Funder Data-sharing Policies: Overview and Recommendations

Stephanie Wykstra

September 2017

Executive Summary

Introduction

Research transparency – sharing data and other materials for further research and re-analysis – is on the rise. Many people and institutions in the research ecosystem have pointed to potential benefits of transparency, from accelerating scientific progress to increasing reliability of research. Over the past two decades, dozens of funders and hundreds of journals have adopted data-sharing policies, data repositories have sprung up to meet the growing demand, and norms among researchers have also started to change.

In spite of the increasing support for transparency, many questions remain. On the funder side, these questions often pertain to data-sharing polices: what data should funders ask grantees to share? Where and when should they share? And what support will funders provide for data-sharing? This report, commissioned by Robert Wood Johnson Foundation, reviews current funder policies and practices and provides recommendations.

Methods

Research for the report consisted of several different activities: (1) A scan of open data policies of 27 research funders and 17 journals; (2) A scan of 7 open data repositories; (3) Key informant interviews with 18 open data stakeholders; (4) An assessment of the Health and Medical Care Archive (*for RWJF's internal use*).

Recommendations

Recommendation #1: Adopt a data-sharing policy to promote transparency in research covering grants which produce research data. At a minimum standard, the policy can ask grantees to publicly share the de-identified data, metadata and code underlying their articles in a data repository, at the time of publication.[1] Adopting this requirement will allow the foundation to bolster the requirement that some journals already adopted, to join dozens of other funders, and in doing so, to make it clear that the underlying data, not just summary results, are an important part of the foundation's investment in research.

Recommendation #2: Ask grantees to share the larger collected dataset on a case by case basis. Preparing a large dataset and all associated materials for public use is a significant investment

[1] Preprints are not discussed here, but as preprints gain momentum in various disciplines, it will be worth considering whether data-sharing should be a requirement not just for "published" work but for work that is shared with the public in any form (preprint, research report, etc).

of time for researchers, and is a bigger request than reinforcing the norm of transparent results (as in recommendation #1). Therefore, it makes sense to consider which projects are producing data which is likely of wide value for re-use, and to request that those datasets be shared.

Recommendation #3: Ask that grantees submit a plan for how they will prepare and share data, along with their grant proposal. The plan should address questions such as: what data will be produced in the course of research? Where will it be shared, and who will be in charge of preparing the data to share and ensuring confidentiality of subjects? Which metadata will be released (i.e. information about the variables and also key information about the study)? What are the expected costs of preparing and sharing data?

Recommendation #4: Commit to financially supporting data preparation activities. For data underlying published research, this would mean supporting research staff or researcher time. For larger collected datasets, this could mean a partnership with an external data curator, or it could mean a substantive financial commitment to the research team.

Recommendation #5: Promote ways to reward grantees professionally for sharing data: ask grantees to ensure that their data has a digital object identifier (DOI) assigned, and that the data is cited in the article, so as to facilitate data citation. Consider other potential rewards such as asking that grantees note their shared data in funding proposals, alongside publications.

Recommendation #6: Ensure that grantees and foundation staff are aware of and have guidance in understanding and implementing the policy. The policy should be shared accessibly on the foundation website, as well as in requests for proposals and other documentation about grant requirements.

Recommendation #7: Consider mentioning that compliance with the policy may be a factor in future funding decisions.

Table of Contents

I. Introduction

Over the past two decades, there has been a growing movement towards research transparency. Funders and journals are increasingly adopting data-sharing policies, and organizations to support data-sharing are on the rise as well, including data repositories and curation services.

While internet-based infrastructure is recent, the idea of data as a public good goes back centuries. In the 17th century, Isaac Newton pushed the astronomer John Flamsteed to share his observations of the stars. As the work had been funded by the Royal Treasury, Newton insisted that the data was a public resource that should be shared for the purpose of accelerating scientific discovery.[2]

Beyond maximizing the re-use value of data, another rationale for sharing is to enable other researchers to check summary results using the underlying materials. This purpose for data-sharing isn't new either.[3] But over the past decade, worries about the reproducibility of research have made the latter motivation particularly salient.[4] Data-sharing isn't a guarantee of reproducibility (in any sense of the term), but many different members of the research ecosystem - researchers, funders and journal editors – argue that it is an important way to increase the reliability and credibility of research.

While the purposes of data-sharing are fairly clear, the details can be complex. Research funders and journals that adopt data-sharing policies face a number of decisions about the timeframe for sharing data, how much data should be shared, where to ask researchers to share, and whether to invest in curation services. Funder data-sharing policies vary widely in how they address all of these questions.

As the transparency ecosystem grows and matures, it is well worth stepping back to consider next steps for improving transparency. The report – commissioned by Robert Wood Johnson Foundation (RWJF) – draws on interviews with stakeholders and key experts to provide recommendations on its data-sharing policy and practices. It will also serve as a shared resource for other funders to draw on as they develop their own approaches to research transparency.

[2] Lisa Jardine, Ingenious Pursuits: Building the Scientific Revolution, 17-19.
[3] At the beginning of the 20th century, Francis Galton wrote "I have begun to think that no one ought to publish biometric results, without lodging a well arranged and well bound manuscript copy of all his data...to those who desire to verify his work" (Galton 1902).
[4] See for example Ioannidis 2005.

Background and central questions

RWJF has been involved in efforts to make scientific research more open and accessible to the public, including convening a meeting on open access, co-hosted with the Scholarly Publishing and Academic Resources Coalition (SPARC) in 2015. The initial meeting, with over 50 foundations and other groups represented, developed into the Open Research Funders Group (ORFG).

For the past several decades, the foundation has partnered with Inter-university Consortium for Political and Social Research (ICPSR) -- a data archive based at the University of Michigan -- to share selected grantee data within the foundation's official data archive, the Health and Medical Archive (HMCA). The data at HMCA is freely available to the public, and RWJF supports ICPSR's work in curating and preparing grantee datasets. RWJF has an internal process for requesting that certain grantees work with ICPSR to share data[5], and prioritizes sharing data that is likely to be of high value for re-use by other researchers.

RWJF is interested in potentially re-evaluating and standardizing its policy,[6] and as the Foundation plans its next steps, has raised a few central questions:

- Should the scope of a policy include all data-generating research projects, or only research which yields data likely to be of high value for secondary analysis?
- What requirements have other funders adopted in their data-sharing policies (e.g. what, where and when should grantees share data)?
- How should RWJF assess the value of data curation, particularly when it comes to a wider scope of projects sharing their data?
- Other recommendations: what advice do other funders have in adopting a policy, based on their experiences working with their own staff and grantees?

Methodology

Research for the report consisted of several different activities: (1) A scan of open data policies of research funders and journals; (2) A scan of open data repositories; (3) Key informant interviews with open data stakeholders; (4) An assessment of the Health and Medical Care Archive (this part of the report is for RWJF's internal use only).

[5] One grantee has also shared data with Roper Center.

[6] As part of considering this step, RWJF solicited views from programs officers and grantees in a 2016 consultation on open access and open research.

The list of 27 funder policies (see Appendix II) grew out of online research as well as conversations with funders. Several federal funders, including NIH and NSF, have organization-wide policies that apply to all of their institutes/directorates, and these organization-wide policies were considered,[7] along with other international governmental policies, and all the private foundation policies located.

The journal scan included 17 journals in which RWJF grantees often publish, and included an assessment of whether they had a data-sharing policy and its major components if so (see Appendix III). Further research on journals draws on a published review of 318 biomedical journal data-sharing policies (Vasilevsky et al 2017).

The research on open data repositories included an assessment of 7 repositories, of which 5 are commonly used general repositories (see Appendix III).

18 interviews took place between February and April 2017, and included key stakeholders and experts at 14 organizations (see Appendix I). The majority of the interviewees were contacts at foundations, but also included representatives of other organizations (e.g. Center for Open Science and several repositories).

II. Funder data-sharing policies

History

Throughout the past couple of decades, there's been a growing consensus among research funders around a few points. First, that the internet has made it much easier to share underlying research data, and that there are clear-cut cases in which this ease of sharing data has accelerated scientific progress. The Human Genome Project and other genomics research is often cited as an example (and in many ways, genomics researchers have led the way in showing the value of data-sharing and laying out principles for sharing).[8] For funders, data-sharing is one way to pursue a greater return on the investment in research. Data shared by researchers may be used by others for further analysis and meta-analysis, drawing out a greater

[7] We did not compile policies of the specific institutes/directorates here, as this was beyond the scope of this project (and would amount to several dozen further policies). Some federal funders had, as of May 2017, issued drafts (called "plans") of their data-sharing policies. We did not include these plans in our review either, since they are not yet finalized policies.

[8] See Fort Lauderdale Agreement, for example (Wellcome Trust 2003).

value from the original data collection. As one funder noted in conversation, "I think [data-sharing] is a terrific practice for us to get additional value from our philanthropic investments."

Funders also often point to transparency as one way to bolster the reliability of research. Throughout the research process, researchers make decisions in how they prepare and analyze data which they draw on in their research papers. If the underlying data (and the statistical code) is not available to others, it can be impossible to check their work. Journals, including journals such as Nature and Science, have started to require that the materials required to reproduce (i.e. re-analyze) results are shared upon publication of articles. But norms can be slow to change: many journals do not yet have such a requirement, or do not enforce the requirement that they do have (see Appendix III for examples). Thus, funders may play an important role in re-enforcing and bolstering journal requirements. Many funders may also be in a position to ask grantees to share more data (i.e. larger, more comprehensive datasets) than journals commonly request.

Research funders began to adopt policies in the early 2000s. Some of the first research funders to adopt data-sharing policies included several of the UK Research Councils (Economic and Social Research Council and the Medical Research Council), which adopted policies in 2000 and 2005.[9] The National Science Foundation in the US adopted a data-sharing policy in 2008. Over two dozen major governmental and private foundations have followed suit in the subsequent decade.

Alongside the organizational policies, there have been several noteworthy initiatives. In 2007, OECD (a forum of 30 countries) issued "Principles and Guidelines for Access to Research Data from Public Funding."[10] In 2008, Wellcome Trust and the World Health Organization convened a meeting with a number of research funders and other stakeholders, and put forth a joint statement on sharing research data to improve public health. And in 2013, under President Obama, the White House's Office of Science and Technology Policy issued a memorandum on data access policies, asking all federally funded agencies in the United States which receive $100M+ per year in federal funds for research to adopt policies on data access.

Challenges for funders

There is widespread agreement among funders that data (and other materials such as statistical

[9] Sarah Jones, International Journal of Digital Curation (2012), 7(1), 114–125.
[10] http://www.oecd.org/science/sci-tech/38500813.pdf

code, software and surveys) are potentially valuable research outputs. But there are also big challenges in designing and implementing a data-sharing policy. Starting with these challenges will set the stage for examining how funder policies attempt to address them.

Challenge #1: Diversity in research norms

Many funders support a range of research projects from various fields. Transparency norms vary widely across different fields/subfields within the natural and social sciences. For example, while genomics researchers began to formulate open data standards twenty years ago, other domains such as clinical trials research has started the conversation about data-sharing much more recently (Vickers 2016). Where a field has just recently embarked on a conversation about data-sharing, there may be a dearth of infrastructure such as subject-specific data repositories, metadata standards and familiarity/support of data-sharing. Surveys of researcher attitudes towards data-sharing show these differences between fields (Federer 2015, Tenopir 2015, Van den Eynden 2016).

Challenge #2: Poor incentives

Researchers are primarily rewarded professionally for publishing articles in peer reviewed journals. Even though the conversation about rewarding data-sharing rather than publications alone has begun, few tenure review and promotion committees acknowledge and give credit for research activities beyond publication, such as data-sharing.[11] Researchers express concern that if they share their data before they publish results, others may use their data for their own publications (and render the original researchers' conclusions redundant). (Panhuis 2014, Tenopir 2015).

Challenge #3: Cost of data-sharing

While researchers may be able to share data quickly, data which is hastily put online may be poor quality. Data curators who work on making materials usable, including staff at ICPSR and Roper Center, report that they often find missing metadata. For example, datasets may be missing variables names which are essential for understanding what the variables represent. Other issues are missing or corrupt files, messy statistical code without clear connection to tables in publications, file formats which are sub-optimal for preservation purposes, and many other problems which could diminish the data's re-use value. Another big issue is that if data are not carefully checked for identifying information, the shared data may allow subjects to be identified. The importance of quality and confidentiality raises the question of what funders should invest, not just towards funding data collection and researcher time, but towards support for curating and sharing data.

[11] This is changing at a few institutions; see for example an initiative from researchers in the Netherlands on the Science in Transition initiative.

Elements of data-sharing policies

The policy review covers 27 data-sharing policies shared by private foundations and governmental funders on their websites.[12] Funder policies are heterogeneous: they differ not just in the substance of the requirement, but also in the elements that they include. Some are very thorough, covering where data should be shared, how long it should preserved, whether additional materials (e.g. metadata, statistical code) should be shared, and so on. Others are a few sentences and do not elaborate on each of these categories. Here we cover the basic elements that policies contain.

1. Principles about the value of sharing research data

Most policies (20 out of 27) describe the funder's motivation for asking grantees to share data. This element is a general statement at the beginning of the policy mentioning the value of research data for re-use and/or re-analysis.[13]

2. Requirement that grantees provide a data management plan (DMP)

The idea behind DMPs is that grantees describe their plans for sharing data, usually at the time that they apply for funding. 21 out of 27 funders with data-sharing policies require grantees to submit plans, though there is variation in what they require grantees to include in their plans. Most funders specify at a minimum that grantees share basic information such as:

- The data and other materials (e.g. statistical code and software) that will be generated in the course of the research.
- Where the data will be shared, whether it will be open or shared under restricted access, licenses that will be applied, and how confidentiality of participants will be safeguarded.
- Who will be responsible for managing the data and sharing process, both during the project and afterwards.
- If applicable, an explanation of whether there are any restrictions on sharing data (due

[12] All policies appear on funder websites with the exception of RWJF (which shares its requirement with select grantees in a grant agreement). All funders' policies are listed in Appendix II and in greater detail in a full spreadsheet here.

[13] For example, the Laura and John Arnold Foundation statement is: "To the extent datasets are not subject to confidentiality requirements, we believe that datasets should be shared as freely as possible. At the most basic level, even the most well-intentioned scholars can make mistakes that would never be revealed unless someone else could double-check the code against the actual dataset. Data sharing also enables scholars to check others' work for sensitivity to the assumptions or model, as well as to extend it via further analyses."

to confidentiality, legal or proprietary issues), and a request for exemption for the relevant data.[14]

- The estimated costs of preparing and sharing data, as well as costs for sharing data within a repository (if applicable).

3. Ensuring confidentiality

As one of the greatest potential risks of data-sharing is inadvertently identifying research participants, 20 out of 27 funders note the importance of safeguarding confidentiality when sharing data publicly.

4. Informed consent

Six policies mention obtaining informed consent from research participants. For example, the PCORI policy states that researchers should obtain, "Appropriate documentation of patient consent that permits data collected as part of the study to be de-identified, used for future research purposes and shared broadly with researchers not affiliated with the institution conducting the study."

5. Instructions about where to share data

The majority (19 out of 27) of funders specify that data should be shared in a publicly accessible data repository. 8 funders specify that grantees should share data in a subject-specific repository (also called a "domain" repository), if a relevant one is available. If there is no domain repository available, some funders mention general repositories which can be used, such as Figshare, Dataverse or Open Science Framework (OSF). For example:

Howard Hughes Medical Institute: "If a public repository has been agreed upon by the research community for a specific type of dataset (such as GenBank for DNA sequences, the Protein Database for X-ray structure coordinates and structure factors, or the Bio-Magnetic Resonance Bank for NMR data), the author(s) should use that repository to optimize the ability of others to compare, search, merge, and build upon the data."

6. Timeframe on when to share

[14] An example of a funder discussing this is DFID: "Exceptionally, exemptions may be granted to specific policy requirements. Generally, these will be granted only if doing so would lead to better development outcomes. Exemptions may also be granted on grounds of security, legal, ethical or commercial constraint. If you believe there are good reasons not to make some research outputs openly accessible or to delay their release, then these must be explained in the Plan. DFID will consider these requests and may grant an exemption"

The vast majority of funders (24 out of 27) specify when data should be shared. But funder policies vary widely on timeframes.[15] The most common timeframe is to require data-sharing with publication or soon after (12 out of 27 policies), or 12 months following publication (2 policies). In 3 cases, funders merely mention that data should be shared within a "reasonable time" or "as soon as possible." Other policies ask grantees to share data 6-12 months after the grant period (3 policies) or 12 months after data collection (1 policy). The remaining policies specify a timeframe combining the two: either by publication or within a certain period after the end of the grant period (or end of data collection), whichever comes first.

7. Specifying what is meant by "data" and how much should be shared

Some funder policies clarify what is meant by "data" in the policy, and what should be shared. The most common way of describing what should be shared is asking grantees to share the materials needed to validate findings shared within research publications (11 out of 27). Several funders go further and ask grantees to share collected data, whether or not the data is used in a publication (6 policies). Or they specify that they will consider the publication data as a minimum standard but may ask for further data to be shared depending on its likely re-use value (6 policies). In the remaining cases, the policy does not specify what is meant by data in the policy.[16]

8. Requirement to share metadata and other materials

Some funders (18 out of 27) specify that further materials should be shared, including metadata needed to interpret the data. Metadata include key information about the variables that are shared (e.g. variable names, value codes), and also information about the study (including details on the methodology and how data was collected).[17] Shared materials can often be unusable without metadata, so this element is a request that researchers include what is needed to interpret the data properly.

Funders do not specify in their policies exactly which elements about the data or study should be shared, in the vast majority of cases. Instead, they refer to metadata very generally. For example, NIH's policy is characteristic of the language used in many policies:

[15] See Appendix II.
[16] See Appendix II.
[17] For more information on metadata, see ICPSR's webpage, which provides a guide to one metadata schema (the Data Documentation Initiative) as an example. DDI is generally used for social science data. Other fields – particularly fields with robust data-sharing standards such as astronomy and genomics – have their own specialized schemas for which elements are important to share, in order to allow re-use of shared data.

"Regardless of the mechanism used to share data, each dataset will require documentation. (Some fields refer to data documentation by other terms, such as metadata or codebooks). Proper documentation is needed to ensure that others can use the dataset and to prevent misuse, misinterpretation, and confusion. Documentation provides information about the methodology and procedures used to collect the data, details about codes, definitions of variables, variable field locations, frequencies, and the like. The precise content of documentation will vary by scientific area, study design, the type of data collected, and characteristics of the dataset."

There are a couple of exceptions in which funders are a bit more specific about the metadata to be shared. For example: AHRQ specifies that "A published data set consists of at least one formal metadata document, the digital scientific data described by that metadata, and supplemental information provided to assist the data user. The metadata for scientific data will include, at a minimum, the common core metadata schema in use by the Federal Government, found at https://project-open-data.cio.gov/." And Arnold Foundation specifies that randomized experiments should share study-level metadata in keeping with CONSORT standards.

9. Covering costs for preparing and sharing data

Many funders (16 out of 27) mention covering the costs that grantees incur while preparing and sharing data. All but two of the funder policies ask that grantees describe these anticipated costs within their data management plan and budget for them in the initial funding application.

10. Requirement to share statistical code

Relatively few funders (8 out of 27) mention statistical code explicitly, though many more may be taken to implicitly refer to it by specifying that "materials needed to reproduce research publication results" should be shared.[18]

11. Enforcement mechanisms

11 out of 27 policies mention that they will consider compliance with the data-sharing

[18] See full spreadsheet for details. An example of a funder which does mention code explicitly is the Arnold Foundation: "Researchers should already, as a matter of course, produce well-annotated scripts to clean and analyze data. The final version of these scripts should be uploaded to OSF and made publicly available. Ideally, the final code scripts should enable another researcher to take the original raw dataset(s), clean and merge them as was originally done, and re-run the original analysis."

requirement in future funding applications.[19] The most common method of potential enforcement is to consider failure to comply in future funding decisions. The CDC policy mentions that "Awardees who fail to release data in a timely fashion will be subject to procedures normally used to address lack of performance (e.g., reduction in funding, restriction of funds, or grant termination)." NOAA states that "Past performance regarding data sharing and manuscript submission shall be considered when reviewing new awards." The American Heart Association likewise states that it will "spot check" compliance with the policy, and failure to comply may affect future funding.[20]

12. Persistent identifiers and data citation

A handful of funders (9 out of 27) mention that grantees should obtain a persistent identifier (e.g. a DOI) so that their datasets may be uniquely identified and cited, and that they should ensure that their publications cite and link to their data.

13. Preservation period

5 funders specify the minimum duration of time that data should be made available. The timeframes include: Research Councils UK and Institute of Education Sciences (10 years), PCORI (7 years), DFID and Cancer Research UK (5 years).

14. Registration of studies

3 funders mention that grantees conducting particular kinds of studies (e.g. clinical trials, randomized controlled trials) should register their study and analysis plan in a public registry prior to data collection.[21] This requirement goes beyond data-sharing, but is related, in that it is another effort to increase the reliability of research through transparency requirements.[22]

What should funders require?

Some elements of data-sharing policies, as described above, are very common and

[19] See full spreadsheet for details.
[20] The policy is too recent for AHA to have had experience with checking and enforcement, since none of the grantees have yet met the deadline to share data.
[21] Macarthur and Arnold Foundations mention pre-registration, as does AHRQ. Other funders may encourage or require registration elsewhere, but did not mention in their data-sharing policies.
[22] A group of funders including Wellcome Trust and Gates Foundation recently signed a common statement on a related requirement to report clinical trial results regardless of outcome in a public registry.

uncontroversial, and all of these are recommended elements of a funder policy.

- **Share metadata along with data:** Asking grantees to share metadata along with data is a commonsensical reminder to make data usable. While specific metadata standards vary by field, at the very least this information should consist in clear names for variables, value codes, a readme file describing key information about shared files and key information about the study (such as when it took place, methodology, how study participants were selected).

- **Unique identifier for datasets:** Asking grantees to acquire digital object identifiers (DOIs) facilitates citations of their datasets.

- **Data underlying an article should be cited within that article, and it should be clear where to access the data.**

- **Researchers should ensure that their consent form language allows for de-identified data-sharing.** A useful guide from ICPSR suggest recommended language for this purpose.[23]

On the other hand, there are aspects of policies which vary widely across funders, particularly on key questions such as how much data to share, where and when. We discuss these questions below, outlining reasons for making different decisions.

Should funders ask grantees to share a data management plan?

Most funders with policies (21 out of 27) ask that their grantees submit a data management plan (DMP) answering key questions (see the Moore Foundation's and Wellcome Trust's list of questions as examples). Several funders described DMPs as a very useful exercise to prompt grantees to think about both their plans for sharing and any associated costs to budget for, well ahead of time.

One funder that asks grantees to share plans said, "We've had people tell us, 'My research is better later because we had to think about this up front.' Writing a plan forces them to think through questions like: where am I going to share my data? What data am I producing? If you

[23] ICPSR describes the goal of the informed consent language as follows: "Promises in the informed consent can appear to limit an investigator's ability to share data with the research community. In reality, investigators can inform study participants that they are scientists with an obligation to protect confidentiality and still share the study data with the broad scientific community."

know someone is going to view your data, that makes you track it carefully and keep metadata. The mindset is different if you're thinking ahead. If you go back a year into the award, you can't reconstruct it."

While requiring a DMP is standard practice, one funder did worry that grantees would "use a DMP as a substitute for sharing data," and that they would just fill in the DMP saying that they weren't able to share data for some reason. Another funder that does require DMPS said that there's a danger that this process can be seen by grantees "as a box you tick" as opposed to an important part of doing research. However, one solution would be to emphasize that DMPs are not an end in themselves, but are merely preparation for sharing. Along these lines, a recent review of funder policies suggests that DMPs should be treated less as a paperwork requirement, and more of a "living document that form the basis of collaboration between researchers, funders, and data managers throughout the life of a research project." (Neylon 2017).

As far as timeframes for sharing data management plans, funders request DMPs along with the funding application. Any costs of preparing and sharing data are then included in the budget. The advantages of doing so at this early stage are: Asking that grantees plan early on and budget for any costs as well considering the data management plan as factoring into the merits of their proposal overall. One funder noted that if a grantee was collecting valuable data, but did not explain a thoughtful plan for sharing it, "that application would not get very far."

Recommendations:

- **Ask for data management plans along with funding applications to help researchers to plan key logistics ahead of time:** Particularly if funders do ask that grantees share collected data, it will be important to consider logistics ahead of time (e.g. in a data management plan). The burden of preparing data for public sharing after it is collected can be onerous, particularly if a strategy is not developed in advance and costs are not budgeted.
- **Emphasize that sharing should be the default:** While exceptions to open data are possible if data cannot be shared publicly due to confidentiality or proprietary/legal restrictions, the expectation is that data will be shared (and the DMP should describe a plan for doing so).

How much data should grantees share?

11 out of 27 funder policies ask that grantees share "publication data" (i.e. the data underlying

published research), at a minimum. On the federal level, the OSTP memo of 2013, which required federal funders receiving $100M+ in research funds to create data-sharing policies, described data as "material…necessary to validate research findings including data sets used to support scholarly publications." This language is echoed in many of the federal funder policies (and plans for policies).[24]

6 out of 27 funders go further to require that data is shared beyond the publication data. For example, the American Heart Association asks for all data, regardless of whether the data is used in a publication. 6 funders, including Arnold Foundation and Wellcome Trust, ask grantees to share the data underlying publication at a minimum, but do sometimes ask for more data to be shared on a case by case basis, taking into consideration the value of the larger dataset for re-use. (See appendix II for details of each policy).[25]

As one funder noted in conversation, what funders ask grantees to share "really comes down to a question for the funder: Why do they care about it? If mostly for reproducibility of published results, then probably asking for what journals ask for makes sense. If for re-use of data, then asking for more makes more sense."

One funder that asks for the collected data rather than publication data noted that at first, applicants were startled, since this wasn't the norm they were used to. But, the funder said, "Most of big concerns that funders have, that everyone is going to opt out, that no one is going to apply, wasn't a problem. We got past that so quickly, and it didn't affect application volume. Definitely it was a burden initially to educate people. Just like anything else that's new there were some hurdles, but wasn't the amount of revolt we were worried about."

There are advantages and disadvantages of adopting a more narrow scope such as sharing publication data, as opposed to asking grantees to share the larger set of collected data:

[24] Note that the distinction between "publication data" and "collected data," while common and useful for practical purposes, can easily be more complex than we have space to discuss. The materials considered to be "underlying" the publication could be interpreted differently from field to field or even within the same field. For a helpful discussion of this complexity, see Borgman 2016, Chapter 10.

[25] We can think of data-sharing as a spectrum. On the far left of the spectrum is only an article (often a summary based on underlying materials), without any of the underlying data or code used to create the tables/figures in the article. More transparent is to share the variables used to produce the summary results included in the article. Still more revealing is to share both data and code used to produce the results. Sharing the larger collected dataset(s) in addition can add considerably more information, since the larger dataset would (in many cases) include data that goes beyond what was summarized in the article. Finally, most transparent of all would be to share the original (raw) data, along with all the instructions (ideally in statistical code) used to transform the original data into the publication dataset, and the analysis code used to produce summary results. It is possible to embed the underlying data and transformations into the summary results themselves (making the results computationally reproducible without additional effort). For one group working on this, see Code Ocean (https://codeocean.com/).

Category	Materials underlying publications (data, metadata, code)	Collected data (i.e. de-identified data collected during the course of research, even if never used in a publication)
Reproducibility	Helps to address some reproducibility concerns by enabling others to try to re-analyze results using shared materials. (However, may be limited when it comes to "checking" results, since important information such as omitted observations and variables may not be included).	May not address reproducibility concerns, if researchers do not also share publication data and code, since others may not be able to re-trace the steps that researchers took to clean, transform and analyze their collected data in order to produce results in articles.
Re-use for secondary analysis	Some grantees may not publish at all for a long period of time, or they may publish one article using a small subset of their data. A lot of the potential value of the data may go to waste.	Permits secondary analysis, assuming that data is of sufficient quality (e.g. study-level and variable-level metadata is included to aid interpretation)
Cost	For some studies, may require less time to clean, de-identify and otherwise prepare publication data (since only a small subset of data will be used in the publications.) This varies widely, however.	For some studies, may require more time to prepare (varies widely, however).
Researcher incentives	Researchers have already used data, so not as much fear of having others "scoop" them.	Researchers may strongly argue that they haven't had time to publish (depends on timeframe).
Research norms	This standard is increasingly expected by journals in many fields (with some exceptions such as clinical trial research)	Journals do not ask for data beyond what is used in a publication. Not yet a norm in many fields to share all data. Re-use value may be clearer in some fields than in others.

Recommendations:

- **Publication data, shared in a public repository and de-identified, as a minimum standard:** The considerations point to a minimum requirement of sharing materials needed to re-analyze the results in an article. This "publication data" would include not just the final dataset underlying results, but also the statistical code used to analyze the data, and the metadata needed to interpret the data and other materials.
- **Consider asking grantees to share more data on a case by case basis, considering the likely value for re-use:** Some funders ask for publication data as a minimum standard. They then require more extensive data-sharing on a case by case basis, depending on the likely re-use value of the data. This can allow a good balance of benefits from shared data and cost/time.

Where should grantees share data?

There are several questions when it comes to where to share data. First, there's the question of storing data within a public repository rather than informal channels such as available upon request or on researchers' websites. Second, there's a choice between an open data repository versus restricted access. Open data is available to anyone to download (though it may be either free or cost a fee). Restricted access data is available only to some: the restriction may be based on applying for access, or it may be available only to members of certain institutions. Finally, there is a question of which repository to use. There are hundreds of repositories, from subject specific repositories (GenBank for DNA sequences, for example) to institutional repositories hosted by universities, to general repositories which can be used for any virtually any kind of data.[26]

Many funder data-sharing policies (19 out of 27) require that data is shared in a repository of some kind, as opposed to via request or researcher websites. The 2013 OSTP memo, one requirement of US federal funder data-sharing policies is that they "promote the deposit of data in publicly accessible databases, where appropriate and available." The reasons for preferring a repository to availability upon request or on a researcher's website are clear: if only available upon request, researchers often fail to respond to or deny requests (Vines 2013). Repositories offer more stable and long-term storage solutions than a link to a personal or university website which may easily break over time. Many repositories also offer digital object identifiers (DOIs) which facilitate data citation (see more on the importance of citation in the

[26] See re3data.org for a list of 1,500 research data repositories.

"incentives" section below), and have Creative Commons re-use licenses (CC0 and CC-BY) as a default for share materials (see Appendix III). Finally, storing in a repository allows better data discoverability than simply placing on a website, as repositories frequently have search functions which allow data users to search by keywords or other metadata.

Funders are aiming to gain the widest possible data re-use, in asking grantees to share in public repositories. The exception to this may be where a community of researchers is likely to have improved data-sharing incentives if data is initially shared within a data enclave available only to members.[27] In addition, in some cases, data must be available only upon request because of the potential to identify participants (this can be a concern particularly in sharing sensitive health data). In other cases, it is possible to create both a public use version of data as well as a restricted use version which has more detailed information.[28]

Finally, on the question of which repository to use: some funders which award grants to diverse research projects give general rules for which repositories to use, rather than specifying particular repositories.[29] Eight funder policies recommend that grantees share in a domain repository where possible, so as to improve the discoverability of data. A few funders mention guidance in selecting a repository (see AHA guidelines) or provide a list of recommended repositories (see Wellcome Trust's list). We discuss available repositories at greater length in section IV.

Recommendations:

- **Require that grantees share data in an open data repository**, rather than a researcher website or available upon request.
    - If data cannot be de-identified and/or are extremely sensitive, then sharing in archives with restricted access is one solution.[30]

[27] One interviewee reflected that while the goal of having data dispersed as widely as possible makes sense as a value, there are other ways of sharing which might provide better incentives to researchers. One such model is that of a repository whose members get first, priority access to their jointly shared data. An example is the case of the Sloan Sky Survey (Borgman et al 2016). Researchers share data with others in a closed network (within a data repository) and are granted exclusive rights to access data for one year. After that point, the datasets are available to the wider public. However, we don't discuss this at greater length because it wouldn't necessarily affect recommendations for a funder policy.

[28] This dual-purpose solution has often been the case with RWJF data shared in the HMCA archive.

[29] The exception is funders with a relatively specific domain, such as certain NIH institutes, which require their grantees share within a domain repository. Another noteworthy exception is the Economic and Social Research Council in the UK, which funds the UK Data Archive, where it requests that its grantees share data.

[30] Whether it is most appropriate to de-identify data, share data in a restricted access repository, or request an exception to a data-sharing policy, depends on details of a case. Funders often mention that grantees should work

- **Where available, domain repositories are preferable in order to promote data discoverability**. If one is not available, then a general repository should be used (see section IV on repositories below for discussion).

What should the timeframe be for sharing data?

Different funders give very different answers to this question in their policies. Part of the reason for the differences is certainly that they vary in the data they're asking grantees to share. As we saw above, some funders ask only that grantees share publication data. Of 18 policies asking grantees to share publication data at a minimum, 13 specify that data should be shared at the time of the publication or soon after (see Appendix II). The remaining do not specify a clear timeframe or mention that data should be shared "as soon as possible."

On the other hand, if asking grantees to share collected data, the time of data release need not coincide with the time of publication. Funders that ask for the larger dataset as opposed to publication data commonly ask for the data to be shared between 6-12 months following the end of data collection or following the end of the grant period. Examples of timeframes include policies of the American Heart Association, which asks grantees to share data 12 months following the end of the grant award period, as well as Department for International Development (DFID), which asks that grantees share data 1 year following the end of data collection.

Another option is to have a shorter timeframe for sharing data, but to offer the possibility of an extension upon request. Some Research Councils in the UK have a shorter timeframe, such as Economic and Social Research Council, which asks grantees to share data within 3 months following the end of the grant period. However, the policy also allows grantees to request an embargo period of 12 months following the end of the award, if needed in order to publish results.

There's a trade-off involved in setting a timeframe. For the collected data especially, researchers want to be able to use the data for their own publications, rather than opening up the data for anyone to use. As one funder mentioned, "[Especially] if you're funding in a LMIC (low or middle income country), there are all sorts of issues there. One of the main ones being that there's often little capacity in those areas to do a lot of analysis very quickly. The danger is that if they share all the data on day one, they won't have time to do analysis."

with program officers to develop a data management plan which maximizes the value of shared data while taking into account the importance of maintaining confidentiality of participants.

On the other hand, there are also risks with delays. As another funder said, "With regard to open data: researchers are going to sit on studies for years. We should have that data [much sooner]…It's about saving lives, getting the research to the people that can use it." Selecting a timeframe for sharing is a balance between allowing time for the original team to analyze their data and the value of the public good.

Recommendations:

- **Publication data shared at the time of publication.** If the requirement is that grantees share publication data, asking materials to be shared at the time of (each) publication is reasonable.
- **Collected data, if requested, shared 6-12 months after the end of the grant.** Allow some additional time for grantees to share their collected data (if applicable). A timeframe of 6 months to 12 months after the end of the grant, or after the completion of data collection, is in keeping with other funder requirements.

What can funders do to improve data-sharing incentives?

In surveys of research attitudes towards data-sharing, many researchers identified professional career concerns as some of the major reasons not to share data (Tenopir 2011). First, researchers worry that others will publish using their shared data, before they do so. Second, they worry that they will not receive credit when others use their data, either in the form of co-authorship (if appropriate given their field's norms) or through data citation. Another area of concern is the cost of preparing data to share.

On the issue of professional credit, funders can do a couple of things to try to improve incentives. First, they can require that grantees share data with a unique identifier such as a DOI, so that their data can be more easily cited by those that rely upon it. Second, funders can ask that researchers cite their data in their papers so that others are able to use and cite their data more easily.

Another potential way to reward data-sharing is for a funder to consider shared data as one part of grant applications, alongside publications. Wellcome Trust is an example of a funder that invites prospective grantees to list data and other scholarly research outputs in addition to journal articles within funding applications.[31] By considering shared data to be a valuable research product, funders can promote the norm that shared data, and not just

[31] Wellcome Trust noted in conversation that work is ongoing to refine this further.

publications, counts as a professional contribution.[32]

The second concern – regarding the financial cost – is the one that the funders can address by fully supporting the costs of fulfilling their requirements. Many funder policies state that they will cover costs associated with data-sharing (16 out of 27 policies), and ask that grantees budget for staff or researcher time to share data, when they apply for funding. Covering costs associated with data preparation should be something that funders commit to covering, if they require data-sharing.

Some funders also independently pay for data curators to ensure the quality and completeness of shared materials.[33] The clearest benefit of paying for third party data curation is certainly in improving datasets which will be used by many researchers for their own analyses in the future. Another big advantage is the security of having an additional check that no identifying information is accidently released. (One open data expert mentioned that when researchers share data on their own in portals, "the responsibility for not releasing confidential data rests with depositor… and they can make mistakes.") Data curation can also play a crucial role of standardizing metadata across a number of similar datasets (and this is often the purpose of specialized repositories which share only a very specific kind of data, such as the polling data shared by Roper Center).

Finally, funders can consider how they are willing and able to enforce their data-sharing requirements. 11 out 27 funder policies mentioned some form of potential enforcement. Where funders do mention penalties, the most common (7 out of 11) is to say that they will consider compliance with data-sharing requirements when grantees apply for future funding. Another potential mechanism is to withhold the last grant payment (say 10% of the total) if grantees fail to share data (1 out of 8 policies).

However, enforcement is clearly a work in progress. None of the funders that I spoke to has enforced their policies (though for some of them, the policy is so new that the timeframe for sharing has not yet been reached by any grants).[34] Researchers share data in many different repositories, sometimes years after grant periods end. As a result, it is a serious commitment to following up long-term with grantees.

The best approach may be to follow the lead of the funders which indicate that failure to

[32] Another possibility mentioned by one funder is to develop a system where they make data management plans publicly available, connect them to researcher IDs (such as ORCID), and then credit researchers for having fulfilled the plan laid out in the DMP. This idea was discussed, but not yet implemented by the funder.

[33] We are aware of the examples of Economic and Social Research Council in the UK and RWJF. (There may be others, particularly Institutes or Directorates within NIH/NSF which maintain specific-specific repositories for their grantees).

[34] Nor are there – to my knowledge – public accounts of such enforcement on the part of funders.

comply may affect future funding decisions. This route does not require tracking compliance on every article that a grantee publishes (which may not be feasible due to burden of time on staff). Rather, it leaves open that a funder would ask grantees to demonstrate compliance when they apply for future funding, and that this would be one factor when considering the merit of such decisions.

Another helpful measure is to ensure that grantees are aware of the policy. In a recent survey, 20% of researchers were unaware of whether their funder required them to make their data open, and over half expressed that they would welcome more information on how to comply with a funder's policy (Fane et al 2016). In conversation, funders recommended clearly communicating with grantees. In addition to posting the policy on a website and in grant agreements, funders mentioned a couple of strategies for communication. One funder said, "What helped the most is when we put out sample data plans—at first, 80% of the plans had to be re-done. We had to tell them, 'Do not reference presentation or publications in your data plan. What data are you going to produce and where are you going to put it?'" Another funder stressed talking with researchers about their concerns: "If it's just the concern that I can't get a grant because I won't have enough publications, then I'd work on rewarding data-sharing. If it's concern around data being misused, address that. We need to understand what the issues are."

Recommendations:

- **Ask that grantees specify the costs for data-sharing in their data management plans**, and commit to covering those costs. Particularly if data beyond the publication data is shared, it can take considerable staff or researcher time to prepare data for public use. Also take into account any costs for sharing within repositories.
- **Require that grantees acquire a unique identifier when sharing data (such as a DOI)**, so that others are able to cite their data more easily and give them professional credit, and to cite their data in articles.
- **Consider professional incentives for rewarding data-sharing,** such as asking grantees to list their shared data in funding applications (following the example of Wellcome Trust).
- **Clearly communicate with grantees about the policy and how they can comply with it.** Consider sharing sample data management plans and offering resources (potentially training and workshops) on data-sharing.
- **Consider stating in the policy that failure to comply may affect future funding decisions.**

III. Journal data-sharing policies

Journals, as well as funders, have a role to play in the shift towards more transparent and reproducible research. A few key questions: What do journals in these areas require researchers to share, and when? How do the roles of journals and funders complement each other? In order to review journal policies, we compiled a list of 17 journals in which RWJF grantees often publish (Appendix III).[35] We also drew on the findings of a systematic review of biomedical journal policies (Vasilevsky 2017).

Of 17 journals which we reviewed, 8 require data-sharing and 3 journals encourage it (the remaining 5 do not mention data-sharing). The list includes journals which publish a range of scientific research – including Nature, Science, and PLOS (all of which have policies) – as well as some journals specific to public health.

All of the journal policies which we reviewed require authors to share materials underlying the published results. A few journals, including Science, also require that researchers share large datasets (going beyond publication data) which are of high value for reuse, including microarray data, protein or DNA sequences, atomic coordinates or electron microscopy maps for macromolecular structures, and climate data.

The journals vary in whether they ask researchers to share upon request (1 policy out of 8) or to deposit data in public repositories (4 out of 8 required sharing in repositories and 3 encouraged it). Enforcement strategies also vary: of the 8 journals with the requirement to share data, only 4 mentioned a penalty for failing to share.[36]

A recent systematic review of biomedical journal data-sharing policies (Vasilevsky 2017) found that of 318 journals, about 20% require data-sharing, with 12% explicitly stating that articles would not be published without the underlying materials shared. About 15% of journals only mentioned data-sharing for genomics or other omics data. Another 23% of journals encouraged but did not require data-sharing.

Our review demonstrates a couple of points regarding the complementary role of journals and

[35] An RWJF program officer compiled this list. For fuller details, see also a shared spreadsheet.

[36] One journal, Milbank Quarterly, mentions a range of penalties for failing to share upon request, up to removal of the article from the journal. PLOS enforces the policy by refusing to publish the article unless the materials have been deposited. And Nature mentions that if researchers fail to release materials to readers, the journal may post a statement of correction stating that "readers have been unable to obtain necessary materials to replicate the findings."

funders. First, journals requiring data-sharing are still a small fraction of the total pool (20% of biomedical journals, and 43% in our review of RWJF-relevant journals). Policies vary in whether they have any enforcement strategies in place, and whether they require grantees to share data in public repositories. Thus, funder policies can cover a gap left by many journals which haven't adopted policies, or which don't ask that grantees share data in public repositories.

Second, very few journals ask that researchers share more than the data used in the publication (as opposed to the wider collected dataset). As we have seen above, this is something which some funders do ask of selected grantees, particularly where the collected data (as opposed to a small subset underlying a publication) is of value for re-use. Funders can fill a role there by going beyond the purview of journals.

Third, funders are involved at a much earlier point in the research lifecycle (before data collection) than journals. The timing allows funders to ask grantees to think about data-sharing before they even begin to collect data, in asking for data management plans. This can greatly help researchers in being more transparent downstream, and can raise the reliability of the research by making researchers aware of their workflow and data management strategies.

Finally, changing norms is certainly going to require involvement of all stakeholders, not just journals. Where funders ask grantees to share data, they demonstrate that transparency is a valuable part of research.

IV. Data Repositories

The majority of funder data-sharing policies ask that grantees share data in a public repository (19 out of 27). Repositories are either "domain repositories" which host data from particular fields or subfields, institutional sites (often hosted by universities), or general repositories. 8 of the policies ask that grantees share in a domain repository if available, which refers to a repository which houses subject-specific data. Sharing in a domain repository can boost discoverability of datasets by being a place where researchers know to search for the data. Several funders and journals provide lists of recommended domain repositories (for example, see Wellcome Trust's and Nature's pages. A catalog of repositories is also available at Registry of Research Data Repositories (Re3data.org). Institutional data repositories such as those hosted by universities are another option to consider.

If a domain repository is not available, there are a number of general repositories, most of which are open to any kind of research data: these include Dataverse, Dryad, Figure, Open Science Framework and Zenodo (see Appendix III for a list of these repositories and some of

their features, as well as further details here).

What are the features to look for in choosing a repository, or in recommending a repository to grantees? The Data Curation Centre in the UK provides a useful checklist for considering repositories (Whyte 2015). Its five basic questions include:

- Is the repository reputable?
- Will it take the data you want to deposit?
- Will it be safe in legal terms?
- Will the repository sustain the data value?
- Will it track data usage?

Reputation: there are several ways to assess repository reputation. One way is to check whether the repository has been certified. There are several international bodies which certify repositories, including the Data Seal of Approval. However, the process for certifying repositories is relatively new, and there are many repositories which meet high standards of quality and yet are not certified. An alternative is to check whether the repository is recommended by funders and journals (see Wellcome Trust and Nature lists).

Data type: Some repositories restrict the kind of data or file format that they accept.

Legal terms: Repositories sometimes have default license agreements. Many of the general repositories listed below have adopted Creative Commons 4.0 licenses, with a default of the permissive licenses (CC-BY and CC0) which enable others to re-use and build upon the shared materials. This is important for permitting other researchers to use the data for further analysis.

Another legal issue is protection of research subjects. Repositories vary widely in the degree of protection for restricted access data they offer. In many cases, funders and journals ask that researchers de-identify data and share what they can openly, omitting any direct identifiers such as names, as well as indirect identifiers such as narrow geospatial data. Yet in some cases, it is not possible to remove potentially identifying information (PII) without much reducing the value of the data to other users. Many repositories include capacity to limit data-sharing to restricted-use files which are password protected, in which users can be given access only upon permission. Some repositories, including **ICPSR**, have several levels of protection, including high levels of protection such as a “virtual data enclave” and “physical data enclave.”

Sustainability: Is the repository likely to be sustained over the long-term, and is there a plan place for safeguarding data if the repository shuts down?

Preservation: Does the repository make files available to users in non-proprietary file formats, which are more accessible and less likely to become unusable over time? Does the repository offer version control, in order to store prior versions of files?

Discoverable and accessible: Are files discoverable by widely used search engines? Does the repository enable searching by metadata such as keywords?

Citable: Does the repository assign DOIs so that datasets can be uniquely cited and attributed when re-used?

Tracking data usage: some repositories track and display use metrics such as downloads, unique users, and views. These are helpful for gauging the use value of shared data. Other repositories (such as Dataverse) enable data providers to enter their own questions for those who download datasets, which can be useful for finding out more information about who is using data and for what purposes, without restricting the datasets.

Recommendations:

- **Ask grantees to share in a subject-specific repository which is reputable in their discipline if one is available.** Consider offering grantees a list of reputable general repositories which they can use, if no subject-specific repository is available (see Appendix III).
- **Require that the repository assign unique identifiers (DOIs)** so that data can be cited more easily. Regardless of which repository is used by researchers, it is important that the paper cite the dataset, so funders should also ask that grantees cite their own associated data clearly in publications.

V. Conclusion

Data-sharing is still relatively new to researchers, journals and funders in most fields. While this means we're on an exciting forefront, it also means that there's a challenge of discerning best policies and practices, often in the absence of strong empirical evidence. There's a danger of trying to change too much too soon, but there's also a danger of waiting. In order to move towards greater transparency, no single funder, journal or association can act effectively alone. The ecosystem is tied together and must move forward together.

The recommendations presented throughout and distilled in the executive summary are neither radically beyond what other funders have taken on, nor are they conservative. They will require additional investment from RWJF, but will also allow the foundation to keep supporting the broader transparency movement and maximizing the values of its investment in research.

Appendix I: Interviewees

Person	Organization	Position
Alon Axelrod	Inter-university Consortium of Political and Social Research	Archive Manager, Health and Medical Care Archive
Belinda Orland	American Heart Association	Senior Manager of Research Operations
Brian Quinn	Robert Wood Johnson Foundation	Associate Vice President, Research-Evaluation-Learning
Carolyn Miller	Robert Wood Johnson Foundation	Program Officer
David Mellor	Center for Open Science	Project Manager, Journal and Funder Initiatives
Greg Tananbaum	Open Research Funders Group (as well as SPARC)	Consultant
Jason Gerson	Patient-Centered Outcomes Research Institute (PCORI)	Senior Program Officer
Jennifer Hansen	Gates Foundation	Senior Officer, Knowledge & Research
Joshua Greenberg	Sloan Foundation	Program Director for Digital Information Technology
Kathleen Weldon	Roper Center	Director of Data Operations and Communications
Kelly Hunt	Hunt Strategy Group	Senior consultant
Margaret Tait	Robert Wood Johnson Foundation	Research Associate
Maryrose Franko	Health Research Alliance	Executive Director
Matthew Trujillo	Robert Wood Johnson Foundation	Program Officer
Oktawia Wojcik	Robert Wood Johnson Foundation	Program Officer
Robert Kiley	Wellcome Trust	Open Research Development Lead
Sindy Escobar	Doris Duke Foundation	Senior Program Officer for Medical Research & Co-Chair of the Health Research Alliance Open Science Task Force

Stuart Buck	Arnold Foundation	Vice President of Research Integrity

Appendix II: Funder Policies[37]

For the sake of concision, this table is condensed from a more comprehensive version (which includes additional elements such as enforcement mechanism and retention period) available here.

Definitions: "Publication data" refers to data and other materials (e.g. code) underlying published result. "Collected data" includes the data that was collected in the course of the research, with the exception of confidential and/or proprietary data.

Funder (clickable link)	Which data to be shared?	Timeframe (deadline for sharing)	Where to share?[38]	Policy specifies that data management plan required?
Agency for Heathcare Research and Quality (AHRQ)	Publication data and collected data where feasible	With publication (for publication data; beyond that timeframe varies)	Public repository	Yes
American Heart Association (link)	Collected data	One year after grant period	Approved public repository	Yes
Arnold Foundation (link)	Publication data & data beyond that case by case	With publication or end of grant period (whichever comes first)	Public repository	No
Bill and Melinda Gates	Publication data	With publication (12 month	"Accessible and open"	No

[37] This list of includes private foundation policies identified through online research and conversations, as well as "umbrella" government funder policies (i.e. which apply to all sub-groups of the funder). There are many individual institutes/directorates within NSF, NIH and UK Research Councils with their own policies, but they were not included here individually. In addition, there are some federal funders that have plans to adopt policies as of May 2017, but had not yet adopted policies – these plans are not included in the table below.

[38] Here a "public repository" can mean a site intended for data-sharing, whether a domain repository, institutional repository or general repository. Where funders specify one repository or type of repository in particular, this is noted in the full spreadsheet. Public repositories can also include repositories which allow for restricted access to data as needed to safeguard identities of participants or proprietary data.

Foundation (link)		embargo may be applied)		
Canadian research funders (link)	Not specified[39]	By publication, for publication data	Public repository	Yes
Cancer Research UK (link)	Not specified	By acceptance of publication (or according to norms in the field)	Various methods accepted	Yes
Center for Disease Control (CDC)	Collected data	One year after quality evaluation	Public repository or sharing with partners (where there are confidentiality concerns)	No (optional)
Department for International Development (DFID)	Collected data ("Raw or derived datasets")	One year after final data collection or with publication (whichever comes first)	Public repository	Yes
Department of Education (link)	Publication data	With publication	Various methods accepted	Yes
European Research Council (ERC) Horizon 2020	Publication data & data beyond that case-by-case	"As soon as possible"	Public repository (preferred)	Yes
Food and Drug Administration (link)	Publication data	With publication	Public repository (preferred)	Yes
Howard Hughes Medical Institute (link)	Publication data	"Following publication"	Public repository	No
Macarthur Foundation (link)	Publication data	Not specified	Not specified	No
Moore Foundation (link)	Publication data	"As soon as possible" (with 6 month timeline	"Openly and freely available"	Yes

[39] "Not specified" means that it wasn't clear from the policy whether the funder is requesting data underlying publications or collected data; in some cases, the funder states that the appropriate data to share depends on the community and varies (see full spreadsheet for details).

		for conservation and genomic data)		
National Aeronautics and Space Administration (NASA)	Publication data	With publication or "reasonable time" after	Public repository (can also be supplemental information)	Yes
National Institute of Standards and Technology (NIST)	Publication data	Not specified	Public repository	No
National Oceanic and Atmospheric Administration (NOAA)	Collected data	With publication or 2 years after data collection or 2 years after grant (whichever comes first)	Public repository	Yes
National Science Foundation (NSF) See future plan	Varies[40]	"Within reasonable time"	Not specified (depends on community norms)	Yes
National Institutes of Health (NIH) (See future plan)	Publication data and collected data; only applies to grants of $500K+ in a single year	By publication (for publication data)	Various methods accepted	Yes
National Institute of Standards and Technology (NIST)[41]	Publication data	Within 12 months of publication	Public repository	Yes
Office of the Director of National	Publication data	At time of publication	Public repository (preferably the DTIC repository)	Yes

[40] In the policy, NSF refers to "the primary data, samples, physical collections and other supporting materials created or gathered in the course of work under NSF grants." However, in the FAQ: "What constitutes such data will be determined by the community of interest through the process of peer review and program management. This may include, but is not limited to: data, publications, samples, physical collections, software and models."
[41] Note: document is called a "plan" but specifies it is effective as of Jan 1, 2017

Intelligence (ODNI)				
Patient-Centered Outcomes (PCORI)[42]	Collected data	By completion of study	Public repository (list of approved repositories in progress)	Yes
Research Councils (RCUK)	Publication data minimally; beyond that data that is valuable for re-use	With publication	Public repository (specific councils have more specific guidelines – see link)	Yes
Robert Wood Johnson Foundation (language used in grant agreements[43])	Collected data	Within 12 months of end of grant period	HMCA archive (curated by ICPSR)	No
Sloan Foundation (Grant guidelines Pg 8-9)	Reviewed case by case	Reviewed case by case	Reviewed case by case	Yes
Smithsonian Institution (link)[44]	Publication data (for journals which require it); further data case by case	12 months after publication	Public repository	Yes
U.S. Geological Survey (USGS)	Publication data	With publication	Public repository	Yes
Wellcome Trust (link)	Publication data & data beyond	With publication (or as agreed-	Public repository	Yes (for some grants)

[42] Some aspects still in development, such as a pilot of repositories

[43] "If one of the deliverables described in Section 1 is a public use data set for inclusion in the Foundation's Health and Medical Archive, you shall, at no additional cost to us, cause public use data files to be constructed (with appropriate adjustments to assure individual privacy) in accordance with the specifications of the Inter-University Consortium for Political and Social Research, University of Michigan, including the full documentation outlined in the Consortium's current data preparation manual. Unless we otherwise specify, such public use data files shall include all data files used to conduct the analysis under the grant. You shall transmit one computer-readable copy of such public use data files and documentation to the Consortium within 12 months of the expiration of the grant period. A portion of your final payment up to 10 percent of the grant award amount may be withheld until this deliverable has been received."

[44] This document notes that it is a "plan" but that it is effective as of Oct 1, 2015

	that case by case	upon in DMP)		

Appendix III: Repositories

See more detailed spreadsheet here.

Repository	Scope	Public archive?[45]	Curation available?	Use Metrics	Cost & File size limit[46]	License
Dataverse (link)	Accepts data from any field[47]	Yes	No	Downloads	No charge for up to 2 GB per file; no restriction on number of files	CC0 default[48]
Dryad (link)	Data underlying international scientific and medical literature	Yes	Yes[49]	Downloads	$120 for 20 GB (max file size of 10 GB)	CC0 default
Figshare (link)	Accepts data from any field	Yes	No	Number of downloads, views, times shared, and	No charge for up to 20 GB total, 5 GB file size	CC0 default

[45] This category refers to whether members of the public have access to data (with the exception of data that is restricted access for confidentiality reasons).
[46] These are size limits for normal individual usage; some repositories may permit larger file sizes for members of partner institutions or work with users on a case by case basis.
[47] Dataverse notes that about half of datasets are from the social sciences. "Medicine, health and life sciences" comprise 12.2% of the archive's data. https://dataverse.org/metrics
[48] "CC0" refers to the Creative Commons Zero Public Domain Dedication Waiver. https://creativecommons.org/publicdomain/zero/1.0/legalcode.
[49] "A curator will check your files for technical problems before they are released." More details in FAQ: "can the files be opened? are they free of viruses? are they free of copyright restrictions? do they appear to be free of sensitive data?). The completeness and correctness of the metadata (e.g. information about the associated publication, the date on which any embargo is to be lifted, indexing keywords) are checked and the DOI is officially registered."

				citations		
ICPSR (link)/ Open ICPSR	Largely social sciences, as well as public health	Yes for OpenICPSR; ICPSR is typically restricted to members (except where funder requests open)	Yes	Downloads / unique users / institutions of users[50] For ICPSR: list of publications that use data	Prices vary (OpenICPSR self-deposit is free up to 2GB)	OpenICPSR: Attribution 4.0 Creative Commons license
Open Science Framework (OSF)	Accepts data from any field	Yes	No	Unique visits, downloads (link)	No charge for up to 5 GB per file; no limit on number of files	Select from a variety of licenses
Roper Center (link)	Public opinion data	Mostly accessible only to member institutions (datasets available to others for a fee)	Yes	Downloads available on request	No charge to data contributors (there is a charge for accessing data)	N/A
Zenodo (link)	Accepts data from any field	Yes	No	None publicly visible	No charge for up to 50 GB per dataset	Select from a variety of licenses

[50] Note that these are the metrics that are visible for the HMCA archive (used by RWJF-funded researchers)

Appendix IV: Journal Policies

RWJF grantees publish in the following journals regularly (put together with the input of RWJF program officers). The review gives a sense of the prevalence of data-sharing policies of journals which are relevant to the foundation (and is not meant to be a systematic review of journal policies or a representative sample of public health/biomedical journals). See spreadsheet with further details here.

Journal	Data-sharing policy?	Which data and where to share
Cell (link)	Yes	Must deposit some kinds of data to a public repository; other kinds must be available upon request or shared publicly at publication.
Milbank Quarterly (link)	Yes	Make available on request
PLOS (link)	Yes	Share all data underlying published results either on request or publicly; sharing in repository strongly recommended.
Proceedings of the National Academy of Sciences (PNAS)	Yes	Share all data underlying published results either on request or publicly; sharing in repository encouraged.
Science (link)	Yes	Large datasets must be deposited in a subject-specific or institutional repository prior to publication. All underlying materials must be available after publication (may be shared on request) to "any reader of Science."
Nature (link)	Yes	"Supporting data" must be available to editors and peer reviewers at the time of submission. For specific kinds of datasets (laid out in policy), must be made public in repositories. All publications must include data availability statements and

		provision of "minimal data set" underlying publication is encouraged.
F1000Research (link)	Yes	Data underlying results must be shared in an open repository (subject-specific if available)
Journal of the American Medical Association (JAMA)	Required for some datasets	Large genomic datasets must be shared in relevant repositories.
Lancet (link)	Not required but encouraged	Encouraged to include a digital object identifier (DOI) at the end of the Methods section.
Preventive Medicine (link)	Not required but encouraged	Encouraged to include data references in the paper including: name, dataset title, data repository, version (where available), year and global persistent identifier.
Social Science and Medicine (link)	Not required but encouraged	Encouraged to include data references in the paper including: name, dataset title, data repository, version (where available), year and global persistent identifier.
American Journal of Preventive Medicine (link)	Not required	N/A
Health Affairs (link)	Not required	N/A
American Journal of Public Health (AJPH)	Not required	N/A
Journal of Epidemiology and Community Health (link)	Not required	N/A
Journal of Public Health Research (link)	Not required	N/A
New England Journal of Medicine (NEJM)	Not required	N/A

Appendix V: Further Resources

There are a number of resources which may be useful to funders and grantees, and this section highlights a variety of these guides and initiatives.

Resources on data curation and data-sharing:

- ICPSR's guide to data preparation and archiving: http://www.icpsr.umich.edu/files/ICPSR/access/dataprep.pdf
- UK Data Archive's Guide to creating and managing data: : http://www.data-archive.ac.uk/create-manage
- Digital Curation Centre, metadata standards guide: http://www.dcc.ac.uk/resources/metadata-standards
- Nine simple ways to make it easier to (re)use your data: https://ojs.library.queensu.ca/index.php/IEE/article/view/4608/0
- Ten Simple Rules for the Care and Feeding of Data: http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003542
- FAIR data principles developed by FORCE11: https://www.force11.org/group/fairgroup/fairprinciples

Case studies on data-sharing from funders, repositories and others:

- LEARN Toolkit of Best Practice for Research Data Management: http://learn-rdm.eu/wp-content/uploads/RDMToolkit.pdf?pdf=RDMToolkit
- Digital Science Report: The State of Open Data https://figshare.com/articles/The_State_of_Open_Data_Report/4036398

Organizations that offer guidance for funders on data-sharing policies:

- Open Research Funders Group (ORFG), see resources page.
- Center for Open Science (TOP Guidelines, also adapted here for funders).
- Expert Advisory Group on Data Access (EAGDA)
- Research Data Alliance (RDA)

References

Barbui, Corrado. (2016). “Sharing all types of clinical data and harmonizing journal standards” *BMC Medicine* 14:63. 10.1186/s12916-016-0612-8

Borgman, Christine. (2015). Big Data, Little Data, No Data: Scholarship in a Networked World. The MIT Press.

Borgman, Christine et al. (2016). “The durability and fragility of knowledge infrastructures: Lessons learned from astronomy.” *Proceedings of the Association for Information Science and Technology* 53:1. 10.1002/pra2.2016.14505301057

Clemens, M.A. (2015). “The meaning of failed replications: A review and proposal.” *J. Econ. Surv.*10.1111/joes.12139

Digital Curation Centre, “Overview of funders’ data policies.” http://www.dcc.ac.uk/resources/policy-and-legal/overview-funders-data-policies. Accessed May 2017.

Fane, B. et al. (2016). “Open Season for Open Data: A Survey of Researchers.” State of Open Data. figshare. https://doi.org/10.6084/m9.figshare.4036398.v1

Federer LM et al. (2015). “Biomedical Data Sharing and Reuse: Attitudes and Practices of Clinical and Scientific Research Staff.” PLoS ONE 10(6). https://doi.org/10.1371/journal.pone.0129506

Goodman A et al. (2014) Ten Simple Rules for the Care and Feeding of Scientific Data. PLoS Comput Biol 10(4): e1003542. https://doi.org/10.1371/journal.pcbi.1003542

Holdren, J. (2013). “Increasing Access to the Results of Federally Funded Research.” Office of Science and Technology Policy. http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf

Hunt, Kelly. (2016). “Robert Wood Johnson Foundation Grantee and Staff Member Perspectives on Open Access Publishing and Research.”

http://www.rwjf.org/content/dam/farm/reports/reports/2016/rwjf432351/subassets/rwjf432351_1

Ioannidis J.P.A. (2005). “Why Most Published Research Findings Are False.” PLoS Med 2(8): e124. doi:10.1371/journal.pmed.0020124

Institute of Medicine (2015). “Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk.” https://www.nap.edu/catalog/18998/sharing-clinical-trial-data-maximizing-benefits-minimizing-risk

Jardine, L. (1999). Ingenious Pursuits: Building the Scientific Revolution. New York: Doubleday.
Jones, Sarah. (2012) “Developments in Research Funder Data Policies.” International Journal of Digital Curation, 7(1). http://dx.doi.org/10.2218/ijdc.v7i1.219

LEARN (2016). “LEARN Toolkit of Best Practice for Research Data Management.” http://learn-rdm.eu/wp-content/uploads/RDMToolkit.pdf?pdf=RDMToolkit

Neylon C. (2017) “Compliance Culture or Culture Change? The role of funders in improving data management and sharing practice amongst researchers.” Research Ideas and Outcomes 3: e14673. https://doi.org/10.3897/rio.3.e14673

Nosek, B., et al. (2015). “Promoting an open research culture.” Science 348, 10.1126/science.aab2374.
OECD (2007). “OECD Principles and Guidelines for Access to Research Data from Public Funding.” https://www.oecd.org/sti/sci-tech/38500813.pdf

Office of Budget and Management (1999). “Circular A-110.” http://www.whitehouse.gov/omb/circulars_a110

Panguis, WG (2014). “A systematic review of barriers to data sharing in public health.” BMC Public Health 14. 10.1186/1471-2458-14-1144

Peer, L, A Green and E Stephenson, (2014). “Committing to Data Quality Review.” (2014). The International Journal of Digital Curation. http://dx.doi.org/10.2218/ijdc.v9i1.317

SPARC, “Data Sharing Requirements by Federal Agency.” (2016). http://sparcopen.org/wp-content/uploads/2016/05/SPARC-JHU-Digest-of-Federal-Data-Sharing-Requirements_v2_20160510.xlsx. Accessed May 2017.

Tenopir C, Dalton ED, Allard S, Frame M, Pjesivac I, Birch B, et al. (2015) “Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide.” PLoS ONE 10(8): e0134826. https://doi.org/10.1371/journal.pone.0134826

Tenopir, C. Allard, S., Douglass, K. (2011). “Data Sharing by Scientists: Practices and Perceptions,” PlosOne, 6(6).: e21101. doi:10.1371/journal.pone.0021101

Treadway, J., et al, (2016) “The State of Open Data Report.” figshare.https://doi.org/10.6084/m9.figshare.4036398.v1

Van den Eynden, Veerle et al. (2011). “Managing and Sharing Data: Best Practices for Researchers.” UK Data Archive. http://www.data-archive.ac.uk/media/2894/managingsharing.pdf

Van den Eynden, V et al. (2016). “Survey of Wellcome researchers and their attitudes to open research. figshare.” https://doi.org/10.6084/m9.figshare.4055448.v1

Vasilevsky NA, et al. (2017) “Reproducible and reusable research: are journal data sharing policies meeting the mark?” PeerJ 5:e3208https://doi.org/10.7717/peerj.3208

Vickers, Andrew. (2016) “Sharing raw data from clinical trials: what progress since we first asked ‘Whose data set is it anyway?’” Trials 17:227 10.1186/s13063-016-1369-2

Vines et al. (2013). “Mandated data archiving greatly improves access to research data.” Arxiv. http://arxiv.org/abs/1301.3744/ doi: 10.1096/fj.12-218164

Wallis J.C., Rolando, E., & Borgman, C.L. (2013). If we share data, will anyone use them? Data sharing and reuse in the long tail of science and technology. PLoS ONE 8(7), e67332. doi:10.1371/journal.pone.0067332

Wellcome Trust. “Sharing research data to improve public health: full joint statement by funders of health research.” http://www.wellcome.ac.uk/About-us/Policy/Spotlight-issues/Data-sharing/Public-health-and-epidemiology/WTDV030690.htm.”

Wellcome Trust, Fort Lauderdale agreement (2003). “Sharing Data from Large-Scale Biological Research Projects: A System of Tripartite Responsibility.” https://www.genome.gov/pages/research/wellcomereport0303.pdf

White, E.P., Baldridge, E., Brym, Z.T., Locey, K.J., McGlinn, D.J., & Supp, S.R. (2013). Nine simple ways to make it easier to (re)use your data. Ideas in Ecology and Evolution 6(2), 1–10. doi:10.4033/iee.2013.6b.6.f

Whitmire, Amanda et al (2015) "A table summarizing the Federal public access policies resulting from the US Office of Science and Technology Policy memorandum of February 2013." figshare. http://dx.doi.org/10.6084/m9.figshare.1372041. Accessed May 2017.

Whyte, A. (2015). "Where to keep research data: DCC checklist for evaluating data repositories" v.1.1 Edinburgh: Digital Curation Centre. Available online: www.dcc.ac.uk/resources/how-guides

Wynholds, Laura, et al. (2012). "Data, Data Use, and Scientific Inquiry: Two Case Studies of Data Practices." http://works.bepress.com/borgman/264/