't a great excuse. Did you note data restrictions in the manuscript?" and referred to Nature's policy that, on publication, authors should make their data, code and protocols "promptly" and publicly available. Therefore, it appeared that Gallant has violated *Nature's* policy and fundamental principles of open science. De Domenico further complained that Gallant's paper has given him several ideas that he would like to test but not having access to Gallant's data he is not able to do that. To this Gallant answered: "And why do you assume that your project is better than the ones that we are continuing with these data? My students and postdocs are an awesome group of people, the stuff they have in the pipeline is great! But I can't afford for them to be scooped." Gallant then affirmed his commitment to opne science, that he had shared many datasets in the past and fave further explanation of why he has not yet shared this particular data set - that comlex data takes time to understand and his team wanted to work on data more before releasing it and that since his lab has competeed for and has won the money to collect the data and then worked to collect it, they should should be able to work on first before others do it. Many academics on Twitter were not happy about Gallant's answer. They called it a "nonsense excuse", "scandalous", etc. Someone on *Nature's* website wrote that "Jack Gallant refuses to share the data (in violation with Nature's Journal Policy and with his NSF grants)." Some called to boycott Gallant and to retract his paper.

Questions for discussion:

1. Who is right in this debate? Are the objections to Gallant's position justified? What do you think about Gallant's reasons of not sharing the data set? Does he violate the principles of open science?
2. Why might scientists have reservations about sharing their data?

This project has received funding from the European Union's Horizon 2020 research and innovation programme under GA No 101006430