a collection of

OPEN SCIENCE USE CASES

MEET / SHARE / INSPIRE / CARE

THE OPEN SCIENCE USE CASE AWARDS

The Open Science Awards recognize researchers or research students who have used Open Science to make their research more accessible, transparent or reproducible. In the context of the National Open Science Festival a call for Use Cases was published. This call was open to all researchers and Phd students from Dutch universities, UMCs and research institutes. The call was looking for use cases that explored challenges and difficulties as well as positive experiences and successful outcomes. This collection of Use Cases is a result of that call. More info about the Dutch National Open Science Festival can be found at:

www.opensciencefestival.nl

nationaal
programma
open
science

COLOPHON

The National Open Science Festival is organized in the context of the National Programme Open Science.

Date June 2020
DOI 10.5281/zenodo.3903257

Publishing coordination Melanie Imming, SURF
Design www.spresso.nl

The Reviewing Open Science Festival Programme Committee
Prof. Dr. Casper Albers
Kristijan Armeni
Dr. Anna Besse-Lototskaya
Dr. Loek Brinkman
Dr. Maria Cruz
Tess van Doorn
Dr. Anita Eerland
Margo de Groot Coenen
Melanie Imming
Dr. Ir. Bianca Kramer
Esther Maassen
Dr. Carlos Martinez-Ortiz
Prof. Dr. Frank Miedema
Dr. Peter Novitzky
Esther Plomp
Alexandra Sarafoglou
Prof. Dr. Rens van de Schoot
Jeroen Sondervan
Anke Versteeg
Dr. Egon Willighagen
Dr. Anneke Zuiderwijk

The Reading University Call for Use Cases for their Open Research Awards inspired us for this Call.

OPEN SCIENCE
USE CASES

The Use Cases in this Collection are authored by:

Open science practices in Majorana research
André Melo, Sebastian Rubbert, Dr. Anton Akhmerov,
Kavli Institute of Nanoscience Delft,
Delft University of Technology

Studies of Populations of Individuals birds (SPI-Birds) Network and Database
Dr. Antica Culina, Netherlands Institute of Ecology (NIOO-KNAW)

A webtool for interactive data visualization and data sharing
Dr. ir. Joachim Goedhart, University of Amsterdam
Dr. Przemek Krawczyk, Amsterdam University Medical Centers
Dr. Martijn S. Luijsterburg, Leiden University Medical Center

Open science and open data for human factors research
Dr. Pavlo Bazilinskyy, Dr. Ir. Joost de Winter,
Delft University of Technology

cBiT: The Compendium for Biomaterial Transcriptomics
Dr. Dennie Geert Anne Jozef Hebels, MERLN Institute,
Maastricht University
Prof. dr. Jan de Boer, Dept. of Biointerface Science,
Technical University Eindhoven

The Student Initiative for Open Science (SIOS)
Myrthe Veenman, Karoline Huth, Maike Dahrendorf, Lea Schumacher, Sandra Geiger, Iris Smal and 12 other SIOS team members, University of Amsterdam

A practical tool for standardising future costs in economic evaluation
Klas Kellerborg, Meg Perry-Duxbury, Linda de Vries,
Dr. Pieter van Baal, Erasmus School of Health Policy & Management, Erasmus University Rotterdam

Making open psychological datasets more accessible and useful for research and teaching
Dr. Cameron Brick, University of Amsterdam

An open-source, open-participation competition for Fast Radio Burst detection
Dr. Liam Connor, University of Amsterdam
Dr. Joeri van Leeuwen, ASTRON/University of Amsterdam
Dr. Adriënne Mendrik, Dr Alessio Sclocco, Tom Klaver,
Maarten van Meersbergen, Pushpanjali Pawar,
The Netherlands eScience Center
Giuseppe Gianquitto, Dr. Annette Langedijk, SURF

OPEN SCIENCE PRACTICES IN MAJORANA RESEARCH

WINNER
USE CASE
AWARD

NAMES & AFFILIATIONS OF THE TEAM
André Melo, Sebastian Rubbert, Dr. Anton Akhmerov
Kavli Institute of Nanoscience Delft, Delft University of Technology

INTRODUCTION, PROVIDING A BRIEF DESCRIPTION OF THE CASE STUDY

Majorana bound states (MBS) are a promising candidate for building fault tolerant quantum computers due to their non-abelian exchange statistics and topological protection. Most experimental efforts to create MBS require applying strong magnetic fields. While this approach is straightforward to implement, it also limits device performance and scaling. In this publication, we proposed a solution to this problem using an alternative setup that can host MBS without the need for an external magnetic field. We made all of the code and data, necessary to reproduce the simulation results, available online under a permissive license. Additionally, we published the paper on SciPost, a leading open-access journal with open peer review. Finally, we submitted our work to a science reproducibility hackathon, where three teams were able to independently replicate our results.

DESCRIPTION OF THE RESEARCH CONTEXT IN WHICH THE OPEN PRACTICES WERE EMPLOYED

Our group's research revolves around the fields of topological condensed matter and transport in mesoscopic systems. Numerical simulations and algorithm development play a large role in our work. We are enthusiastic about open science and open source software. We maintain several packages (kwant, adaptive, zesje) and online courses (topocondmat, Solid State Physics lecture notes).

WHAT OPEN PRACTICES WERE USED AND WHY

1) We published the code and data necessary to reproduce our results on Zenodo under the BSD 3-clause, which imposes minimal restrictions on their use and distribution. We made sure to thoroughly document the code and computational environment we used, enabling anyone to verify (or even extend) our results. Moreover, because all Zenodo uploads are assigned a DOI, our work will always be easily retrievable and citable.

2) We published the manuscript on SciPost, an open access journal with open peer review. This ensures anyone is able to access our work, provide feedback, as well as read the referee's criticism and how we addressed it.

3) We submitted our paper to ReproHack, a science reproducibility hackathon held at Leiden University. Three participating teams were able to independently download our code and data, and use them to reproduce the results outlined in the manuscript. The teams also provided us with valuable feedback on their experience, which we will incorporate in future publications.

WHAT BARRIERS OR CHALLENGES WERE ENCOUNTERED, AND HOW THESE WERE HANDLED

Properly documenting code and ensuring it is reusable can require a significant investment of time. To minimize effort spent on this we attempted to write clean, tested code from the start. Additionally, we used tools such as version control, continuous integration, Python environments and Jupyter notebooks to put together a robust and transparent simulation workflow.

It was intimidating to publish research code for the first time and to then submit it for scrutiny at a hackathon. However, it was reassuring to realize that most people are appreciative of open science efforts and happy to provide constructive feedback. The response from people both in and out of our field has been overwhelmingly positive and encouraging.

WHAT BENEFITS WERE REALISED, AND FOR WHOM, AS A RESULT OF USING THE OPEN PRACTICES

While writing robust simulation code requires time and effort, it gave us more confidence in our results and ultimately allowed us to progress faster with our research. Writing reusable code makes it easier for other researchers (both internal and external to our group) to build on our results. Currently a master's student in our group is working on extending the results of the paper. Starting from code that is easy to run and modify will allow him to focus on working on new ideas, rather than reproducing old results.

WHAT LESSONS HAVE BEEN LEARNT FROM THE EXPERIENCE

Developing a robust, version controlled simulation pipeline may seemingly slow down research initially, but will ultimately allow for faster progress. Furthermore, it makes it substantially easier for other researchers to build on your results.

The ReproHack feedback showed that we should be even more thorough about commenting our code and data. In particular, we should have been more verbose in our Python notebooks, namely providing more detailed comments and descriptive variable names. The hackathon participants also suggested we consider how to support users who do not have access to a cluster so that they can still engage with the work and partially reproduce our results.

CONCLUSION, SUMMARISING THE MAIN TAKE-AWAY MESSAGE

DEVELOPING AND PUBLISHING HIGH QUALITY CODE IS A WORTHY INVESTMENT OF A RESEARCHER'S TIME: IT ENABLES HIGHER QUALITY RESEARCH, SAVES TIME AND ALLOWS OTHERS TO EASILY BUILD ON YOUR RESULTS.

STUDIES OF POPULATIONS OF INDIVIDUALS BIRDS (SPI-BIRDS) NETWORK AND DATABASE

WINNER USE CASE AWARD

NAME & AFFILIATIONS OF THE TEAM
Dr. Antica Culina
Netherlands Institute of Ecology (NIOO-KNAW)

INTRODUCTION, PROVIDING A BRIEF DESCRIPTION OF THE CASE STUDY

The idea of SPI-birds was born when I spent a good one year trying to identify all of the studies that I could potentially use for my research project. Why would others spend so much time trying to locate populations that are relevant to their work? How about those datasets that are no longer available because the main person collecting the data has retired? Finally, how much time will it take to understand each dataset and to convert them to the same format?

Long- and short-term data on birds have been collected across the globe, ranging from basic breeding data to individual data such as personality scores, genetic information, or parasite loads. Research based on these data has been of a great significance for ecology and evolutionary biology. Although some collaboration has been achieved among researchers who are working on hole-nesting passerines (a key species group with a large number of monitored populations), a focused, centralized, large-scale effort to establish a well-defined community working on wild populations of individually marked birds has been lacking. This is becoming ever more important because the only way to understand, mitigate, and prevent effects of global phenomena (such as climate change or urbanization) on wild populations, is by using data on large spatial and longer temporal scales.

OPEN SCIENCE USE CASES

DESCRIPTION OF THE RESEARCH CONTEXT IN WHICH THE OPEN PRACTICES WERE EMPLOYED

The lack of an overview of monitored bird populations and their attributes (e.g. type of data collected, length of study), combined with the lack of data standards, is hampering collaborations and data exchange and is generating bias in spatial representation of populations in larger-scale studies. Further, there is an increasing need for appropriate and systematic data archiving, facilitated by the open science movement. Such data archiving is a serious problem in many areas of science, including studies on individual animals, where datasets become ‘extinct’ at a fast rate. The SPI-Birds network creates such a much-needed platform for data archiving, standardization and access, as well as to act as a central hub to facilitate scientific collaboration and data exchange. I have initiated this project mid 2019, and it turned out extremely successful with a high degree of community participation. Currently, SPI-Birds host data on 77 populations covering close to 30.000 ha of land with 30.300 nest boxes, with close to 400.000 breeding attempts of 800.000 individuals in cumulative 1700 years.

WHAT OPEN PRACTICES WERE USED AND WHY

In designing the working model for SPI-Birds it was important to consider that many of the researchers collecting the data are not willing to make their data fully open and would not want to participate in the network if Open data were the requirement. Thus, our main goal was to still make data FAIR (findable, accessible, interoperable, and reusable), including archiving the datasets that do not have a sustainable archiving in place. Second, we made all of the other project components open, such as all of our documentation (e.g. standard format) and the code pipelines that convert data formats used by different groups into a standard format.

All datasets in SPI-Birds database are accompanied by rich meta-data that describe the properties of the studied populations, data collected on these populations, as well as field protocols applied. Meta-data also describe the conditions of data use. While for most populations this is 'in agreement with the data owner' some researchers have opted to make their data fully open. In this case, the only condition of data use is to acknowledge the effort of the data owner in collecting the data. Anyone can access the meta-data and search for populations based on these meta-data. Once the population(s) of interest is identified, the user can send the data request to the SPI-Birds database. This request goes to the data owner (if not stated otherwise) and once approved, the data in a standard format are sent to the user.

We also run quality checks on the data (many research groups don't have automated data checks) to highlight the potential errors in the data and improve data quality. These errors are sent to the data owner for verification, and if solved, the records are updated, while keeping version controlled 'old' dataset(s).

WHAT BARRIERS OR CHALLENGES WERE ENCOUNTERED, AND HOW THESE WERE HANDLED

I have encountered two main barriers at the start of the project:

1) Locate all (as many as possible) populations – once the project received enough support, the members themselves 'snowballed' the known populations
2) Convince the data owners that they can only gain by participating, and that their data will be safe with us. This also included highlighting all the benefits of participating in SPI-Birds. Once some 'big' groups gave us support, more and more other groups started to trust us.

This project benefits, first, scientific research as it enhances collaboration and drastically reduces the waste of time in locating and standardizing data, preserves data, and creates equal visibility for all datasets.

It benefits researchers too, especially the data owners because it archives their data, exposes their population to potential collaborators and reduces time they usually need to invest in formatting data for each data request.

WHAT LESSONS HAVE BEEN LEARNT FROM THE EXPERIENCE

Everything can be done by carefully considering and acknowledging all the stakeholders needs, providing a sense of a community and a common goal, and being persistent.

CONCLUSION, SUMMARISING THE MAIN TAKE-AWAY MESSAGE

SPI-BIRDS NETWORK AND DATABASE IS AN ONGOING PROJECT AND WE PLAN TO CONTINUE GROWING, TO ACT AS A DATA HUB ULTIMATELY CONNECTING ALL RESEARCHERS WORKING ON POPULATIONS WITH INDIVIDUALLY MARKED BIRDS. OUR PLAN IS TO EXTEND OUR SCOPE TO HOST NOT ONLY STANDARD BREEDING SEASON DATA, BUT OTHER TYPES OF DATA TOO (E.G., DATA COLLECTED OUTSIDE THE BREEDING SEASON, HORMONAL DATA, BIOCHEMICAL DATA, BEHAVIOURAL DATA, GENETIC DATA, DIET AND FOOD AVAILABILITY DATA).

NEXT, WE ARE CONNECTED WITH OTHER, ONGOING CENTRALIZED EFFORTS TO MAP THE FULL SPECTRUM OF DIFFERENT TYPES OF DATA ON BIRDS THAT CAN COMPLEMENT EACH OTHER. THIS CAN PROVIDE VERY COMPREHENSIVE INFORMATION ON INDIVIDUALS OVER THEIR FULL LIFE-CYCLE.

WE ALSO BELIEVE THAT SPI-BIRDS CAN BE AN EXCELLENT PLATFORM TO ENABLE BETTER RESOURCE ALLOCATION BETWEEN RESEARCH GROUPS. FOR EXAMPLE, WHILE A DATA OWNER MIGHT HAVE THE DATA, HE MIGHT LACK FUNDS TO ANALYSE THEM. ON THE OTHER HAND, A DATA USER MIGHT HAVE FUNDS, BUT LACK THE DATA. IN THIS CASE, PULLING THE RESOURCES (DATA AND FUNDS) TOGETHER IS THE BEST WAY TO ENABLE SCIENTIFIC PROJECTS, COLLABORATION AND PROGRESS.

A WEBTOOL FOR INTERACTIVE DATA VISUALIZATION AND DATA SHARING

WINNER USE CASE AWARD

NAMES & AFFILIATIONS OF THE TEAM

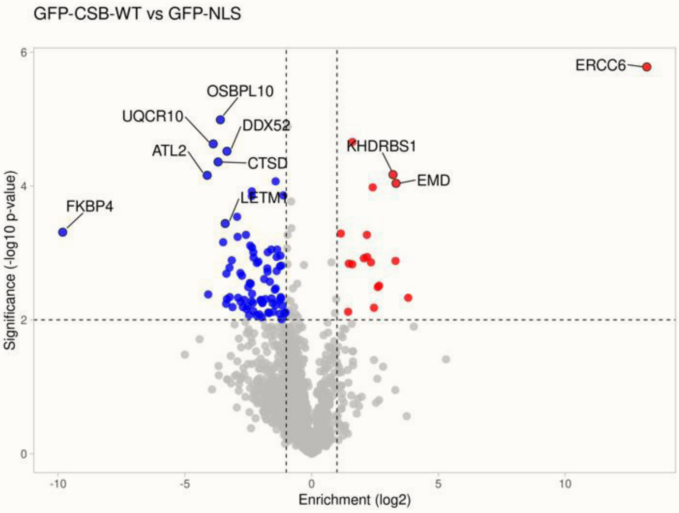
Dr. ir. Joachim Goedhart University of Amsterdam
Dr. Przemek Krawczyk Amsterdam University Medical Centers
Dr. Martijn S. Luijsterburg Leiden University Medical Center

INTRODUCTION, PROVIDING A BRIEF DESCRIPTION OF THE CASE STUDY

Scientific publishing has moved away from traditional print and is almost completely replaced by online, interactive platforms. This revolution is not accompanied, however, by a change in the way that figures or data are published. The underlying reason is that the presentation of results is based on a unidirectional workflow that consists of data acquisition, generating figures and embedding of the figures in manuscripts. As a consequence, figures are only static representations that often do not capture the underlying depth of the experimental results on which they are based. We envision a new, interactive way of publishing results, enabling the audience to inspect the figures, access the data and modify the figures for personal re-use. To facilitate this, we have created online, open-source shiny apps (PlotsOfData, PlotTwist) that can be used to generate the figures and enable user interaction after publication. Now we extend this approach to a new, powerful web app, VolcaNoseR, which enables the interrogation and generation of interactive plots and data sharing for relatively large datasets, such as results of proteomic or genetic screens.

DESCRIPTION OF THE RESEARCH CONTEXT IN WHICH THE OPEN PRACTICES WERE EMPLOYED

Quantitative methods have a prominent role in the life sciences and the (medical) biology domain. The scale at which these methods are applied is rapidly increasing. An example from our own field is the use of large-scale, i.e. genome- and proteome-wide, screens to examine biological systems. The datasets that are generated contain a wealth of information on thousands of components (proteins or genes). The data is often presented as a 'volcano plot' in which only a handful of these components are annotated, as in the example below. In this figure all the other data is inaccessible.



To enable access to, and interaction with all the information and to facilitate the re-use of this information, we have developed an online, free web tool. The web tool is dubbed VolcaNoseR and is available here.

An example of a shareable version of the figure shown above that is generated with the web tool is accessible through this link. The interactive plot that connects the figure with the data is part of a publication that is currently in revision (van der Weegen et al, 2020, Nature Communications).

WHAT OPEN PRACTICES WERE USED AND WHY

The web tool was written using free, open source software (R, shiny) and the code is posted on Github. Releases are archived at Zenodo under the DOI: 10.5281/zenodo.3625857. The web tool is freely accessible, without restrictions. Early versions were announced on twitter and users responded with comments, which were used to improve the webtool. A preprint reporting the web app is in preparation and will be deposited on bioRxiv.

WHAT BARRIERS OR CHALLENGES WERE ENCOUNTERED, AND HOW THESE WERE HANDLED

In our experience some users fail to enter their data, mostly because it contains incompatible characters or values (space in header names or NaN instead of NA). This is partially solved by adding code to the app that replaces incompatible characters and values. Another challenge is speed, as it takes several seconds to launch the app, retrieve the data and generate the plot. This can probably be solved by higher bandwidth connections and more efficient code.

WHAT BENEFITS WERE REALISED, AND FOR WHOM, AS A RESULT OF USING THE OPEN PRACTICES

The web tool allows simple customization of the way large-scale data are visualized, which makes (re)plotting straightforward and provides clear benefits for the researcher by enabling a user-friendly interface to interact with generated data without requiring programming knowledge. More importantly, readers of the publication benefit because they can interact with the figure and have full access to the data. The data can be easily re-used and new plots can be created that can be stored and shared.

WHAT LESSONS HAVE BEEN LEARNT FROM THE EXPERIENCE

To develop a practical and useful web tool, a tight collaboration with frequent communication between developer and user is essential. Since user feedback is crucial, it is also beneficial to share early versions of the tool with potential users. We have used twitter to announce early versions. This has resulted in feedback from several different users which was extremely helpful and motivating.

CONCLUSION, SUMMARISING THE MAIN TAKE-AWAY MESSAGE

USER-FRIENDLY WEB TOOLS LOWER THE BARRIER FOR SCIENTISTS TO GENERATE GRAPHS THAT ARE REPRODUCIBLE AND SHAREABLE. HYPERLINKS TO INTERACTIVE PLOTS ENABLE SIMPLE ACCESS TO BOTH THE FIGURE AND THE DATA.

SIMPLIFYING ACCESS TO DATA ALSO INCREASES THE POSSIBILITY TO INTERACT WITH DATA THROUGH INTERACTIVE PLOTS, WHICH ENABLES STRAIGHTFORWARD RE-USE OF PUBLISHED DATASETS.

THE DEVELOPMENT OF WEB TOOLS, SUCH AS VOLCANOSER, ARE THEREFORE IMPORTANT INSTRUMENTS TO INCREASE THE TRANSPARENCY AND REUSABILITY OF LARGE-SCALE PUBLISHED DATA IN LIFE SCIENCES AND BEYOND.

OPEN SCIENCE AND OPEN DATA FOR HUMAN FACTORS RESEARCH

WINNER USE CASE AWARD

NAMES & AFFILIATIONS OF THE TEAM
Dr. Pavlo Bazilinskyy, Dr. Ir. Joost de Winter
Delft University of Technology

INTRODUCTION, PROVIDING A BRIEF DESCRIPTION OF THE CASE STUDY

Traditionally, human factors research is conducted in a lab. We employ open-science crowdsourcing techniques for research in the domain of automated driving. Namely, we have conducted a series of experiments that feature surveys as well as reaction time and key-press tasks. Our datasets are openly available, to promote online crowdsourcing research with large cross-cultural sample sizes.

Our studies assist in the advancement of online research supported by open data.

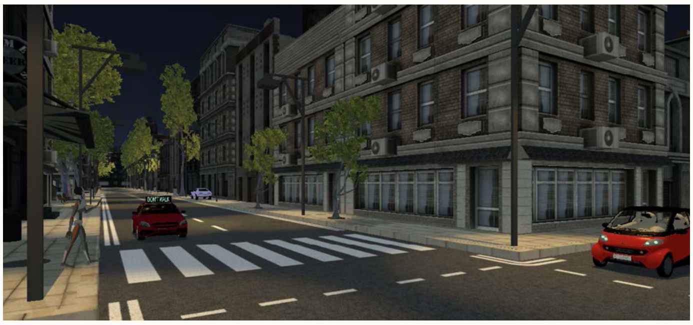


Figure 1. The open-source simulator for traffic research.

DESCRIPTION OF THE RESEARCH CONTEXT IN WHICH THE OPEN PRACTICES WERE EMPLOYED

We work in the very fast-paced domain of automated driving. It is a world with the automotive industry in charge, where questions are answered within months and new challenges are raised constantly. Sometimes, hasty decisions are taken, driven by marketing and fast cycles of releasing a new model. Often such decisions are based on studies involving small sample sizes of engineers employed in the company. We offer a different angle. The general public is the future users of automated driving. We argue that research involving the general public is essential at the current step of development of automated driving, where laws are not yet in place and the opinion of the public has not yet formed. Since 2014, we have published seven journal articles based on crowdsourced data, with another two papers in preparation. The studies range from textual surveys to high-precision reaction time measurement. They feature a gender-balanced sample of at least 1.000 participants from all age groups and representing at least 50 countries. With such large data samples, we were able to derive strong conclusions and even disprove well- cited and accepted findings that had been based on the experiments featuring a dozen of all-male 22 years old psychology students.

WHAT OPEN PRACTICES WERE USED AND WHY

All our articles are available online on services like ResearchGate and supplemented by publicly available code and data stored at the 4TU repository. We are also developing a next-generation open-source simulator for traffic research involving multiple agents. The simulator is already available on Github and free to use by both the scientific community and the public, see Figure 1. The development of the simulator was assisted by a freelance programmer, who was hired through an open and public call. Working with him allowed us to greatly optimise the use of time and resources.

WHAT BARRIERS OR CHALLENGES WERE ENCOUNTERED, AND HOW THESE WERE HANDLED

Once your data is published in a repository like 4TU, it stays there. Preparing the dataset and making sure it is ready for publication is difficult and requires an extra effort.

Not everyone in academia accepts the use of crowdsourcing. The most common piece of criticism concerning the technique that our colleagues express to us is about the "lack of control". During the last years, we have developed a framework to achieve a comparable to the laboratory control of the environment. Such a framework yields competitive results because we have developed mechanisms of asking questions that help to spot participants that engage in the task purely to receive compensation and use such metrics as the time of execution and patterns in collected data to remove participants that had not adhered to the instructions.

WHAT BENEFITS WERE REALISED, AND FOR WHOM, AS A RESULT OF USING THE OPEN PRACTICES

We believe that whenever research is performed in the public domain, all of its outcomes should be public by default. Whenever possible it should be done already from the early stages of the project, and not two years later when the article with results is finally accepted.

When code and data become publicly available, the scientific community does not just receive the researchers' interpretation of gathered data but also means to reproduce the experiment. Which, in the end, is also beneficial for the creators of the dataset, as the likelihood of misinterpretation and ambiguity is lowered.

WHAT LESSONS HAVE BEEN LEARNT FROM THE EXPERIENCE

Conducting open-science and open-data research is challenging. Presenting a crowdsourced study and publishing a dataset upon its completion provides a new arena for questions to the authors and uncovering possible weaknesses and shortcomings of the study. Because of this reason, we learnt to take more care about the study from day one, which benefits all aspects of the project and makes dissemination easier.

Employing crowdsourcing and the principles of open science also help to save tremendous amounts of time and resources. A decade ago conducting a study with a sample size of 3.000 people in one month was infeasible, today it is a reality that should be promoted.

Outsourcing parts of development, as we are practising with the development of the multi- agent simulator, has large potential. The world of academia and especially people in engineering should learn to be more open about their projects and use the skills of qualified professionals in the industry. The throughput of such symbiotic projects is very high.

CONCLUSION, SUMMARISING THE MAIN TAKE-AWAY MESSAGE

WE ARGUE THAT CROWDSOURCING MAY ALLOW FOR LARGE-SCALE HUMAN FACTORS RESEARCH, AND THIS MESSAGE SHOULD BE CONVEYED TO THE ACADEMICS AND THE PUBLIC.

IT IS SILLY NOT TO HAVE DATA AND CODE IN OPEN ACCESS. IN OUR VIEWS, ALL SCIENTIFIC OUTCOMES MUST BE COMPLEMENTED BY PROPERLY FORMATTED AND ANONYMISED DATA AS WELL AS CODE USED FOR ANALYSIS. MAKING DATA AND SOURCE CODE PUBLICALLY ACCESSIBLE IS RELATIVELY EASY IN TODAY'S WORLD OF EVERYTHING BEING CONNECTED AND SHARED. THERE ARE NO EXCUSES NOT TO DO SO.

OF COURSE, NOT ALL HUMAN FACTORS RESEARCH CAN BE CONDUCTED ONLINE. CLASSIC STUDIES THAT REQUIRE CONTROLLED CONDITIONS AND SPECIALISED EQUIPMENT LIKE EEG AND HIGH-PRECISION EYE TRACKERS WOULD REMAIN IN THE LAB FOR YEARS TO COME. HOWEVER, WE ARE READY TO CHALLENGE THIS THESIS AS WELL AND EXPLORE NEW APPLICATIONS OF OPEN SCIENCE IN HUMAN FACTORS RESEARCH.

CBIT: THE COMPENDIUM FOR BIOMATERIAL TRANSCRIPTOMICS

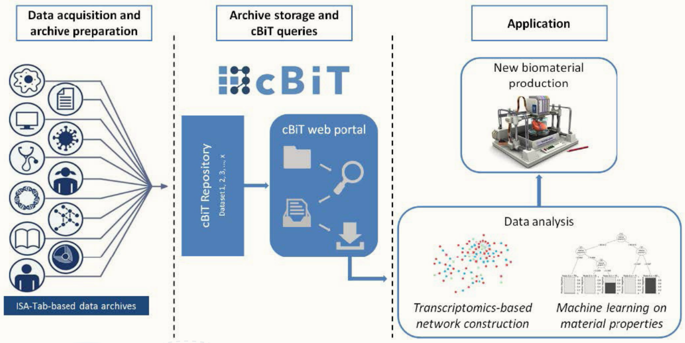
WINNER USE CASE AWARD

NAMES & AFFILIATIONS OF THE TEAM

Dr. Dennie Geert Anne Jozef Hebels, MERLN Institute, Maastricht University
Prof. dr. Jan de Boer, Dept. of Biointerface Science, Technical University Eindhoven

INTRODUCTION, PROVIDING A BRIEF DESCRIPTION OF THE CASE STUDY

The MERLN Institute at Maastricht University and the BiS department at the Technical University Eindhoven are tissue engineering laboratories where we develop new techniques to help the human body repair damage. A key aspect of our research is the development of new biomaterials that integrate better with the surrounding tissue. For years already, we had been using transcriptomics, a technology that studies the expression of thousands of genes at the same time and generates Big Data, to understand how the body reacts to certain biomaterials and how we can improve them. However, the generation of such large amounts of data requires a proper data handling strategy, which was never before systematically implemented in our institutes. Data were scattered, inadequately and inconsistently coded, difficult to compare, and not shared with anyone. This motivated us to develop a publicly accessible data repository to systematically store data while simultaneously making them available to other researchers: The Compendium for Biomaterial Transcriptomics (cBiT: here and here.)



Data acquisition and archive preparation
Archive storage and cBiT queries
Application
cBiT
cBiT web portal
cBiT Repository
Dataset 1, 2, 3, ..., x
ISA-Tab-based data archives
New biomaterial production
Data analysis
Transcriptomics-based network construction
Machine learning on material properties

Figure 1. Data in cBiT are archived in a standardized way, allowing for efficient data analysis strategies.

OPEN SCIENCE USE CASES

DESCRIPTION OF THE RESEARCH CONTEXT IN WHICH THE OPEN PRACTICES WERE EMPLOYED

The transcriptomics community has been one of the front runners in Open Science, publicly sharing research data already for almost two decades, with two key repositories being NCBI's Gene Expression Omnibus and EMBL-EBI's ArrayExpress. However, the obligation or incentive to share transcriptomics data when publishing a paper is still highly dependent on the field the scientific journals operate in. For example, in the toxicology field sharing transcriptomics data (or other types of data) is a widespread policy, but in the tissue engineering field no such policy exists. To make matters even more complicated, the existing transcriptomics repositories are not always able to handle the complex metadata and supplementary data associated with tissue engineering transcriptomics studies (e.g. data on biomaterial classification, material properties, etc.). While the lack of Open Science in the tissue engineering field is slowly starting to change, we noticed that the current situation is holding back the development of new biomaterials that can be used to cure a wide variety of diseases and tissue damage. Transcriptomics generates a vast amount of data and its full potential is never utilized in any single research paper. Keeping these data locked away on an institute's server is therefore a huge waste of effort and money considering that these data sets are often only used once and that the transcriptomics technique is very time consuming and expensive.

WHAT OPEN PRACTICES WERE USED AND WHY

The cBiT repository is publicly accessible and comprises a central warehouse containing data from transcriptomics studies (Figure 1). This enables researchers worldwide to access the data without any restrictions. cBiT data sets are prepared in an interoperable and reusable ISA-Tab format and include all relevant study details (metadata) and data files, implementing established ontology terms wherever possible. The ISA-Tab format is a tab-delimited text format that can be accessed on any computer without the need for specific software that may become obsolete. The transcriptomics data files associated with each data set (both raw and processed) are also based on the tab-delimited text format and other supplementary data files, such as biomaterial measurements, are included in commonly used formats like Microsoft Excel. The ontology terms help to standardize the metadata thereby drastically improving the comparability of studies and removing ambiguity that can lead to confusion. After data set preparation, the data archive is imported into cBiT, given a persistent identifier (Handle ID), and processed and indexed to enable search queries and downloads. Data can subsequently be used for data analysis approaches directed at developing new and improved biomaterials. The Handle ID enables researchers to always find a data set since this identifier is updated whenever a data set is moved to another location. The search query tool makes it easy to find specific data and researchers have the option to filter data sets based on a wide variety of characteristics, such as experimental conditions, type of biomaterial, etc.

WHAT BARRIERS OR CHALLENGES WERE ENCOUNTERED, AND HOW THESE WERE HANDLED

During the exploratory phase of the project, we realized that the standardization in the tissue engineering field is basically non-existent. This required setting up a metadata standard based on official ontology terms wherever possible which proved to be very time consuming and required a lot of discussion with experts in the tissue engineering field and research into existing ontology databases (we used the EBI Ontology Lookup Service. Other smaller challenges were deciding on a suitable persistent identifier (the Handle ID is widely used), creating a user friendly application programming interface, and coming up with a file template that could be imported into cBiT without errors (we contacted specialists to set up the website, focusing on user demands, and used their programming skills to successfully import files).

WHAT BENEFITS WERE REALISED, AND FOR WHOM, AS A RESULT OF USING THE OPEN PRACTICES

By developing cBiT, we managed to switch from data that were scattered, inadequately and inconsistently coded, difficult to compare, and not shared with anyone in the tissue engineering field to a system that offers data to researchers all over the world according to the FAIR guidelines.

WHAT LESSONS HAVE BEEN LEARNT FROM THE EXPERIENCE

What started as a simple idea to improve data sharing of transcriptomics studies among researchers in tissue engineering, turned out to be much more complex to implement. We realized how important it is to make sure data follow one consistent standard. This mostly came to light when trying to determine which ontology terms to use and creating infallible data set templates. Importantly as well, the whole experience has been fun!

CONCLUSION, SUMMARISING THE MAIN TAKE-AWAY MESSAGE

WE PRESENT THE CBIT REPOSITORY AS A FAIR-BASED TOOL TO HELP RESEARCHERS WITH FINDING STANDARDIZED KNOWLEDGE ON THE INTERACTION OF COMMONLY USED BIOMATERIALS WITH DIFFERENT CELL TYPES AND INSIGHT INTO THE UNDERLYING GENE EXPRESSION RESPONSES.

WE ALSO INVITE OTHER RESEARCHERS TO ADD THEIR DATA TO CBIT, THEREBY BECOMING THE GO-TO RESOURCE FOR BIOMATERIAL- ASSOCIATED TRANSCRIPTOMICS DATA.

IN DOING SO, WE EXPECT TO INCREASE THE RE-USE OF EXISTING DATA SETS AND MAKE A MAJOR CONTRIBUTION TO A MORE EFFICIENT DEVELOPMENT OF NEW AND BETTER MATERIALS THAT SHOW IMPROVED INTEGRATION IN THE HUMAN BODY.

STUDENT INITIATIVE FOR OPEN SCIENCE (SIOS)

ENCOURAGEMENT AWARD

NAME & AFFILIATIONS OF THE TEAM

Myrthe Veenman, Karoline Huth, Maike Dahrendorf, Lea Schumacher, Sandra Geiger, Iris Smal and 12 other SIOS team members University of Amsterdam

INTRODUCTION, PROVIDING A BRIEF DESCRIPTION OF THE CASE STUDY

The idea for the Student Initiative for Open Science (SIOS) began over a year ago with the staunch realization that many of the papers and findings taught in our undergraduate courses cannot be replicated. The urge to do something about it, further developed after taking a course about good (and questionable) research practices in which we learned about all the exciting solutions and initiatives that are offered by the open science movement. We thought that such knowledge should not only be given to us Research Master students but need to be made more available to a wider population of students. While most other open science (OS) initiatives focused on academics and professors, at the time, there were not a lot of resources for students. Consequently, we founded the Student Initiative for Open Science (SIOS) which provides information in many forms (talks, lectures, online material, blogs, workshops) for students on open science . We aim to widely promote information about issues that are currently impeding the validity of current research and offer advice on what to look out for in assessing the quality of research. Furthermore, we want to inform about how OS might be able to solve such issues and how students, themselves, can engage in OS practices. In this, we want to strongly emphasize the students' perspective and the applicability to their lives. We believe by promoting OS to students at the early stage of their career, we might be able to improve research practices in science in the long-term.

DESCRIPTION OF THE RESEARCH CONTEXT IN WHICH THE OPEN PRACTICES WERE EMPLOYED

When providing information on OS, we focus on its applicability within the student context.

WHAT OPEN PRACTICES WERE USED AND WHY

These are some of the main events that we organized to promote OS within the student community:

1) TWO TALKS INTRODUCING OPEN SCIENCE:
In these talks, we discussed causes of the replication crisis, how OS could solve these and why it is relevant to students. Here, no prior knowledge was needed and we focussed on how OS can be beneficial for students specifically and how they can implement OS practices within their studies. We are holding these low-level talks to encourage students who might not even have heard about OS to become engaged with these topics.

2) A TALK ABOUT THE DISTINCTION OF CONFIRMATORY VERSUS EXPLORATORY RESEARCH:
In this talk, Alexandra Sarafoglou explained the difference between confirmatory and exploratory research and why it is important to distinguish them. Here, she had an emphasis on what this distinction means for student thesis projects.

3) A TALK ABOUT THE INCONVENIENT TRUTH OF THE P-VALUE:
In this talk, Eric-Jan Wagenmakers illustrated drawbacks of using p-values and introduced Bayesian hypothesis testing as a valuable alternative. We organized this talk to encourage students to critically reflect on what they learn in their statistics classes and become aware of the drawbacks of currently common statistical practices.

4) A PANEL DISCUSSION ON “OPEN SCIENCE PUT TO PRACTICE”:
For this panel discussion we invited three PhD students from different departments and a methodological student advisor to discuss the benefits and challenges of applying OS practices within the different areas of Psychology. We organized this event as we wanted students from various psychological backgrounds to see how OS can be applied in their specific field.

5) TWO OPEN SCIENCE MOVIE NIGHTS:
Here, we organized a “public viewing” of different OS related videos with subsequent discussions. We thought this would be a nice way to encourage students to utilize OS online resources.

6) PRE-REGISTRATION WORKSHOP:
In this workshop, we showed how students can pre-register their thesis on osf and why this can be beneficial for them. We aimed to give them practical skills so that they can implement pre-registration more easily in their thesis.

WHAT BARRIERS OR CHALLENGES WERE ENCOUNTERED, AND HOW THESE WERE HANDLED

The main challenges we had were in regard to funding and getting a large enough team to organize all the events. In regards to funding, we handled it by organizing events that did not require money (as luckily, speakers did not request payment and we could use rooms of the University of Amsterdam). With the start of the academic year, we were able to recruit more students to join our team, especially after the “Introduction to Open Science” lecture.

WHAT BENEFITS WERE REALISED, AND FOR WHOM, AS A RESULT OF USING THE OPEN PRACTICES

The more we looked at OS from the student perspective, the more we realized how much they can benefit from these ideas. For example, although pre-registering your thesis causes more work at the start, the subsequent process can be much smoother as a pre- registration is basically a very concrete plan. Further, if more researchers engage in OS practices, this can also help students to a large extent. For example, access to literature is a crucial point and shared study materials, code and data can help students to build their own research thesis.

WHAT LESSONS HAVE BEEN LEARNT FROM THE EXPERIENCE

Students are very open and excited about OS ideas and practice. It can encourage them to further belief in research even after the many flaws that the replication crisis revealed.

CONCLUSION, SUMMARISING THE MAIN TAKE-AWAY MESSAGE

ALTHOUGH OS MIGHT NOT SEEM APPLICABLE TO STUDENTS AT FIRST SIGHT, IT IS A GROUP THAT WOULD HIGHLY BENEFIT FROM A BETTER KNOWLEDGE OF OS.

THIS ALSO SUMMARIZES THE FEEDBACK WE GET FROM STUDENTS. OS CAN SEEM SCARY AND ABSTRACT FOR THEM BUT ACTUALLY HAS PRACTICAL APPLICATIONS TO THEIR STUDENT LIFE.

FURTHER, IF WE LAY THE GROUNDWORK FOR OS IDEAS AND PRACTICES EARLY IN PEOPLE’S CAREER, THIS HAS THE OPPORTUNITY TO CHANGE RESEARCH PRACTICES IN THE LONG RUN.

A PRACTICAL TOOL FOR STANDARDISING FUTURE COSTS IN ECONOMIC EVALUATION

HONORABLE MENTION

NAME & AFFILIATIONS OF THE TEAM

Klas Kellerborg, Meg Perry-Duxbury, Linda de Vries,
Dr. Pieter van Baal Erasmus School of Health Policy & Management
Erasmus University Rotterdam

INTRODUCTION, PROVIDING A BRIEF DESCRIPTION OF THE CASE STUDY

We present (currently) two open access tools, made in Shiny in R Studio, both under the name of 'PAID' for the Netherlands and the UK. They are designed to promote and aid the standardization of applying future unrelated costs in economic evaluation. This will be expanded to at least two more tools for Germany and Greece.

DESCRIPTION OF THE RESEARCH CONTEXT IN WHICH THE OPEN PRACTICES WERE EMPLOYED

When deciding on what interventions to reimburse within the healthcare budget, decision makers often refer to the cost-effectiveness of said interventions. This is estimated using economic evaluation, in which costs and benefits of an intervention are estimated and compared. In the Netherlands the guidelines for such evaluation requires the inclusion of future unrelated medical costs. In the UK guidelines are currently under review, and by providing an easy way for researchers to access these costs, it may be that the governing body (NICE) will be more inclined to include these costs.

OPEN SCIENCE USE CASES

WHAT OPEN PRACTICES WERE USED AND WHY

Firstly, the tools are made using Shiny in R Studio, open access software that can be used by anyone. The authors worked collaboratively on the script for the tools. The key open element of these tools is specifically that they are available online, for free and for anyone to access (here and here). The tools are already available in beta test mode so that we can benefit from the experiences and criticisms from first time users.

We are making the tools open access because as researchers we believe that scientific advances should be available to everyone, especially given that we are funded by public money. By making the tools open, more researchers will use them in their work, hopefully furthering our aim which is to include future costs in economic evaluation.

WHAT BARRIERS OR CHALLENGES WERE ENCOUNTERED, AND HOW THESE WERE HANDLED

One main challenge was writing R script with several people – it was easy to open the wrong version of the script, or delete someone's new section of code due to communication breakdowns. This could have been better handled by using a version-control platform (e.g. Github) from the beginning. This is something we are still considering incorporating

Since development, one of the authors is no longer available to work on this project, meaning more time is required of other authors to troubleshoot (we are still in beta testing). We are handling this by maintaining communication with the aforementioned co-author and by bringing new people on board of the project. As these authors will eventually lose their university access we also intend to move our script and papers to the Open Science framework to continue our collaboration.

WHAT BENEFITS WERE REALISED, AND FOR WHOM, AS A RESULT OF USING THE OPEN PRACTICES

One of the first benefits is that students are already using the PAID tools. The costs estimated by the tool are being used in MSc thesis along with scientific articles. This is also a benefit to us, as students have already found some small issues and bugs in the tools that we didn't notice. Another benefit for us as researchers in an academic setting is that by using Shiny in R, the software costs are non-existent, and there is a plethora of useful information on the internet when problems arise.

One of our authors recently attended a congress at which a consultancy firm informed her that they recommend using PAID and/or its approach to ministries trying to set up their Health Economics departments. We can only hope they benefit from this by realising it is possible to standardize and include future costs.

We were approached at conferences with questions about the accessibility of the tools and people were pleasantly surprised that we were providing them for free and for anyone to use. Therefore, we hope that researchers carrying out economic evaluations already can benefit from these tools, as they can now easily incorporate standardized future costs into their models, which otherwise would be hugely time-consuming.

WHAT LESSONS HAVE BEEN LEARNT FROM THE EXPERIENCE

1. Using open software means there is a lot of free support online.
2. Communication, specifically with regards to code writing, is more complicated than it seems at first. Using version-control software and other open-science practices could have streamlined the process and left us with fewer mistakes to clean up.
3. Having an open-access tool means that researchers and students (and others) immediately start using it. This means there are more people beta-testing the tool and that our scientific aims would be accomplished sooner (we hope).
4. In general, creating tools rather than just papers gives scientific work a more tangible output which then furthers the original aims of the research.

CONCLUSION, SUMMARISING THE MAIN TAKE-AWAY MESSAGE

TO CONCLUDE, WE CREATED A SET OF OPEN-ACCESS TOOLS SO THAT RESEARCHERS IN ECONOMIC EVALUATION CAN USE STANDARDIZED FUTURE COST ESTIMATES IN THEIR MODELS. BY MAKING THESE TOOLS OPEN-ACCESS THEY ARE IMMEDIATELY BEING USED BY RESEARCHERS, STUDENTS AND CONSULTANTS.

THE AUTHORS ALSO BENEFIT AS THERE ARE MORE USERS PROVIDING FEEDBACK ON THE TOOL. WE BELIEVE THAT BY MAKING RESEARCH OUTCOMES OPEN, EVERYONE BENEFITS AND SCIENTIFIC PROGRESS WILL BE MADE FASTER.

WE REALISE THAT BY INCORPORATING MORE OPEN-SCIENCE PRACTICES, SUCH AS VERSION-CONTROL VIA GITHUB, WE COULD HAVE AVOIDED SOME HUMAN ERROR AND IMPROVED COMMUNICATIONS BETWEEN AUTHORS.

MAKING OPEN PSYCHOLOGICAL DATASETS MORE ACCESSIBLE AND USEFUL FOR RESEARCH AND TEACHING

HONORABLE MENTION

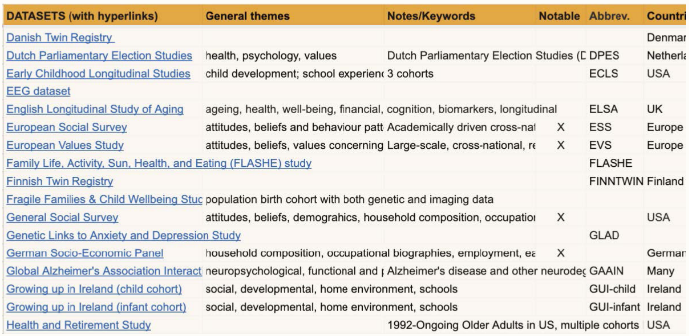
NAME & AFFILIATIONS OF THE TEAM
Dr. Cameron Brick
University of Amsterdam

INTRODUCTION, PROVIDING A BRIEF DESCRIPTION OF THE CASE STUDY

MAKING OPEN PSYCHOLOGICAL DATASETS MORE ACCESSIBLE AND USEFUL FOR RESEARCH AND TEACHING

I have been frustrated that psychologists are not using large, existing, free data to its potential. We don't know what exists, what it contains, or how to use it without a huge investment. In both thesis supervision and research, social scientists mostly collect new datasets that are frequently unsatisfactory in measure quality, sample size, or population.

Most researchers are dimly aware that there are larger-scale, high-quality open datasets available for secondary analysis, but the existing tools for discovery rely on complex websites and often yield too many irrelevant results. A lot of time is needed just to find out which datasets exist and what themes they cover. As this was a barrier to me after years of research, it was likely also a barrier to other social scientists. I believe that better access to existing datasets can be useful both in research and in teaching undergraduates and postgraduates.



DATASETS (with hyperlinks)	General themes	Notes/Keywords	Notable	Abbrev.	Countri
Danish Twin Registry					Denmar
Dutch Parliamentary Election Studies	health, psychology, values	Dutch Parliamentary Election Studies (D		DPES	Netherla
Early Childhood Longitudinal Studies	child development; school experienc	3 cohorts		ECLS	USA
EEG dataset					
English Longitudinal Study of Aging	ageing, health, well-being, financial, cognition, biomarkers, longitudinal			ELSA	UK
European Social Survey	attitudes, beliefs and behaviour patt	Academically driven cross-nat	X	ESS	Europe
European Values Study	attitudes, beliefs, values concerning	Large-scale, cross-national, re	X	EVS	Europe
Family Life, Activity, Sun, Health, and Eating (FLASHE) study				FLASHE	
Finnish Twin Registry				FINNTWIN	Finland
Fragile Families & Child Wellbeing Stuc	population birth cohort with both genetic and imaging data				
General Social Survey	attitudes, beliefs, demograhics, household composition, occupatio		X		USA
Genetic Links to Anxiety and Depression Study				GLAD	
German Socio-Economic Panel	household composition, occupational biographies, employment, ea		X		German
Global Alzheimer's Association Interact	neuropsychological, functional and	Alzheimer's disease and other neurodeg		GAAIN	Many
Growing up in Ireland (child cohort)	social, developmental, home environment, schools			GUI-child	Ireland
Growing up in Ireland (infant cohort)	social, developmental, home environment, schools			GUI-infant	Ireland
Health and Retirement Study		1992-Ongoing Older Adults in US, multiple cohorts			USA

Partial screenshot of the resource

DESCRIPTION OF THE RESEARCH CONTEXT IN WHICH THE OPEN PRACTICES WERE EMPLOYED

The resource I developed is being used by the social science community, particularly supported by the Society for the Improvement of Psychological Science conference (Rotterdam, 2019) and discussion on Twitter (e.g., this popular tweet).

WHAT OPEN PRACTICES WERE USED AND WHY

ACCESSIBILITY OF RESEARCH

I collated a list of openly available psychological datasets. The lack of funding for a bespoke platform turned out to be an advantage. I adopted a service that would be trivial for others to access and edit—a Google Docs spreadsheet.

I listed datasets and descriptions of what they contained rather than collate the data itself. I then organised a hackathon session at the Society for the Improvement of Psychological Science (SIPS) 2019 conference in Rotterdam to give others a chance to search for usable datasets and add them to the list. Laura Botzet, Cory Costello, Anatolia

Batruch, Ruben Arslan, Melissa Kline, Nicolas Sommet, Tobias Dienlin and Hannah Metzler joined me in a successful effort to expand the list and add metadata on themes, keywords, study populations, etc. The hackathon group and other contributors have already identified a range of useful datasets that I had never heard of. The current resource has 124 primary sources, and 43 additional lists of datasets.

WHAT BARRIERS OR CHALLENGES WERE ENCOUNTERED, AND HOW THESE WERE HANDLED

In the spirit of transparency, the list is openly editable for others who discover additional datasets or who could help fill in the blank cells. In addition, this transparency has led to open discussion (largely on Twitter) about the validity of incorporating some datasets in the list, and in one case, debate led to consensus about the removal of an OkCupid dataset because the participants had not provided informed consent.

As the list developed, we also incorporated a sheet with other lists of datasets and journals that publish datasets. Imposing some structure was appropriate to avoid the problem of having thousands of small datasets in the main view, most of which would have no metadata and not be relevant to most users.

On the downside, the list is not a comprehensive list of psychological datasets, and the metadata is incomplete. The flat structure of a spreadsheet also implies similar value between rows, but the entries vary hugely in quality and sample size. A disadvantage of our list (and most datasets) is not being machine-readable. See the exciting Psych-DS project on that topic.

WHAT BENEFITS WERE REALISED, AND FOR WHOM, AS A RESULT OF USING THE OPEN PRACTICES

The main benefit is being able to more quickly navigate existing data in support of research and teaching. There are publications lately that use pre-existing data even in leading journals like Psychological Science, e.g., DOI 10.1177/0956797616660078.

I initially compiled this list for personal use and then realized the benefit of sharing it publicly. Other scientists I'd never met have contacted me to express their appreciation and describe how they are using it. The list now has its own DOI and in personal communication researchers have expressed the intention to cite it. I don't have a record of page views or usage because of the platform, but any time I open the document there are multiple other users present, suggesting wide use across this past year.

WHAT LESSONS HAVE BEEN LEARNT FROM THE EXPERIENCE

In part because of this project, I have started giving invited talks on reproducibility and Open Science within my subject area. This is beneficial to me as a scientist, and was a direct result of giving away my work. For me, this experience has helped illustrate a career path based on transparency, sharing, and reproducibility.

Besides easier access to existing data, a key benefit of this project was the shared experience of participating in an open, collaborative community where tools are happily shared. It is not obvious that giving away one's work can lead to a better outcome even for the individual, but as with open-source code in software development, in my case sharing led to helping others and also dramatically improved the resource. As I develop similar tools, I will definitely share them publicly.

CONCLUSION, SUMMARISING THE MAIN TAKE-AWAY MESSAGE

I AM APPLYING FOR THIS AWARD TO GENERATE ATTENTION AND PUBLIC ENGAGEMENT TO HELP IMPROVE THE LIST AND METADATA. ANYONE IS WELCOME TO CONTRIBUTE, AND THEY DO NOT NEED PERMISSION NOR A LOG-IN TO BEGIN. THE LIST 'MAKING FREE, OPEN PSYCHOLOGICAL DATASETS MORE ACCESSIBLE FOR RESEARCH AND TEACHING' CAN BE EDITED DIRECTLY ON GOOGLE DOCS.

IN AN IDEAL WORLD, A RESEARCH ASSISTANT OR STUDENT INTERN COULD DEVOTE PROJECT TIME TO IMPROVING THE METADATA, AND MIGHT EVEN PRODUCE EVIDENCE GAP MAPS AND GUIDES FOR SPECIFIC CLASSROOM ASSIGNMENTS. HAPPY DATA HUNTING.

AN OPEN-SOURCE, OPEN-PARTICIPATION COMPETITION FOR FAST RADIO BURST DETECTION

HONORABLE MENTION

NAME & AFFILIATIONS OF THE TEAM

Dr. Liam Connor University of Amsterdam
Dr. Joeri van Leeuwen ASTRON/University of Amsterdam
Dr. Adriënne Mendrik, Dr. Alessio Sclocco, Tom Klaver,
Maarten van Meersbergen, Pushpanjali Pawar The Netherlands eScience Center
Giuseppe Gianquitto, Dr. Annette Langedijk SURF

INTRODUCTION, PROVIDING A BRIEF DESCRIPTION OF THE CASE STUDY

Fast Radio Bursts (FRBs) are a mysterious new class of extragalactic explosions that travel to us from billions of light years away, ending their journey at radio telescopes like Apertif, here in the Netherlands. Many astronomy groups from all around the world are trying to detect FRBs, but are using different search algorithms written in various programming languages implemented on different computing architectures. These search algorithms were never validated on the same data, using the same metrics and ground truth, hindering direct algorithm comparison.

In order to standardize the process of searching for these enigmatic bursts, we have created a fully open-source software benchmark, sort of like an FRB Olympics, to gain insight into the performance of FRB search algorithms. This FRB detection benchmark is hosted on the Enhance-Your-Research Alliance (EYRA) benchmark platform (eyrabenchmark.net) set-up by the Netherlands eScience Center and SURF. It allows the international community to benchmark their algorithms and improve their code in order to optimally discover new FRBs.

OPEN
SCIENCE
USE CASES

DESCRIPTION OF THE RESEARCH CONTEXT IN WHICH THE OPEN PRACTICES WERE EMPLOYED

We assembled a team of about 15 software and algorithm-focussed astrophysicists from around the world to participate in the benchmark. In order to do so, their software had to be made public and their packages needed to be containerised using Docker; this alone was a major contribution to the field of FRB discovery.

The FRB detection benchmark is a use case for the EYRA benchmark platform that is actively under development by the Netherlands eScience Center and SURF. A back-end was built that allows users to submit their containerised software packages to search for FRBs in a standardised dataset, but can also be re-used by other benchmarks on the platform. Results are shown on the interactive website that allows for visualisations "beyond the leaderboard" to gain additional insight (using observablehq.com). All code is openly available on github. For the FRB detection benchmark, we generated large datasets with thousands of simulated FRBs injected into them, which the participating software packages then search through, and report their results. The results are evaluated, and fed to our open-source visualisation and analysis packages.

WHAT OPEN PRACTICES WERE USED AND WHY

The FRB detection benchmark is openly available on the EYRA benchmark platform. The benchmark code, including scripts for running the benchmark and visualising the data, is open source and available on github and observablehq.

All the participating software packages for detection algorithms are open source and available. And importantly, the benchmark is open to any further participant who wants to submit their algorithm. This will provide an opportunity for scientists to learn from each other, to better be able to detect Fast Radio Bursts.

WHAT BARRIERS OR CHALLENGES WERE ENCOUNTERED, AND HOW THESE WERE HANDLED

One of the challenges in our benchmark has been managing an international community of astronomers. Many code authors are busy faculty members with teaching duties, so organising the group and getting them to contribute time has been a challenge. While all developers agreed that it would be beneficial to the community to release their code and run this challenge, making it happen in practice required a bit of persuasion.

Another challenge is making platforms, like the EYRA benchmark platform, sustainable. For 2019, we had funding from the alliance project at the Netherlands eScience Center and SURF. Dr. Adriënne Mendrik started the company EYRA, with a non-profit subsidiary EYRA Nova that aims to make the platform sustainable for scientists, such that the benchmark can stay open for future submissions.

WHAT BENEFITS WERE REALISED, AND FOR WHOM, AS A RESULT OF USING THE OPEN PRACTICES

The first great benefit that was realised by our international benchmark was the necessity of all software packages being made public, and each code author putting their software into Docker containers. The community can now install and use all of the relevant FRB search packages very easily, something which almost certainly would not have happened if not for our platform's open-source ethos. The beneficiaries of this are the many early-career astronomers and eScientists who want to use and improve current FRB search algorithms. The fact that the benchmark itself is open means that anybody who wants to participate is welcome to submit their own algorithm.

For example, if an undergraduate in China thinks she has a better FRB-detection package, but does not know anybody in the field, she can submit her own code to the benchmark, as well as check other participants' code.

Finally, future users are a major beneficiary of the open practices associated with our bench- mark. This includes junior scientists from other fields who will be able to look at how our soft- ware benchmark was set-up. This also includes whoever wants to continue our FRB benchmark and expand its scope. For example, the Netherlands has committed €30M to the largest-ever radio telescope, one of whose central science goals is the detection of FRBs. That effort will require benchmarks like ours, and will likely even include our specific software challenge.

WHAT LESSONS HAVE BEEN LEARNT FROM THE EXPERIENCE

One of the central lessons we have learnt from this experience is: Code ought to be written under the assumption that it will be open-source, from the very beginning. Writing a software package for FRB detection that will only be used by you or your group almost guarantees that it will not have sufficient documentation, unit testing, and readability. If one knows in advance that it will be submitted to a benchmark in the future, or used by scientists even after you move on from that project or field, the source code will inevitably be more sustainable.

CONCLUSION, SUMMARISING THE MAIN TAKE-AWAY MESSAGE

WE HAVE SUCCESSFULLY PUT TOGETHER AN INTERNATIONAL SOFTWARE BENCHMARK FOR THE DETECTION OF FAST RADIO BURSTS. THE PROJECT IS OPEN ON MANY LEVELS, INCLUDING OPEN PARTICIPATION, OPEN SOURCE CODE, AND FULL TRANSPARENCY IN HOW THE BENCHMARK IS SET-UP.

THE OPEN PRACTICES WE HAVE FOLLOWED WERE NOT SECONDARY TO THE PROJECT: OUR BENCHMARK SIMPLY COULD NOT HAVE RUN WITHOUT ALL CODE BEING PUBLIC AND ALL PARTICIPATION BEING INCLUSIVE.