Producing Open Data

Caroline Fischer[‡], Simon David Hirsbrunner[§], Vanessa Teckentrup[|,¶]

‡ University of Twente, Enschede, Netherlands
§ University of Tübingen, Tübingen, Germany
| University of Tübingen, Tübingen Center for Mental Health, Department of Psychiatry and Psychotherapy, Tübingen, Germany
¶ Trinity College Institute of Neuroscience, Trinity College Dublin, Dublin, Ireland

Corresponding author: Caroline Fischer (c.fischer@utwente.nl)

Academic editor: Tamara Heck

Abstract

Open data offer the opportunity to economically combine data into large-scale datasets, fostering collaboration and re-use in the interest of treating researchers' resources as well as study participants with care. Whereas advantages of utilising open data might be self-evident, the production of open datasets also challenges individual researchers. This is especially true for open data that include personal data, for which higher requirements have been legislated. Mainly building on our own experience as scholars from different research traditions (life sciences, social sciences and humanities), we describe best-practice approaches for opening up research data. We reflect on common barriers and strategies to overcome them, condensed into a step-by-step guide focused on actionable advice in order to mitigate the costs and promote the benefit of open data on three levels at once: society, the disciplines and individual researchers. Our contribution may prevent researchers and research units from re-inventing the wheel when opening data and enable them to learn from our experience.

Keywords

research data, data sharing, data repositories, open science

Introduction

Modern computing facilities have given momentum to analysing large quantities of human-centred data in many fields of research. Whereas a high number of small individual research grants and projects has promoted the acquisition of many small-scale datasets, the collection of human-centred data can be challenging for study participants, indicating a need to collect data as parsimoniously as possible. Amongst other advantages, open data offer the opportunity to economically combine data into large-scale datasets, thereby fostering collaboration and re-use of data to treat researchers' resources, as well as to

study participants with care. We build on this idea of open data and explore the challenges and benefits of producing open data when studying humans and their behaviour through the lens of distinct research fields: life sciences, humanities and social sciences. We thereby draw on our own experience as open scholars, as well as the literature.

Below, we characterize practices, processes and challenges of opening and sharing scientific research material and we provide guidelines for researchers to make their data accessible for re-use. We aim at contributing to a growing, interdisciplinary body of literature that treats open science both as a research object and a practice (Bartling and Friesike 2014). We draw from experiences in our specific research areas and we hope this contribution is also beneficial for researchers in other fields and disciplines. To this end, we add a practitioner's perspective on the emerging literature on open research data. Such a perspective is needed, as opening up research data might sound easy to some and hard or impossible to others. We share neither of these perspectives, but want to show a very pragmatic perspective to illustrate that producing open data has its pitfalls, but also offers rewards for various stakeholders, including the individual researcher.

In the remainder of this article, we first discuss the characteristics of open data, what qualifies as open data and which formats are reusable. Secondly, we analyse key steps in the research process to produce open data, such as gaining participants' consent, preparing data and metadata and choosing a repository. We specifically acknowledge the effort and resources needed for these steps and highlight the availability of assistance at universities and other institutions that support the production of open data. Finally, we conclude by discussing the impact of producing open data. In doing so, we analyse long-term and short-term consequences on the societal, disciplinary and individual levels to facilitate decision-making regarding open data practices within academic institutions.

Open research data

For a long time, the handling of data has been an integral part of scientific practice. More recently, data have been characterised as the key resource (Kitchin 2014) of the 21st century. Considering the central relevance of the 'data' concept for this article, we begin with a definition of our understanding of 'data', 'research data' and 'open data'.

Data

'Datum' (sg.) and 'data' (pl.) are derived from the Latin word for 'given' or 'something given'. 'Data' typically combine three qualities: a material, symbolic and pragmatic one. From the material perspective, data can both take analogue and digital forms. The focus here is on digital data, which are pieces of information based on binary, electrical pulses. They can be sent from one location to the other as a signal or they can remain in one place, for example, when stored on a medium (Data Ethics Commission 2019: 52). On the symbolic level, data are the results of abstracting the world into categories, measures and other representational forms. These representations can be numbers, characters, symbols, images, sounds, electromagnetic waves and bits. They are understood as the raw material

from which information and knowledge are created (Kitchin 2014: 1). Data are often categorised by form (qualitative or quantitative), structure (structured, semi-structured or unstructured), source (captured, derived, exhausted, transient), producer (primary, secondary, tertiary) and/or type (indexical, attribute, metadata) (Kitchin 2014: 4ff).

Research Data

According to Leonelli (2015): 2), data can be understood as "essentially fungible objects, which are defined by their portability and their prospective usefulness as evidence". In addition to the data definitions shown above, Leonelli highlights the pragmatic nature of data – the aspects of purpose and practice. 'Data' are not an entity simply existing in the environment. Entities only become data when they are treated as such. They are embedded into a certain set of practices. What counts as data and what form data assume depends on the concrete context of application. One of these contexts is scientific research, where data are shaped by comprehensive disciplinary and cross-disciplinary conventions and traditions.

For example, in neuroscience and many other life sciences, datasets are often highly modular. Methods that are used to measure brain function produce data as heterogeneous as lists of voltage values (in the case of an electroencephalogram) on the one hand and greyscale values in three-dimensional image spaces (in the case of functional magnetic resonance imaging) on the other hand. If they are to be re-used, these data points need metadata containing references to underlying technology and settings to structure the data streams in a coherent way.

In many social science fields, such as political science, public administration, organisational studies or management, data are often structured along multiple levels, referring for example to countries, organisations within these countries and individuals within countries and organisations. Such a multi-level structure is only useful when links between levels can be clearly identified (i.e. which citizen belongs to which country).

Moreover, qualitative research tends to work with highly unstructured data, which include artifacts, such as field notes, memos, official documents, images, movie clips, diagrams and tabular data. These data typically include identifiable information about concrete individuals, places or events.

To summarise, we pragmatically understand research data as any data produced in scientific processes and/or used in scientific processes. That might also include big data originally produced by social media companies and repurposed for science or data produced in private sector marketing research. Such datasets might relate to different restrictions concerning data sharing.

Open Research Data

What does 'openness' of data mean in a research context? This question has been answered in many different ways depending on one's understanding of openness in science (Fecher and Friesike 2013). The overall aim of opening research data is to make

them accessible for secondary use (reusage, Steinhardt et al. 2021). This typically involves releasing datasets into a (digital) repository, structuring the data in a common, standardised way, saving data in a portable file format, adding documentation. Crucially, users are not charged when obtaining and re-analysing the datasets. In some situations, data may simply be placed online for download via a URL with no restrictions whatsoever. This contrasts with more restricted conditions of access where only a record referring to the data is published, but access to the dataset is not granted (closed), a record of the data is published and requirements are mentioned that need to be fulfilled to apply for access to the data (mediated) or a certain time point has to be reached at which the data are released automatically (embargoed). Thus, whether data are open or not is not a simple matter of yes or no.

By making data available in any form, stakeholders contribute to making research more transparent and, in the best case, reproducible. However, not all open datasets are equally useful for secondary use. While making data available in any form is conditional to making especially published work more reproducible, not all open datasets are equally useful for secondary use (Markiewicz et al. 2021). The FAIR principles (Wilkinson et al. 2016) list characteristics for maximising the impact of open data, stating that data need to be F indable (i.e. comprising a globally unique and persistent identifier and registered in a searchable resource), A ccessible (i.e. retrievable using a standard, open and free communications protocol), I nteroperable (i.e. using a standard, formal, accessible format for knowledge representation) and R eusable (i.e. enriched with accurate and relevant (meta-)attributes, a valid and accessible licence and formatted according to disciplinary standards). These FAIR principles can serve as high-level guidelines to provide information for implementation choices, independent from the domain and focused on a broad range of scholarly outputs.

Nevertheless, the intention of 'opening' is to make data accessible for the respective research community in a more controlled manner, taking into consideration specific methodological and content-related criteria. This means data access might be restricted for scientific use only or researchers need to identify their research interest or their intended methodology before gaining access. This is foremost a pragmatic decision to limit the focus of this article, because sharing data with amateur scientists and other stakeholders often implies a different quality of accessibility, for example, the need to summarise data, plot data or explain data other than just publishing a raw dataset. Such regulated access can be organised by a research data centre or an institutional repository, for example, by offering on-site access to data, via a remote desktop or sharing of just parts of a dataset. Although these restrictions might impede the notion of openness, especially when working with sensitive data, such as personal data or data from vulnerable groups, this is often the only way to open data up at least partially (Steinhardt et al. 2021). However, in our understanding, the mere claim that datasets are shared by authors upon request cannot count as open data, as authors might decide randomly who will get access and whether requests are reasonable (Tedersoo et al. 2021, Houtkoop et al. 2018). In our understanding, instead of the proactive delivery of open datasets, these claims rather

represent a form of "open-washing". Therefore, stakeholders are beginning to ask for a proactive sharing of data (Morey et al. 2016).

Steps in producing open data

The production of open data does not start with sharing a dataset, but usually with the beginning of each research cycle. In every step of a research project, opening datasets then requires additional considerations. Measures need to be taken and barriers overcome. The following chapter discusses these elements for producing open data along the ideal type of a linear research process. In reality, however, some of the described activities and considerations may overlap or occur in a different order.

Planning: data management plans and pre-registrations

We suggest to consider the possibility of opening up research data already in the planning phase of a project. Although data might be opened up ex-post, such a procedure is often problematic as participants' consent to publication of their data might miss the necessary legal provisions or formatting the datasets for interoperability may be time-consuming.

Researchers should be aware that handling research data and collection and use of data are nowadays highly regulated by national and international legislation. Whenever research data involve information about individual human beings, production and use of the data are currently, for example, regulated by the General Data Protection Regulation for the European Union. The framework stipulates provisions, such as the residence principle, the right to data transfer, the obligation to protect data through system design, a right of complaint and the sanctioning of violations (Roßnagel and Geminn 2020), which are relevant in the context of research data production and (re-)use.

As individual researchers cannot always easily ensure compliance with current or future laws, many funding agencies and ethics boards of universities have taken on this responsibility. Grantees are asked to develop detailed data management plans (DMP) in the application or planning phase. The requirements for these DMPs increasingly address and are favourable to the issue of open data. The German Research Foundation (Deutsche Forschungsgemeinschaft DFG), for example, states that grantees should consider whether and what research data could be relevant for other research contexts and how these data can be made available for reuse. Research data should be “made available as soon as possible” (Deutsche Forschungsgemeinschaft 2015), assuming that the publication of research data does not conflict with data protection or copyright issues. Research data should not only be made available in the short term, but be archived in the researcher's own institution or an appropriate national infrastructure for at least ten years. At the level of the European Union (EU), the funding scheme (Horizon Europe) requires researchers to make data “as openly accessible as possible and as closed as necessary” (European Research Council 2017).

As of 2021, funding requirements are even more formalised regarding an open data policy ‘by default’ and it is not possible to opt out of these obligations completely. It is principally possible to opt out of the requirement to provide open access to data and metadata as long as an explanation is given. The implementation of these funding policies is increasingly supported by infrastructure initiatives and networks. Examples, in this context, are the German national research data infrastructure and research data centres and the plans for a European Open Science Cloud.

Apart from funding bodies, ethics boards of research institutions may also require researchers to produce DMPs. This is mostly the case when research activities involve individuals. When phenomena are studied on the meso or macro level (e.g. studying countries, regions or economic sectors instead of individuals), however, there are still many empirical research projects that plan their data management in an emergent process running in parallel with the project, especially when no ethics approval has to be obtained in advance.

A DMP usually describes which kind and amount of data will be collected in which way and how data will then be stored. Several templates (e.g. for Horizon Europe, from the DFG or any other national funder) and even apps (e.g. DMPTool) exist, making it easy to compile a DMP. Step-by-step guides to preparing a data management plan have been published by many research support units at universities and also by the Science Framework (OSF). DMPs prior to data collection help to pre-plan the different steps of data collection, analysis and storage, to fulfil the requirements for data storage (e.g. a specific form of participants’ informed consent requested by a repository) can be implemented. Many organisational or disciplinary data repositories and data centres are ready to give advice on this planning process. Alternatively, organisational research support teams can be contacted to find out which regulations and aspects need to be taken into account.

Even if not requested, researchers - in our opinion - should always use a DMP to ensure that important opportunities are not missed, for example, with regard to opening up data. DMPs are also helpful in establishing important steps in a research process within a team and to avoid conflict later in case team members have different interests, for example, with regard to opening up the dataset.

Despite all this help being available, a DMP will not be written in a day - especially if researchers lack experience. Depending on the level of detail that is requested by a funder, an ethics committee or other stakeholders, the mere writing of a DMP might take several days - not mentioning planning the data collection and storage as such. Hence, it is important to start such an activity well ahead of any deadlines and planned data collection start dates. If a pre-registration (i.e. submitting your full research plan to a registry to separate confirmatory from exploratory analyses) or registered report (i.e. submitting your full research plan to a journal for peer review and possible in-principle acceptance of the resulting manuscript) is an option for the project, a DMP is mandatory and a good starting point for reflecting on the expected structure of the data and intended outcomes. Step-by-step guides to prepare a data management plan are published by many research support units in universities and also by the OSF.

Collecting: informed consent, opt-in and data collection

Individuals participating in empirical studies need to give their consent to data collection procedures and they need to be informed about the context of the research. How consent needs to be acquired and the scope of information that needs to be given is regulated by the GDPR in the EU context *[1]). According to an EU guideline, an Informed Consent Form (ICF) should at least include the following information:

1. the identity of the data controller (including contact details),
2. the specific purpose(s) of data processing,
3. the subject's rights as guaranteed by the GDPR and the EU Charter of Fundamental Rights (including the subjects' right to withdraw consent and to access their data, respective procedures and the right to lodge a complaint with a supervisory authority),
4. information as to whether data will be shared with or transferred to third parties and for what purposes and
5. how long the data will be stored before they are destroyed (European Commission 2021).

These obligations also have consequences on the open data production process. The ICF must include information on the intention to make the data, or parts of it, accessible as an open dataset. The ICF should further include detailed information about the location the datasets will be uploaded to (e.g. the open data repository). Crucially, current legislation also specifies a requirement to use an explicit opt-in procedure (i.e. participants actively select and agree to a certain procedure regarding data collection) instead of an opt-out (i.e. participants are automatically enrolled in the procedure and need to explicitly state their wish to not take part).

By informing participants about the study and asking them for their consent, we learned that it is useful to separate the opt-in for the general participation from the opt-in for opening up (parts of) the dataset later on. This way, open data constitute no threat to sample size and data from participants who are concerned about opening up data can still be used in analyses. However, such a procedure might result in different datasets (a complete (closed) and a reduced (open) dataset), rendering attempts at reproducing the reported results futile.

Obtaining informed consent from individuals is not always easy as illustrated by members of vulnerable groups who might not be able to understand the given information about a study. For example, individuals suffering from severe cognitive decline, under-age participants or individuals with a limited capacity to communicate, such as coma patients, might not be able to opt-in in a study in an informed way. Solutions are highly dependent on a particular individual, sometimes informed consent being given by parents or other guardians. In other cases, simple language information and a respective consent form might be sufficient.

Another issue to reflect upon is the fact that making participants aware of open data might bias their behaviour within a study. They might be hesitant to talk about sensitive issues fearing that their data might later be de-anonymised. Additionally, insufficient information on which research questions will be pursued, based on their data, may make them cautious about participating. This may especially apply if participants are not motivated to participate in research in general, but have an interest in a certain research topic and are only willing to participate if use of their data is restricted to advancing this topic. On the other hand, opening up their data might also motivate individuals to participate as their efforts can have a higher impact when re-used for several research projects.

While planning data collection, some thought should be given on which data are and are not necessary to collect. Collected data should be as rich as necessary and as sparse as possible to make the best use of the time participants invest. Additionally, researchers should be aware of meta-data that can be automatically collected via certain devices. For example, smartphones that are increasingly used for research on human behaviour have passive sensors (such as geolocation via GPS). These can provide interesting and valuable insight for research projects focusing on movement patterns, for example, but can also pose a risk of revealing sensitive information (such as participants visiting specialist medical healthcare centres). If these data are central to the research conducted, researchers should have planned pre-processing strategies (see next section) to remove any potentially identifying information from the data that are to be opened before release.

Cleaning and preparing: contextualising data and making it ready to publish

Depending on the clarity of the DMP, extensive cleaning should not be necessary (i.e. removing data-points that do not reflect the data acquisition pipeline as described in the DMP due to, for example, hardware failure or human error) of the data. Still, technical errors or disruptions in data acquisition, such as experienced by many researchers during the initial wave of the Covid-19 pandemic, when researchers were not able to invite participants into the lab, can lead to alterations in individual data files (e.g. due to changes in software versions installed on shared hardware used for measurements) that are time-consuming to repair for an individual researcher working on opening their dataset. Importantly, whenever possible, the dataset should only be cleaned to a degree that erroneous data points are removed, but the data has not been fully transformed from raw data into a summary format (for example calculating means instead of reporting the initially collected values). This way, researchers can perform a broad range of data analyses on the open dataset which maximises the utility of the time invested in the first place. Based on the type of data, metadata need to be added which sufficiently describes the data to enable follow-up analyses. In the case of neuroscientific data, the BIDS standard (Brain Imaging Data Structure; Gorgolewski et al. 2016) provides a template data structure that automatically includes the necessary metadata for common analysis approaches.

Eventually, the choice of the file format is important to render the dataset accessible to a broad audience and for a maximum duration. For some data types and depending on the

field, certain file formats can be the quasi-standard (for example (compressed) NIfTI for neuroimaging data) that can be processed by all standard software packages. If the data do not require a certain file format or if no clear disciplinary standards are available, the simplest representation can be chosen that preserves any structure inherent to the data and proprietary formats can be avoided which make it harder for researchers lacking the necessary software licences to open and process a file (e.g. data formats that can only be opened with certain statistics software). Adding an easily accessible, humanly as well as machine-readable description to the data, makes it easy for researchers browsing a repository to decide if the dataset fits their research question.

However, preparing data for publication might also entail a translation of text data (stemming from documents, interviews or surveys, for example) to make it usable for a broader international audience. Translations often entail induced biases, especially when not done by a native speaker and professional translators are expensive and not always affordable. Ideally, these costs related to opening up a dataset are factored in early in a project.

The standardisation of quantitative data might also induce a loss of information. If data are released in a summarised form to save space or to preserve anonymity - for example, means are calculated over repeated measurements of a variable - a subset of analyses depending on the original raw measurements cannot be run anymore, thereby impeding attempts to reproduce results that have been previously reported. However, modern data transfer rates and online server space are steadily increasing, so releasing the full raw data should be the default. For neuroimaging data, a growing number of online databases that invite the submission of raw data, such as OpenNeuro, allow sharing of big datasets without having your own server readily available.

A different problem of making datasets available for reproduction on various systems pertains to either different statistical analysis packages or different versions of these packages leading to different results (Bowring et al. 2019). To mitigate this issue, a detailed report of the methods used to produce results is vital, for example, the statistical software used, its versions, used packages. Moreover, researchers can opt to, in addition to the open dataset, release a version of their own system environment used for the reported analysis via docker (a container solution packaging a full operating system, including all preferences and necessary software).

Opening up: data repositories and data publications

Although researchers should choose a place for publication and long-term storage of their data early in the project, the data storing and sharing, as such, is the last step in the endeavour of producing open data. In general, selecting a repository depends on a lot of different considerations: regulations by the employing organisation, regulations by funders, disciplinary conventions, normative considerations, for example, location of server. Another aspect to consider is whether the repository is owned by a non-profit organisation and restrictions pertaining to the participants. If restrictions are needed - for example, to a 'scientific use only', so-called research data centres might be the best option. These data

centres offer researchers an analysis of restricted data on-site at a workplace in the data centre. They compile scientific use files that differ from public use files or execute the code from external researchers and share results instead of raw data (see, for an example from the social sciences, the Secure Data Center of the German Gesis). If a certain discipline is to be addressed and no restrictions are needed, the best-known repository in a discipline is always a good choice. Usually disciplinary associations or networks can give some advice on preferred repositories. Other multi-purpose archives are, for example, the Open Science Framework, Dataverse and Zenodo. Many universities and research organisations also have developed their own archives or they are collaborating with other institutions to build a common infrastructure. Usually these institutional repositories are intended to ensure long-term storage of data that is nowadays often requested by funders. Advice on general data repositories and the organisational infrastructure can usually be gained from a research organisation's library or research support staff. They will also advise on a suitable data licence (e.g. CC 0 or CC BY) to ensure the least restriction possible as well as data protection, if needed.

A responsible attitude towards data, long-term goals of researchers and lab funding will, sometimes, even lead to the decision that an embargo period is necessary – or producing and maintaining open datasets in a specific case creates a disproportionate burden for the researchers compared to the low possible value and prospect of data re-use. Embargoes can be issued if researchers acquiring data want to or need to make sure that a certain set of analyses can be finalised on the respective dataset before others can publish results, based on this dataset. This can be realised by:

1. keeping the dataset closed for a certain time period, either depending on a pre-defined time window or depending on pre-defined conditions which need to be met in order to release the dataset or
2. releasing the dataset subject to an agreement that new findings on the data can only be published after the pre-defined time window has elapsed or the pre-defined conditions are met.

Issuing embargoes can be sensible if a certain set of analyses is part of a grant funding acquisition of the data or students working towards a degree are dependent on publishing their results first. Towards this end, embargo periods can incentivise opening up datasets in areas where research funding is spent to a great degree on data acquisition and new funding directly depends on high-tier publications, often leading to fully-closed datasets when embargoes are not considered.

With the advent of data publications (Smith 2009) and data centred journals which publish descriptions for datasets are published as articles, researchers can receive proper credit for their efforts in opening up a dataset: mere data publications thus gain weight. In contrast to a typical research article that references an open dataset, these data publications are stand-alone articles themselves which allow researchers to describe their data, data generation and format in more detail. Given the potentially widespread user base, these publications also have a fair chance of attracting a high number of citations.

Additionally, researchers can refer to these publications in later work that builds on the data, thus simplifying the process of describing the dataset in each manuscript.

Impact of producing open data

The production of open data is not an end in itself and especially when we revisit the resources that need to be invested to open up datasets, one might ask whether it is worth the effort. However, there are multiple rationales to open and share data in the research context (Duke and Porter 2013). In the following, we discuss some positive effects of open data production at the levels of the society, the scientific community and the individual researcher in more detail.

Societal level

On the societal level, we might, for example, focus on participants of studies. Data collection always puts a burden on participants, they need to spend their time answering questions or attending lab experiments. In the social sciences, the so-called "oversurveying" of society is highlighted (Weiner and Dalessio 2006). Especially when studies need to obtain representative samples of the population of a country or a region (for example, when validating a psychometric instrument like a questionnaire), it is often resource-consuming to obtain such a sample of a sufficient size . This can become particularly burdensome for participants who are part of a very small population when research questions are investigated that focus on rare disorders, for example. It is for good reason that research in the social and life sciences increasingly incentivises participants and builds on paid online access panels that were initially developed for private sector market studies. Yet, the quality of data collected from these professional study participants is sometimes questionable (e.g. amazon turk). Additionally, participants are incentivised to avoid panel dropouts (participants leaving a panel after only a few survey waves used, for example, for a longitudinal study). However, re-using an existing dataset for a different research question might be problematic in that a sample “overfits” a particular research question. That can lead to biases when analysing the data on a different topic. Similarly, it is also questionable in terms of robustness of findings when a whole body of literature relies on just one shared dataset.

Open data are in the long run less burdensome for society because scholars studying similar questions might build on the same dataset instead of collecting data on their own. In the social sciences, some organisations have long been producers of open datasets which can be exploited for a lot of different research questions. Socioeconomic panel studies (such as the German Socio-Economic Panel), public opinion surveys (such as the Eurobarometer) or data that are available from the national statistical offices (such as the British Office for National Statistics) are efficient ways of collecting and using representative country-level data. In psychology and neuroscience, open data are slowly gaining traction after concerns were raised regarding the replicability of several seminal findings (the so-called replication crisis). By now, extensive data-sharing initiatives, such as the Human Connectome Project, have been founded that often foster the usage of the

available datasets by hosting competitions on a broad range of research questions (for example ADHD-200).

Apart from the burden on study participants, data collection is expensive and, at the same time, research organisations are often lacking resources. A more efficient use of resources - most commonly tax money - can be achieved by teaming up for data collection and sharing data. To this end, open data support the human right to information: everyone may enjoy the benefits of scientific progress and its applications (Article 15 of the International Covenant on Economic, Social and Cultural Rights). Sharing data, not only in small circles, but opening it up widely, can also help to re-distribute resources to countries in which research might not be well funded and researchers might not be able to collect specific forms of data on their own. Especially when data collection is bound to expensive technology, such as in neuroscience, worldwide use of datasets can also help to tear down knowledge barriers and distribute knowledge more equitably. In a similar way, open data may then also be valuable as an instrument increasing public trust in the results of scientific research in the face of controversy.

Disciplinary level

The most direct impact of open data, however, might be observed in the development of the discipline. Thus, open data affects both - research as well as education. In terms of education, open datasets enable teaching, based on real data and practising data analysis, based on real datasets. Realistic problem-based learning is, thus, possible and students are better prepared than working with artificial teaching datasets. Bishop and Kuula-Luumi (2017), for example, show that open data are to a high share downloaded by students on different levels and mostly used for learning and teaching. Students might - in that - also develop new research findings from existing data and help to grow the disciplinary knowledge base, for example, in their theses or research project courses. This can enhance their motivation, based on the notion that they can make a difference. Similarly, completion times for the qualification of specifically undergraduate students can be too short to enable the acquisition of a dataset that is large enough to not be statistically underpowered. Here, open data open new possibilities to investigate a novel research question where students are commonly restricted to literature reviews or small samples.

In terms of research, open data enable innovation and make findings more robust. Whereas innovation is driven by combining datasets and answering new comparative questions, robustness of a finding can only be assessed by re-running the same analysis on different datasets. On the other hand, replications become much easier when data are opened up. The more data-intensive the research, the more dependent replicability becomes on the availability and accessibility of the original datasets for re-analysis. Specifically when using behavioural data, based on task paradigms, slight variations in the implementation of the paradigms are the norm between research groups. If these data are openly available, researchers can investigate the robustness of results across task variations, ultimately providing information for task choice for future studies.

Individual level

Open data - as discussed above - come with several costs for individual researchers or research teams. We argue that, in general, a high intrinsic motivation to publish open datasets is needed. However, we also observe a growing impact of data publications. They count as separate publications instead of just an appendix to an article and might attract additional citations. Still, the value of data publications differs between disciplines and they might be more attractive in the life sciences than in the social sciences. Apart from these metrics, which still count in the recruitment and promotion of academics, open datasets might also generate appreciation and help to build or strengthen an individual network. That again might lead to further collaboration and help in terms of career. Publishing open data might add to a positive reputation of a scholar, either because the discipline relies on shared datasets or because published results can be verified. Being perceived as an expert for open data publication might ultimately also lead to new career opportunities in research support and management.

Conclusion

Due to current incentive structures, producing open data offers a lot of benefit for societies and research fields as a whole, whereas individual researchers doing the work can be at a disadvantage compared to their peers working on closed data sources. Planning for, acquiring and preparing datasets for public release is a serious effort that researchers need to commit to. In comparison, producing closed datasets is less time-consuming as datasets do not need to be made easily accessible and enriched with meta-data. With hiring decisions depending on the number and visibility of publications and funding agencies incentivising the acquisition of new data instead of encouraging the use of existing, extensive open datasets, researchers investing time in optimising their datasets for public release can, thus, lose ground.

Still, apart from the obvious benefits for scientific and societal progress, open research practice is increasingly considered to be a hallmark of "science done right". To this end, individual researchers may benefit from a boosted visibility of their work. Greater transparency supports replicability, thus corroborating trust which is indispensable in scientific discourse. New developments, such as citable data publications, stronger datafication in all scientific disciplines and policy-making towards transparency and replicability by disciplinary associations, national laws and funders also act as incentives for opening up scientific work. Hence, researchers who wish to incorporate open science into their repertoire will find good conditions to do so with ample help available online and within research organisations, as well as academic networks that facilitate building routines for future projects.

Our step-by-step guide summarises the experiences of practitioners across different disciplines and strives to serve as a resource for researchers seeking orientation on how to open up their data. Given the clear benefits for society, disciplines, but also the individual researcher, with current developments in funding, policy-making and hiring decisions increasingly favouring open science approaches, opening up data should be

viewed as a default – where ethically unobjectionable – with clear guidance for scholars entering a field.

Acknowledgements

We want to thank reviewers and editors for their valuable input. We also want to thank Gwen Schulte (DIPF) for proofreading this article.

The publication of this article was kindly supported by RIO. We would like to thank RIO and Wikimedia Deutschland for enabling this collection.

Funding program

The authors were funded by the Open Science Fellows Programme by Wikimedia Deutschland, VolkswagenStiftung and Stifterverband in 2017/2018 (Caroline Fischer and Vanessa Teckentrup) and 2018/2019 (Simon Hirsbrunner).

Author contributions

The authors are presented in an alphabetical order, all authors contributing equally to the manuscript.

Conflicts of interest

References

- Bartling S, Friesike S (2014) Opening Science. Springer [ISBN 978-3-319-00025-1] https://doi.org/10.1007/978-3-319-00026-8
- Bishop L, Kuula-Luumi A (2017) Revisiting Qualitative Data Reuse. SAGE Open 7 (1). https://doi.org/10.1177/2158244016685136
- Bowring A, Maumet C, Nichols T (2019) Exploring the impact of analysis software on task fMRI results. Human Brain Mapping 40 (11): 3362-3384. https://doi.org/10.1002/hbm.24603
- Data Ethics Commission (2019) https://www.bmi.bund.de/SharedDocs/downloads/DE/publikationen/themen/it-digitalpolitik/gutachten-datenethikkommission.pdf?__blob=publicationFile&v=6.. Accessed on: 2022-4-03.
- Deutsche Forschungsgemeinschaft (2015) Guidelines on the Handling of Research Data. URL: https://www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/forschungsdaten/guidelines_research_data.pdf
- Duke C, Porter J (2013) The Ethics of Data Sharing and Reuse in Biology. BioScience 63 (6): 483-489. https://doi.org/10.1525/bio.2013.63.6.10

- European Commission (2021) Ethics and data protection. Non-official guideline drafted by expert panel at the request of the European Commission (DG Research and Innovation). URL: https://ec.europa.eu/info/sites/default/files/5._h2020_ethics_and_data_protection_0.pdf
- European Research Council (2017) Guidelines on Implementation of Open Access to Scientific Publications and Research Data in projects supported by the European Research Council under Horizon 2020. URL: https://ec.europa.eu/research/participants/data/ref/h2020/other/hi/oa-pilot/h2020-hi-erc-oa-guide_en.pdf
- Fecher B, Friesike S (2013) Open Science: One Term, Five Schools of Thought. Opening Science17-47. https://doi.org/10.1007/978-3-319-00026-8_2
- Gorgolewski K, Auer T, Calhoun V, Craddock RC, Das S, Duff E, Flandin G, Ghosh S, Glatard T, Halchenko Y, Handwerker D, Hanke M, Keator D, Li X, Michael Z, Maumet C, Nichols BN, Nichols T, Pellman J, Poline J, Rokem A, Schaefer G, Sochat V, Triplett W, Turner J, Varoquaux G, Poldrack R (2016) The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. Scientific Data 3 (1). https://doi.org/10.1038/sdata.2016.44
- Houtkoop BL, Chambers C, Macleod M, Bishop DM, Nichols T, Wagenmakers E (2018) Data Sharing in Psychology: A Survey on Barriers and Preconditions. Advances in Methods and Practices in Psychological Science 1 (1): 70-85. https://doi.org/10.1177/2515245917751886
- Kitchin R (2014) The data revolution: Big data, open data, data infrastructures and their consequences. Sage
- Leonelli S (2015) What Counts as Scientific Data? A Relational Framework. Philosophy of Science 82 (5): 810-821. https://doi.org/10.1086/684083
- Markiewicz CJ, Gorgolewski KJ, Feingold F, Blair R, Halchenko YO, Miller E, Hardcastle N, Wexler J, Esteban O, Goncavles M, Jwa A, Poldrack R (2021) The OpenNeuro resource for sharing of neuroscience data. eLife 10 https://doi.org/10.7554/elife.71774
- Morey R, Chambers C, Etchells P, Harris C, Hoekstra R, Lakens D, Lewandowsky S, Morey CC, Newman D, Schönbrodt F, Vanpaemel W, Wagenmakers E, Zwaan R (2016) The Peer Reviewers' Openness Initiative: incentivizing open research practices through peer review. Royal Society Open Science 3 (1). https://doi.org/10.1098/rsos.150547
- Roßnagel A, Geminn C (2020) Datenschutz-Grundverordnung verbessern. Der Elektronische Rechtsverkehr https://doi.org/10.5771/9783748920991
- Smith VS (2009) Data publication: towards a database of everything. BMC Research Notes 2 (1). https://doi.org/10.1186/1756-0500-2-113
- Steinhardt I, Fischer C, Heimstädt M, Hirsbrunner SD, Ikiz-Akıncı D, Kressin L, Kretzer S, Möllenkamp A, Porzelt M, Rahal R, Schimmler S, Wilke R, Wünsche H (2021) Opening up and Sharing Data from Qualitative Research: A Primer. WI - Weizenbaum Institute for the Networked Society https://doi.org/10.34669/wi.ws/17
- Tedersoo L, Küngas R, Oras E, Köster K, Eenmaa H, Leijen Ä, Pedaste M, Raju M, Astapova A, Lukner H, Kogermann K, Sepp T (2021) Data sharing practices and data availability upon request differ across scientific disciplines. Scientific Data 8 (1). https://doi.org/10.1038/s41597-021-00981-0
- Weiner SP, Dalessio AT (2006) Oversurveying: Causes, Consequences, and Cures. In: In Kraut AI, et al. (Ed.) Getting Action from Organizational Surveys: New Concepts, Technologies, and Applications. Jossey-Bass, San Francisco. [ISBN 978-0-787-97937-9].

- Wilkinson M, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J, da Silva Santos LB, Bourne P, Bouwman J, Brookes A, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo C, Finkers R, Gonzalez-Beltran A, Gray AG, Groth P, Goble C, Grethe J, Heringa J, 't Hoen PC, Hooft R, Kuhn T, Kok R, Kok J, Lusher S, Martone M, Mons A, Packer A, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone S, Schultes E, Sengstag T, Slater T, Strawn G, Swertz M, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B (2016) The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3 (1). https://doi.org/10.1038/sdata.2016.18

Endnotes

*1 GDPR Article 3(2) applies in two cases when a data processor processes personal data of individuals who are present in the EU, namely:

1. when s/he either offers goods or services to the data subject (market place) or
2. when the data processing serves to observe his/her behaviour (observation place). With this provision, the applicability of European data protection law is no longer linked to the establishment of the controller, but also depends on the location of the data subject in the EU.