Open data surveys: how comparable are they and their policy development applications

David O'Brien (dobrien@idrc.ca) & Daniel Gregoire (dgreg040@gmail.com)

Discussion Paper (v.1) for the RDA Interest Group 'Surveying Open Data Practices'

Comments welcome

Oct. 16, 2019

The RDA Interest Group (IG) 'Surveying Open Data Practices' seeks to promote coherence and coordination among sponsors and designers of open data sharing surveys. Our objective is to advance the understanding of current data sharing practices to inform the policies and initiatives shaping open data practices worldwide. As stated in the IG Charter, our specific objectives are to: 1) convene user communities, 2) develop a community-designed surveys and survey modules to meet specific needs or contexts, and 3) determine how such open survey(s) could be implemented and results analyzed globally.

Contents

1. Context 4

2. Characterizing Surveys 4

Survey overview 5

Table 1: Open data survey characteristics 6

Geographic Coverage 7

Figure 1: Distribution of survey responses by continent (10 surveys) 8

Figure 2: Heat-map distribution of compiled survey respondents by country 9
Differences in data sharing across countries 10
Table 2: Re-coding table for data sharing questions 11
Figure 3A Data sharing for top 10 countries by number of respondents from the 2016 State of Open Data Survey 14
Figure 3B Data sharing for top 10 countries by number of respondents from the 2018 State of Open Data Survey. 15
Figure 3C Data sharing for top 10 countries by number of respondents from the 2014 Wiley survey. 16
Figure 3D Data sharing for top 10 countries by number of respondents from the 2016 Wiley survey. 17
Disciplinary Coverage and Differences 18
Figure 4A: Data sharing across disciplines for the 2014 Wiley survey 19
Figure 4B: Data sharing across disciplines for the 2016 Wiley survey 20
Figure 4D: Data sharing across disciplines for the 2018 State of Open Data survey 22
3. A Design Framework for Open Data Surveys 23
Figure 5: Visualization of the IAD framework (Ostrom 2005:15) 25
Mapping Survey Questions 26
4. Use of Surveys 30
5. Conclusion 32
6. Bibliography 33
Appendix A: Geographic distribution of responses data tables and heat maps 37
Table A1: Total responses from each country for each survey used to generate heat maps 37
Figure A1: Heat-map distribution of survey respondents by country for the Belmont Forum survey conducted in 2016 41
Figure A2: Heat-map distribution of survey respondents by country for the Elsevier& CWTS survey conducted in 2017 42
Figure A3: Heat-map distribution of survey respondents by country for the SpringerNature survey conducted in 2018 43
Figure A4: Heat-map distribution of survey respondents by country for State of Open Data survey conducted in 2016 44

Figure A5: Heat-map distribution of survey respondents by country for the State of Open Data survey conducted in 2017 45

Figure A7: Heat-map distribution of survey respondents by country for the survey conducted by Tenopir et al. in 2009-2010 47

Figure A8: Heat-map distribution of survey respondents by country for the survey conducted by Tenopir et al. in 2014-2015 48

Figure A10: Heat-map distribution of survey respondents by country for the Wiley survey conducted in 2016 50

Example code to obtain number of responses per country and produce heat map 51

A Methodological Note on Interoperability for Appendices B and C 54

Appendix B: Country bar graphs 55

Figure B1: Data sharing across disciplines and through time for Elsevier CWTS survey conducted in 2017 55

Figure B2: Data sharing across disciplines and through time for the SpringerNature survey conducted in 2018 56

Figure B3: Data sharing across disciplines and through time for State of Open Data survey conducted in 2017 57

Example code for country bar graphs 59

Appendix C: Discipline bar graphs 62

Figure C1: Data sharing across disciplines and through time for Elsevier CWTS survey conducted in 2017 62

Figure C2: Data sharing across disciplines and through time for SpringerNature survey conducted in 2018 63

Figure C3: Data sharing across disciplines and through time for State of Open Data survey conducted in 2017 64

Figure C4: Data sharing across disciplines and through time for survey conducted by Tenopir et al. in 2013-2014 65

Example code for discipline bar graphs 66

1. Context

We are in a period of experimentation as government agencies, research funders, research performing organizations, professional societies and publishers introduce open data policies and practices to promote research data sharing. While these parties are united in their goal of making research data open, differences in the incentives, institutional settings and infrastructure capabilities are some of the factors that influence the pace of change (Fecher, Friesike, Hebing 2015).

As open data policies are implemented and as the data sharing norms of research communities evolve, identifying and tracking research data practices has garnered increased interest from the research community at large. This interest is evidenced by the emergence of national and international surveys benchmarking attitudes towards data sharing practices in recent years.

These surveys have revealed considerable variability across disciplines and countries in researcher perceptions and practices relating to open research data. For researchers studying open science and agencies interested in supporting open data, such findings identify gaps between policy intent and actual researcher practices.

We are interested in these findings and the policy questions they raise. We are, however, limited in our ability to compare findings from different surveys. One example of a limiting factor is that surveys we examined ask similar questions but few ask identical question. As such, the findings from one survey do not complement findings from another in a meaningful way.

Bearing these issues in mind, we undertook a comparative analysis of survey instruments. Our approach mirrors similar efforts to compare open data policies (SPARC and DCC 2017, 2018) and infrastructure (Braunschweig 2012). To our knowledge, a comparison of survey findings and the underlying survey instruments has not been undertaken. In the second section of this paper, we introduce ten open data surveys and highlights design differences and select findings. The reported findings point to questions such as what explains different rates of open data adoption across countries, disciplines and changes over time? In section three, we explore how differences in survey designs inform whether results can be used for descriptive or explanatory purposes. We also propose an analytical framework to guide future survey design and improve the comparability of survey instruments. In section four, we illustrate several case-studies demonstrating how survey sponsors have used findings to inform policy development. We conclude by discussing how surveys may be designed to serve the specific needs of sponsors operating in different research contexts and advocate that future efforts promote interoperability.

2. Characterizing Surveys

In this section, we introduce ten global surveys and highlight some of their general characteristics. We discuss their comparability in terms of survey design and the different ways they assess data sharing practices. We also illustrate the geographic distribution of respondents

and disciplinary data sharing practices. This section highlights why these surveys can be challenging to compare. In the section three, we draw on this analysis to identify and categorize questions that could be used to operationalize an analytical framework.

Our sample of global surveys excludes a growing number of national-level surveys, largely commissioned by government agencies. Japan (Allagnat et al 2018), Austria (Bauer et al 2015), Denmark (Danish National Research Foundation 2017), Finland (Enwald, Kortelainen, & Huotari 2017), and the UK (Wolff-Eisenber, Rod & Schonfeld 2016) are recent examples. We also excluded surveys from research performing agencies or research networks in countries like Egypt, Jordan and Saudi Arabia (Elsayed & Saleh, 2018) and South Africa and Kenya (Bezuidenhout & Chakauya 2017).

Survey overview

The surveys included in our sample are presented in **Table 1** and links to the associated questionnaires and datasets are provided in the bibliography. As column 1 indicates, the survey sponsors are mainly scientific publishers and non-governmental organizations (sometimes in collaboration). The survey conducted by Tenopir et al. is an exception as the authors are academics.

Column 3 briefly summarizes the focus of each survey. Only surveys focussing on open data were included in the sample. There are, however, differences in emphasis within this sample. For example, the State of Open Data reports and the survey conducted by CWTS & Elsevier probe researchers' attitudes toward data sharing. The Wiley surveys highlight barriers and types of data that are more likely to be shared. The SpringerNature survey was designed to answer where data sharing occurs, what differences exist at a broad disciplinary level, and asks what services might alter researcher practices. Two of the survey sponsors seek to track changes in researcher's perception and data sharing practices over time (State of Open Data and Wiley).

Columns 4 and 5 demonstrate the recent interest in the topic. The first of the two Tenopir surveys (2015) was the earliest in the sample and the remainder were published in 2014 and after.

Column 6 indicates the recruitment method. The publishers relied on authors listed in their journals. The Belmont Forum and Tenopir surveys relied on listservs and a snowball method of forwarding invitations. The number of invitations sent (where available) and survey respondents varied by a factor of five between surveys. Most surveys generated between 1000 to 2000 responses and the response rate varied between 2-10% where data was available. Notably, the surveys undertaken by Wiley in 2014 and SpringerNature in 2017 had more than double the number of respondents, which may be due in part to the relatively smaller number of questions compared to the more detailed surveys (column 9).

Table 1: Open data survey characteristics

Survey sponsor (abbreviation) [data set citation]	**Title**	**Topics of focus**	**Survey year**	**Publication year**	**Recruitment method**	**# invitations**	**# responses**	**# questions**
Belmont Forum (Belmont) [Schmidt et al 2015]	The Belmont Forum's Open Data Survey	Identify key open data activities, best practices from a user perspective, barriers and incentives for data sharing	2014	2015	Distributed to open access authors Copernicus Publications and forwarded on by first recipients	~29,000	1330[a]	19
CWTS & Elsevier (CWTS & Elsevier) [Berghmans et al. 2017]	Open Data: The Researcher Perspective	Attitudes and behavior of researchers with regard to sharing their research data and using open data in their own research	2016	2017	Researchers that published an article or book (chapter) indexed in Scopus (2012-2015)	50,521[b]	1162	41
SpringerNature, FigShare & Digital Science (SOD 2016) [Nature Research 2016]	Open Data survey	Researcher attitudes and experiences in working with open data	2016	2016	NA	NA	2061	58
FigShare, SpringerNature, Wiley & Digital Science (SOD 2017) [Nature Research et al 2017]	State of Open Data survey 2017	Changes over time researcher attitudes and experience working with open data	2017	2017	NA	NA	2351	95
FigShare, SpringerNature, Wiley, & Digital Science (SOD 2018) [Nature Research 2018]	State of Open Data survey 2018	Changes over time researcher attitudes and experience working with open data	2018	2018	NA	NA	1872	56
SpringerNature (SpringerNature) [Astell et al 2018]	Practical Challenges for Researchers in Data Sharing	Data sharing challenges during publishing and recommendations for support	2017	2018	Registrants to nature.com, biomedcentral.com and springer.com	~249,000	7718	18
Academic / Independent (Tenopir) [Tenopir et al 2015a]	Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide	Data sharing and reuse perceptions and practices among researchers	2009/2010 and 2013/ 2014	2015	Email distributed by DataONE to networks and environmental science listservs and blogs	NA	1329 (2009/2010) 1015 (2013/2014)	76
Wiley (Wiley 2014) [Wiley 2016]	Wiley Data Sharing Survey	Understand how and why researchers make their research data publicly available	2014	2016	Researchers linked to Wiley's journal portfolio	~50,000	2558	22
Wiley (Wiley 2016) [Wiley 2017]	Wiley Open Science Researcher Survey 2016	Understand how and why researchers make their research data publicly available	2016	2017	Researchers linked to Wiley's journal portfolio	~55,000	4668	34

[a] Reported number but there were 1298 responses to the country of origin question

[b] Estimate based on reported response rate

The next section explores the surveys in further depth by reporting their geographic coverage and some of their main findings.

Geographic Coverage

Using our sample, we calculated the representation of researchers by continent and countries. We assessed this distribution by counting the frequency a country was identified by respondents, whether they completed the survey or not. These data tables can be found in **Appendix A** and were used to generate continental distribution and country-level heat-maps of each survey.[1]

Figure 1 shows the continental distribution of the 24,652 responses to the ten surveys. At the continental scale, European respondents were the most numerous followed by respondents in North America and Asia. Outside these regions, the number of responses declines considerably. Africa, for example, accounts for 2.6% of the responses. While the proportion of responses is roughly aligned to global estimates of the distribution of researchers by continent (UNESCO 2015, Figure 1.3) many countries in Asia, Latin America and Africa are under-represented.

[1] The heatmaps were generated using code from the rworldmap package in R (see https://journal.r-project.org/archive/2011-1/RJournal_2011-1_South.pdf and Appendix A for example code).

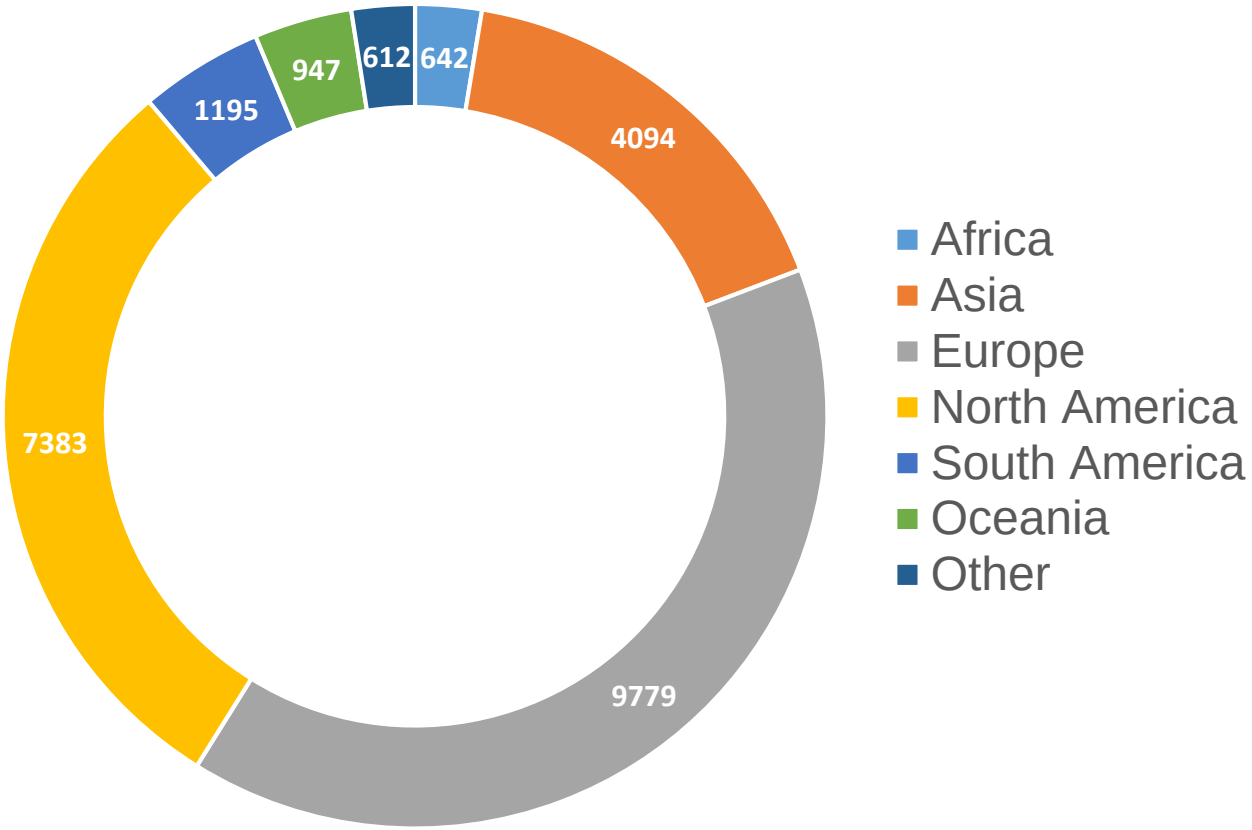

Figure 1: Distribution of survey responses by continent (10 surveys)

The United States has the highest number of respondents compiled across surveys at 6122 (see **Figure 2** and **Appendix A**) and accounted for almost the entirety of the North American dataset. In Asia, the most populous countries of India and China dominate, although these countries still had fewer respondents than in many European countries. European representation was led by Western Europe, specifically the United Kingdom, Italy, Spain, and Germany (see **Figure 2**) although the coverage did vary considerably between surveys (see **Appendix A**). As illustrated, some surveys had broad geographic coverage (e.g., Wiley 2016 see **Figure A10**) whereas as others had a much more limited geographical reach (e.g., Tenopir et al. see **Figure A7**).

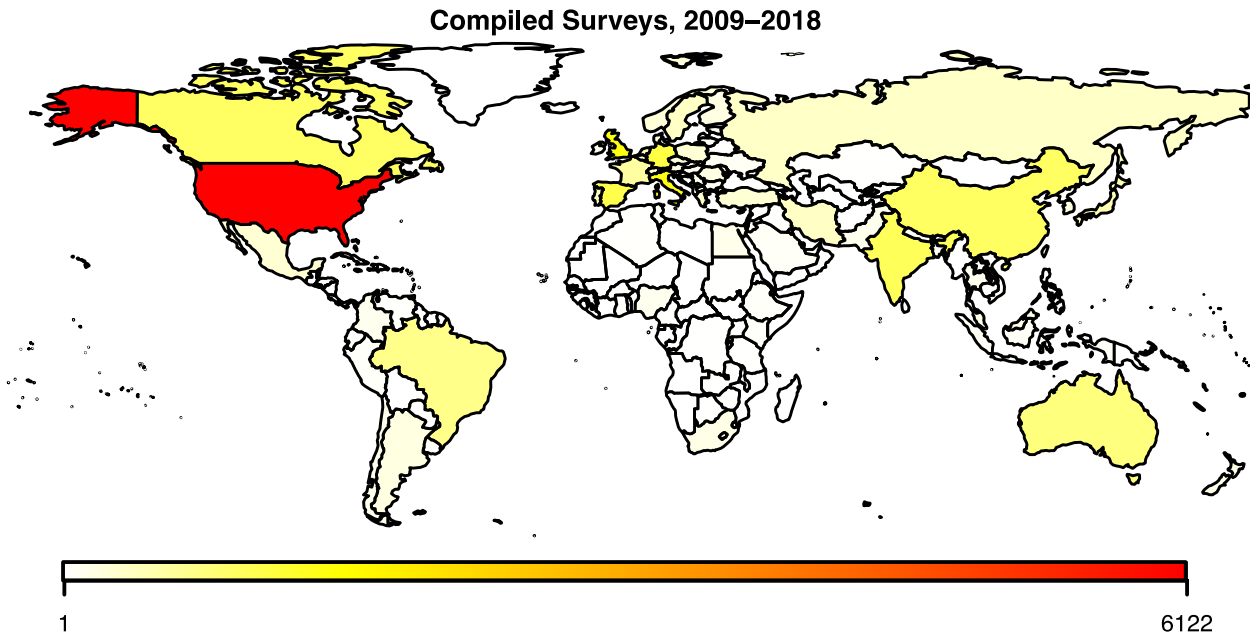

Figure 2: Heat-map distribution of compiled survey respondents by country

In the remaining continents, there are some countries with high representations, such as Brazil with 703 responses, though most countries had less than 50 responses (see below and **Table A1**).

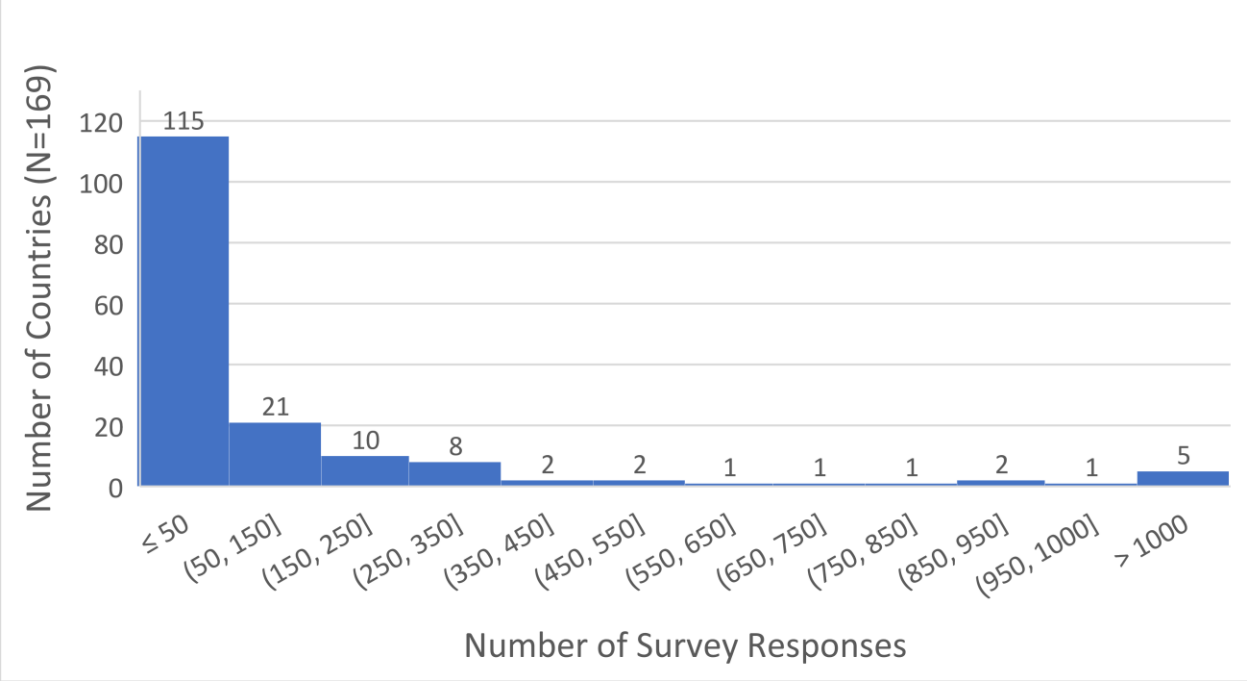

Figure 2a: Histogram of Responses by Country for Sample Surveys

Future efforts to increase country representation should be encouraged and this may require moving away from recruitment avenues linked to major scientific publishers to contacting research institutions or national funding organizations directly.

Differences in data sharing across countries

For countries with numerous responses, what does the data say about their data sharing practices and do countries exhibit different data sharing rates than others? In answering this question, we illustrate an impediment to comparing and analysing survey findings globally.

As discussed, all the surveys are interested in whether researchers share their data, but they all ask this question differently. Wiley (2014) provides a straight-forward question: “Have you ever shared your data publicly?”, yes or no. As shown in **Table 2**, the remaining surveys ask frequency or quantity-based questions such as “how often you share?” or “how much of your data do you share?” to determine whether a researcher shares their data or not.

Such nuances add a dimension to our understanding of data sharing practices but these different framings make comparison between surveys difficult. In an effort to compare the findings from different surveys, we re-coded relevant survey questions as illustrated in the right-hand column to create a binary distinction.

Table 2: Re-coding table for data sharing questions

Survey (abbreviation)	Question number	Question wording	Answer options	Code
Belmont Forum	16	What types of systems/archives do you use to publish your data?	Free text	NA
CWTS & Elsevier	2g	Which of the following locations do you use to archive your research data?	I don't archive	No
			Repository provided by funder	Yes
			Repository provided by publisher	
			Repository provided by institute	
			Repository provided by department	
			Other	
SOD 2016	28	How often have you made your research data free to access, reuse, repurpose or redistribute?	Never	No
			Rarely	Yes
			Sometimes	
			Frequently	
SOD 2017	2.2	How often have you made your research data openly available?	Never	No
			Rarely	Yes
			Sometimes	
			Frequently	
SOD 2018	2.2	How often have you made your research data openly available?	Never	No
			Rarely	Yes
			Sometimes	
			Frequently	
SpringerNature	Fig 2 title	Generally, when submitting a manuscript to a journal what do you do with the data files generated by your research?	Neither	No
			Deposit files in a repository	Yes
			Submit files as supplementary information	
			Both	
Tenopir et al.	13	How much of your data do you make available to others?	None	No
			Some	Yes
			Most	
			All	
Wiley 2014	14	Have you ever made your data publicly available?	No	No
			Yes	Yes
Wiley 2016	21	Where do you make your data publicly available?	I have not made my data publicly available	No
			As supplementary material in a journal	Yes
			Institutional data repository (i.e. university or institute-sponsored)	
			General-purpose data repository (e.g. Dryad, figshare)	
			Discipline-specific data repository (e.g. GenBank, OpenEI, Protein Data Bank, TreeBASE)	
			Personal, institutional or project webpage	
			Informal paths or upon request (email, direct contact etc)	
			At a conference	
			Other	

Figure 3 illustrates the number of researchers per country who share and do not share their research according to the re-coding conventions shown in **Table 2**. The data are from the State of Open Data and the Wiley surveys. These two surveys are illustrated here as they have reasonable distributions across a number of countries and offer the opportunity to examine whether data sharing practices have changed over time.

Figures 3A to D profile the top 10 countries by number of responses to both surveys. When comparing datasets from the 2016 State of Open Data and the Wiley surveys, we see that the countries that make up the top 10 are quite similar. Unique entries are Germany and Canada in the State of Open Data survey, and Australia and Iran for the Wiley survey.

The dominant trend is that the 'yes' responses outnumber the 'no' responses. The ratios do fluctuate between countries and these differences raise policy and practice questions as to why such differences exist. Two main differences are discussed in further detail: between countries in the same survey; and between the same country over time.

Country differences between surveys conducted the same year

In the top 10 countries for the 2016 State of Open Data survey, the percentage of respondents who indicated 'yes' they shared data relative to the total number of responses varied between 63.5 and 85.7%. Spain and Italy have the highest proportion of data sharers with 85.7 and 85%, respectively, whereas Canada has the lowest proportion with 63.5% (Figure 3A).

The Wiley 2016 survey had similar findings. The percentage of respondents who indicated 'yes' they shared data relative to the total number of responses varied between 64.1 and 84.6 %. In this survey, the United Kingdom had the highest proportion of data sharers with 84.6 % whereas the United States and India had the lowest proportion of respondents who shared data with 65.4% and 64.1%, respectively (Figure 3D).

Country differences between the same surveys conducted across multiple years

The State of Open Data and the Wiley surveys allow us to see how data sharing practices have changed over a two-year period. As reported above, the percentage of researchers in the top 10 countries who responded 'yes' they shared research data in the 2016 State of Open Data survey ranged from a low of 63.5 to a high of 85.7%. Two years later in 2018, this range in the top 10 countries declined slightly to a low of 60.5 to a high of 83.8% of researchers who share research data.

By contrast, the Wiley surveys show an increase in data sharing by roughly 10%. In 2014, the percentage of researchers in the top 10 countries who responded 'yes' they shared research data ranged from a low of 55 to a high of 72.4%. Two years later, this range increased to a low of 64.1 to a high of 84.6%.

These overall shifts reflect country-level changes. Comparing the 2016 and 2018 State of Open Data surveys, some countries experience slight declines in data sharing whereas others are

quite sharp. Illustrating the former, respondents in Spain had a data sharing rate of 85.7% in 2016 and this declined to 82.2% in 2018. Similarly, respondents in Canada had a data sharing rate of 63.5% in 2016 and this declined to 60.5% in 2018. These are modest declines. Researchers in Italy on the other hand, reported an 85% data sharing rate in 2016 and this declined sharply to 71.7% in 2018.

The differences between countries and changes within a country over time illustrated above and in the Figures below, raise interesting questions for policy and research. Why do researchers in one country have higher data sharing rates than in others? Can this be explained by changes in the policy environment or might it have to do with the sample population or how the question was asked? Tracking changes at the country level over time offers the potential to measure the effect of interventions designed to shape open data practices.

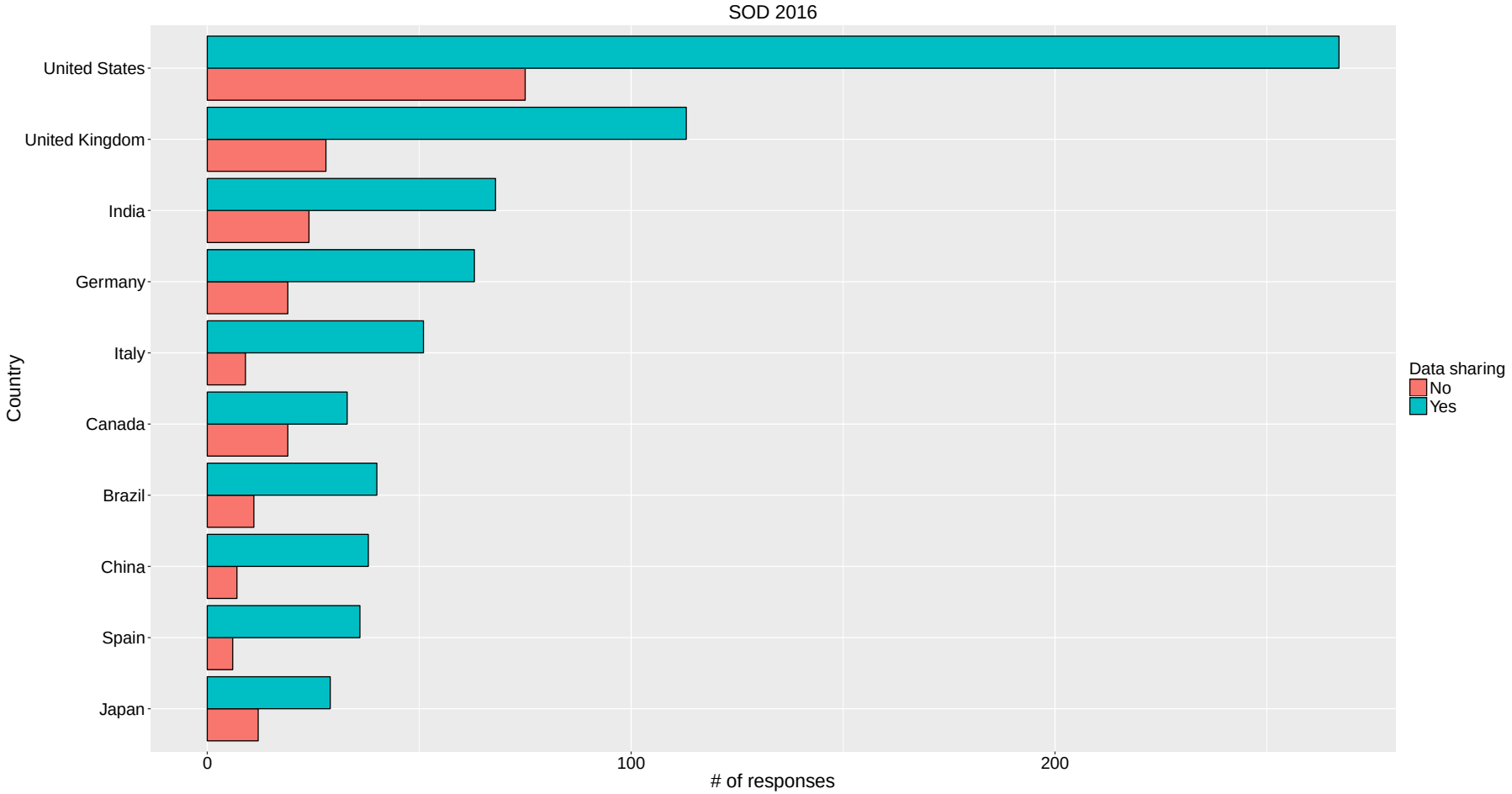

Figure 3A Data sharing for top 10 countries by number of respondents from the 2016 State of Open Data Survey

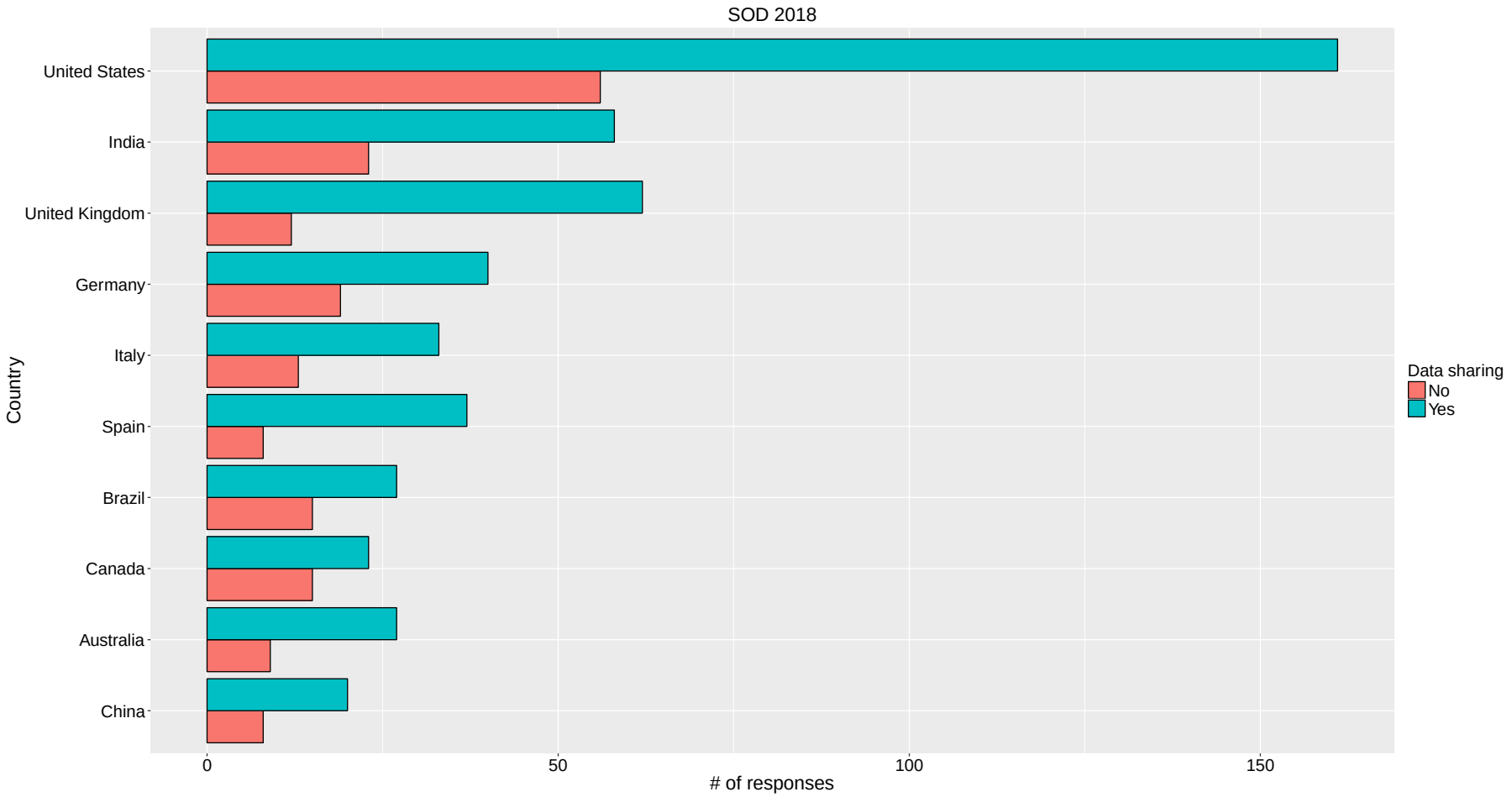

Figure 3B Data sharing for top 10 countries by number of respondents from the 2018 State of Open Data Survey.

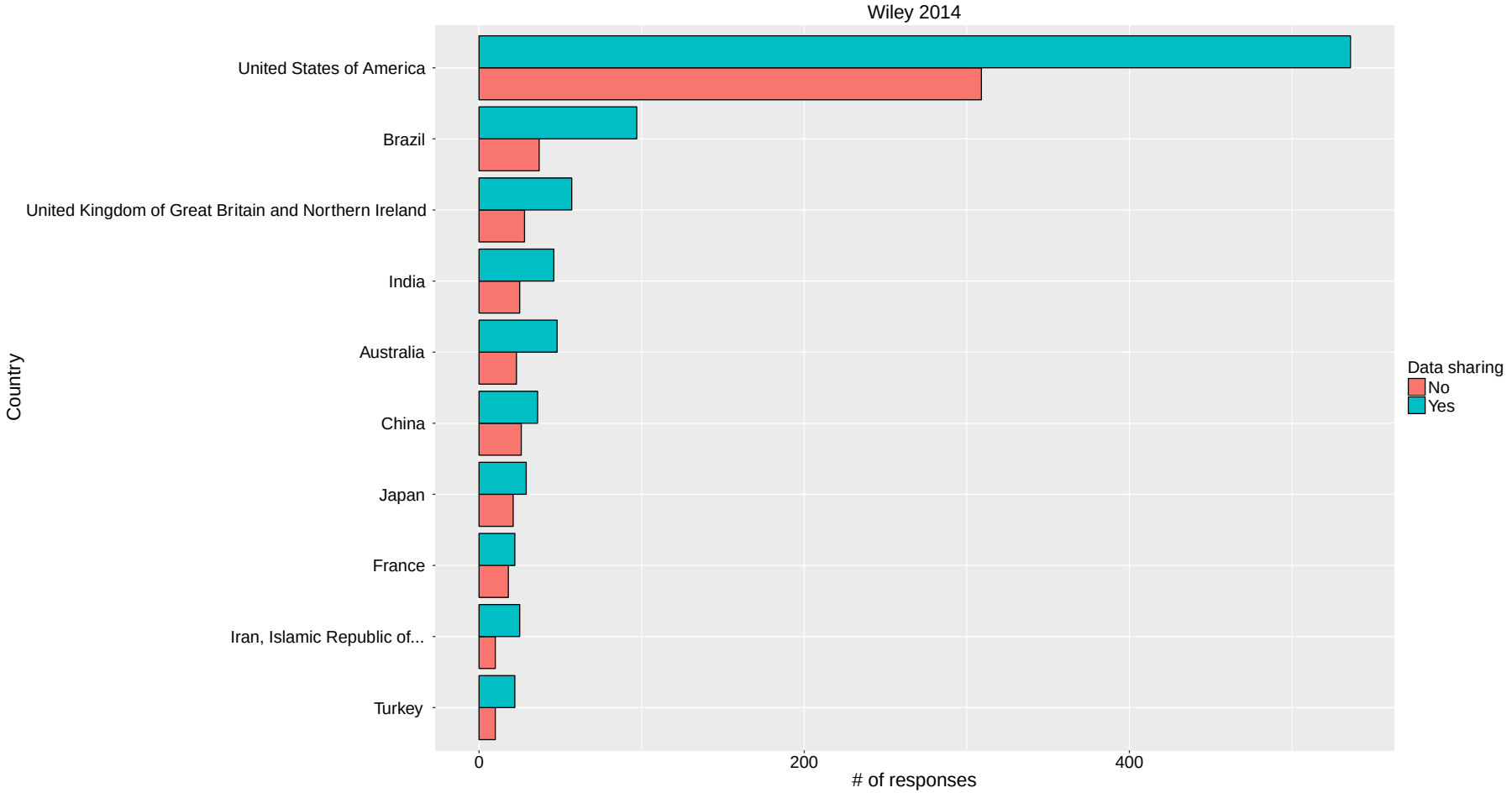

Figure 3C Data sharing for top 10 countries by number of respondents from the 2014 Wiley survey.

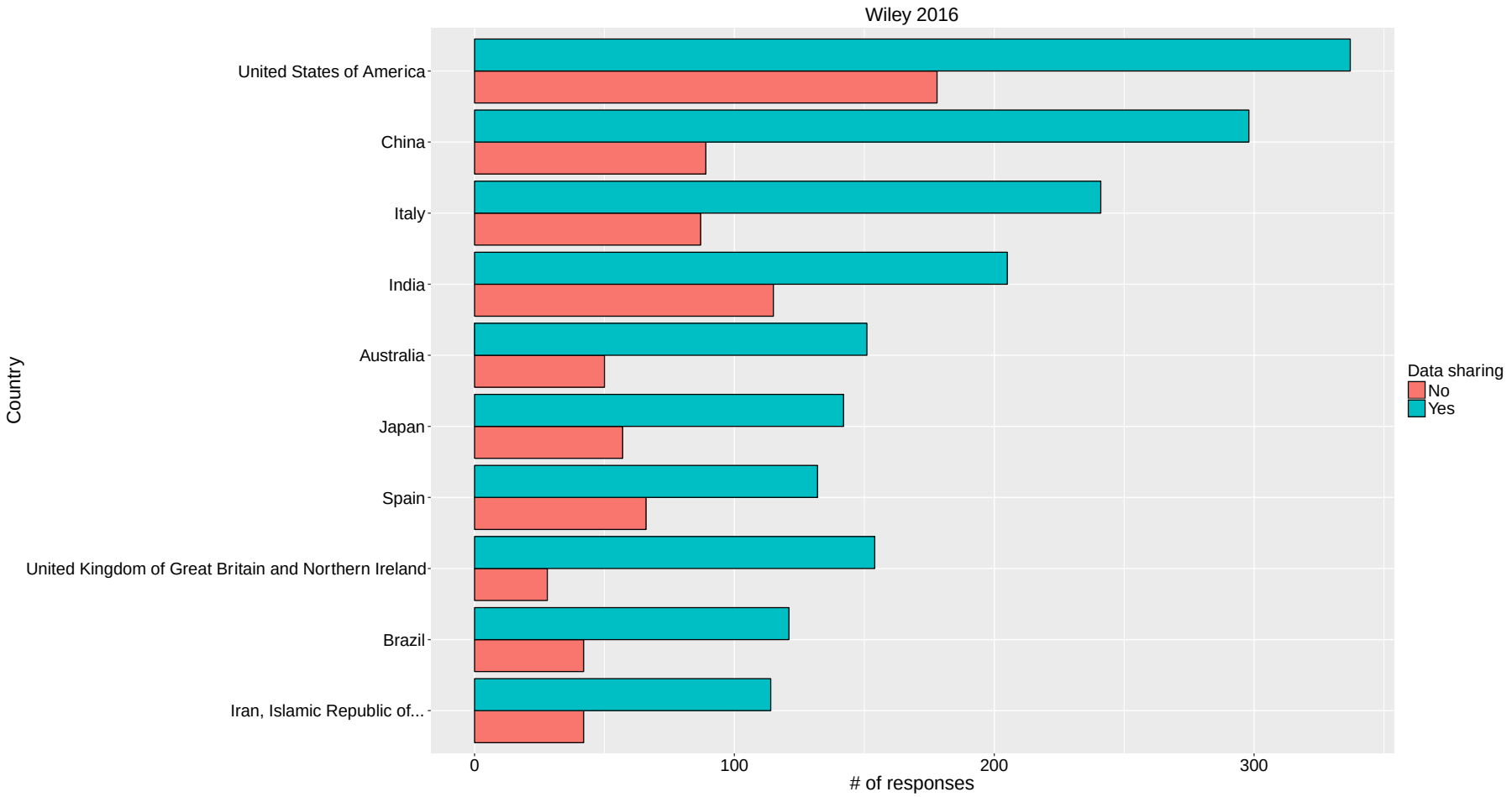

Figure 3D Data sharing for top 10 countries by number of respondents from the 2016 Wiley survey.

Disciplinary Coverage and Differences

Figures 4A-D illustrate the number of researchers by academic disciplines who share and do not share their research using the re-coding conventions in **Table 2**. Data sharing trends by disciplinary field is analysed in all the survey reports. Our rationale for replicating this analysis is two-fold. First, how comparable are the questionnaires and second, what are the differences between disciplines and over time?

For this analysis, we encountered the same challenge of comparability. All the surveys in our sample ask respondents what their disciplinary field is but they all provide different options. To illustrate this point, the SpringerNature (2017) survey lists five broad disciplinary fields whereas the Wiley and Elsevier surveys list over 20 disciplinary options.

Figures 4A & B report findings from the Wiley surveys (2014, 2016) and **Figures C & D** for the State of Open Data surveys (2016, 2018) with unaltered discipline fields. The State of Open Data lists 12 options for disciplines and the Wiley survey lists 24 disciplines. Results for the other surveys can be found in **Appendix C**.

The dominant trend is that 'yes' responses outnumber the 'no' responses in most but not all of the disciplinary fields. Again, these differences raise policy and practice questions as to why these differences exist.

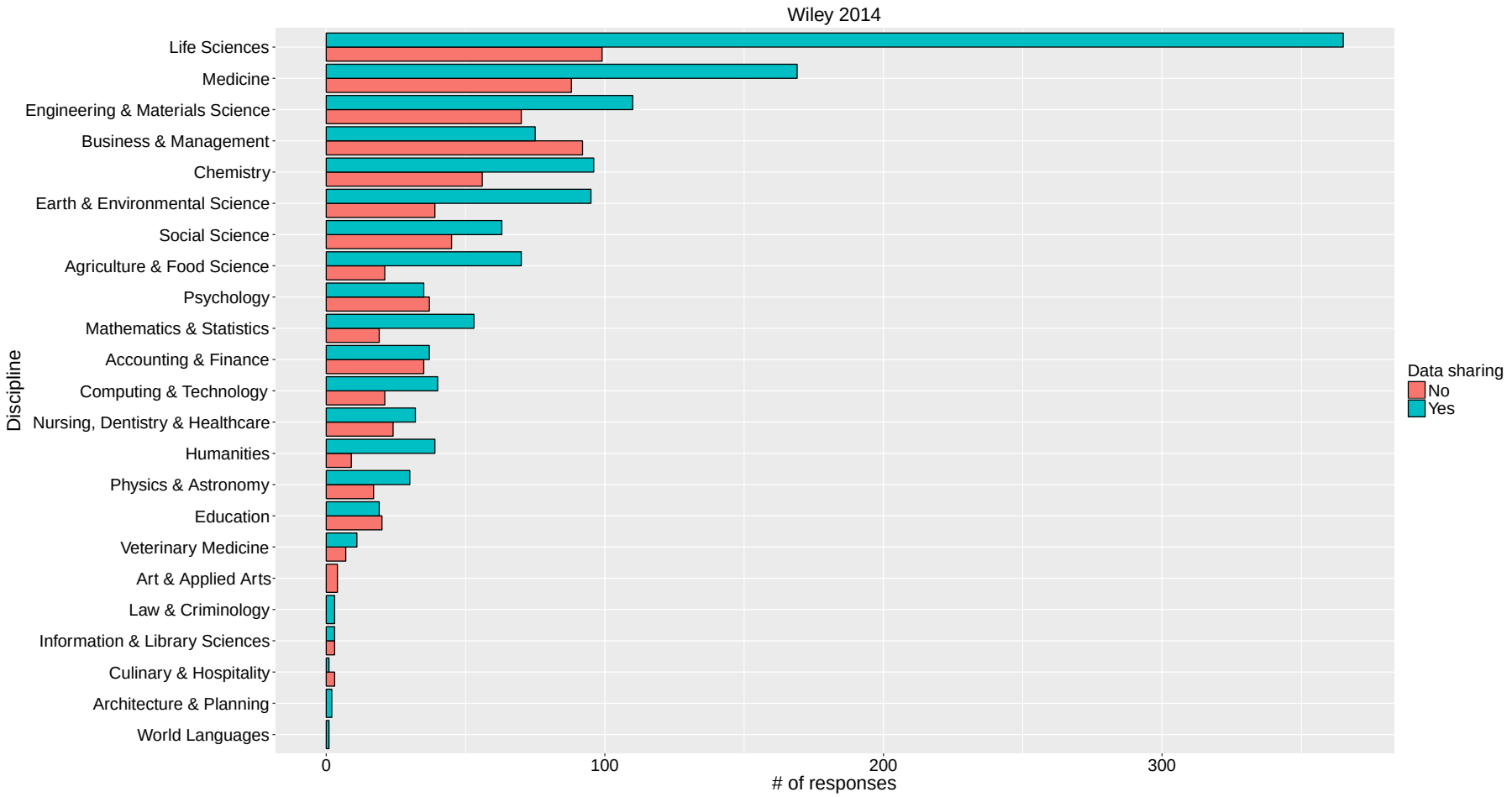

Figure 4A: Data sharing across disciplines for the 2014 Wiley survey

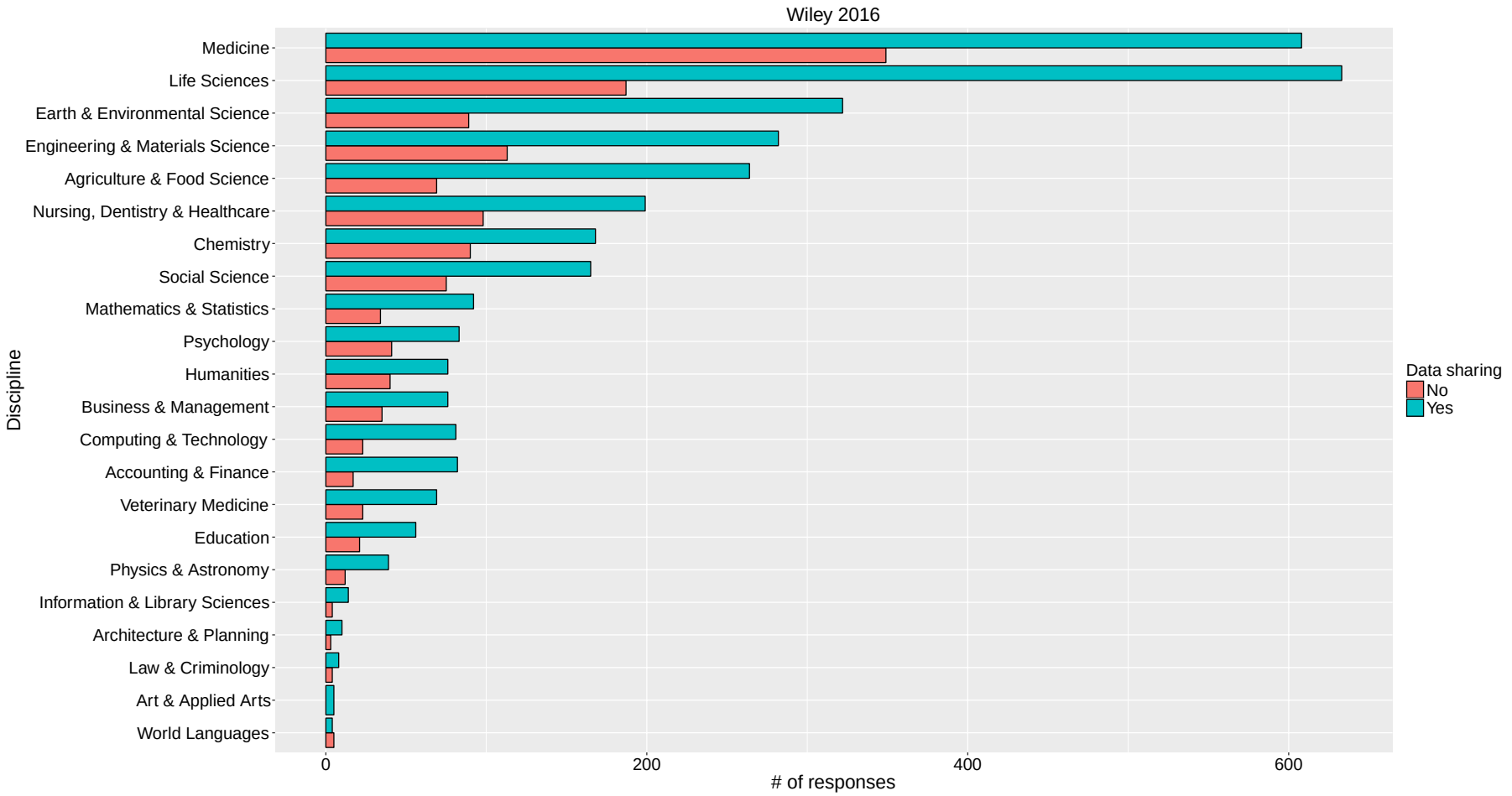

Figure 4B: Data sharing across disciplines for the 2016 Wiley survey

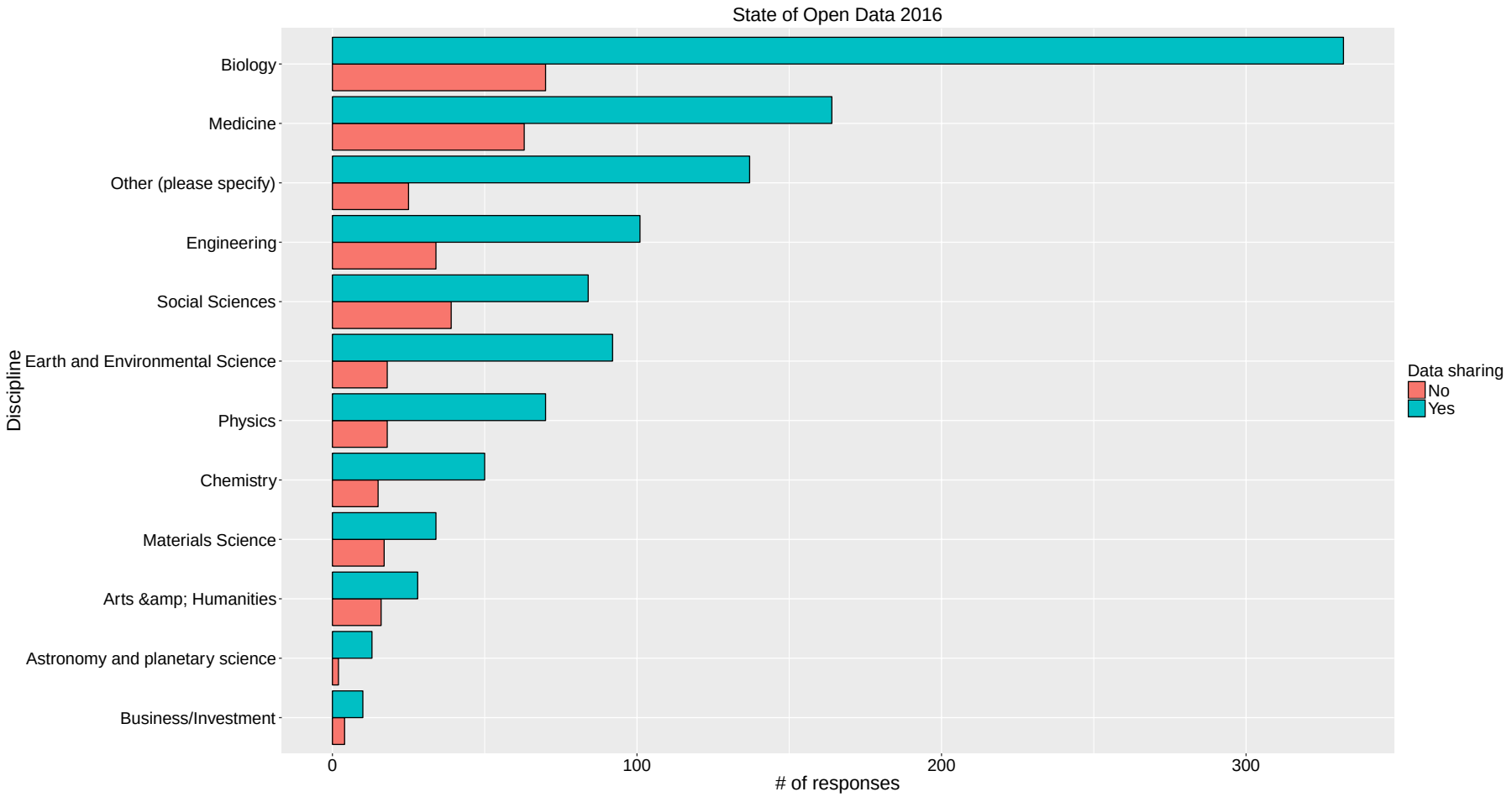

Figure 4C: Data sharing across disciplines for 2016 State of Open Data survey

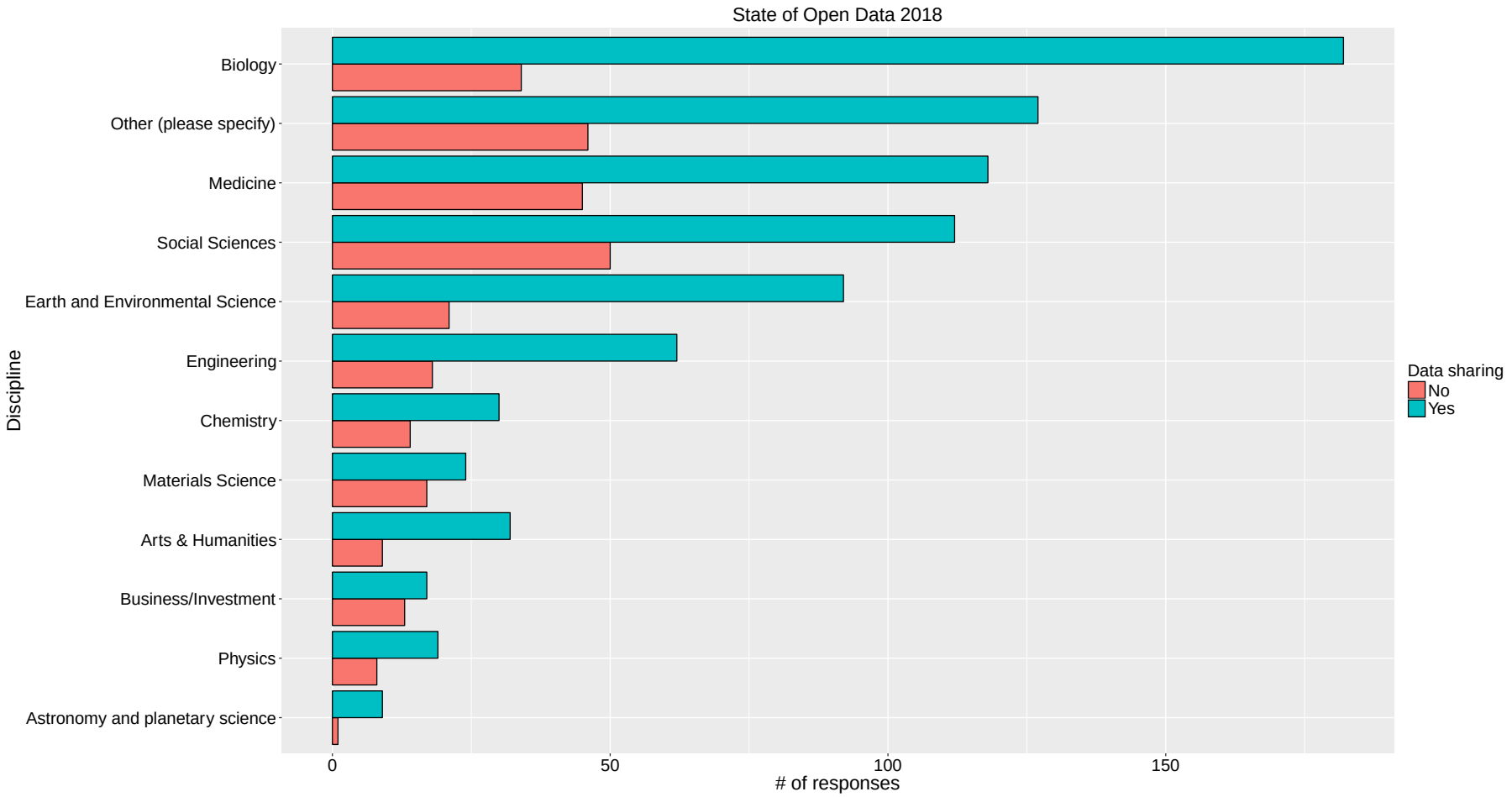

Figure 4D: Data sharing across disciplines for the 2018 State of Open Data survey

Differences between disciplines in the same survey

Overall, the SOD surveys show respondents from all discipline share their data more than they do not. This consistent picture is not reflected in the Wiley survey (2014) where several disciplines (e.g., psychology, education, business and management) have a higher proportion of 'no' data sharing responses (see **Figure 4A**). Similarly, the Elsevier survey reported the following fields with a higher proportion of 'no' data sharing responses: psychology, social science, economics, neuroscience, and nursing (see Appendix C, Figure C1). The "Other Science" category from the SpringerNature survey also reported a higher proportion of 'no' responses (see Appendix C, Figure C2).

Data-intensive disciplines tend to have the highest percent of 'yes' responses. These fields include biology, life science, medicine, and engineering. In some cases, such as the life sciences, the ratio of no vs yes was on the order of 25%, suggesting this discipline leads in data sharing whereas the ratio for medicine is closer to 50% (**Figure 4C & D**).

Differences between the same discipline over time

The SOD surveys are clearer to trace disciplinary trends as the re-coding scheme was applied to the same question in the 2016 and 2018 surveys. When comparing the number of respondents who shared data in 2016 compared to 2018, we found a slight decline in data-sharing. We suspect our re-coding of the data sharing variable and sample (top 10 countries) influenced this finding as the authors of the SOD 2018 report identified a 7% increase in data sharing over the two-year period (SOD, 2018).

The differences in data sharing across countries, disciplines and over time begs the question: why are data sharing outcomes different across countries and academic fields? None of the surveys in our sample purport to explain these differences. That said, surveys are commonly used and designed to identify correlations and causal pathways to help answer these questions. In the following section, we outline an analytical framework that has been applied to a range of contexts to explain and predict how different governance arrangements enable individuals to overcome collective action problems (e.g., public goods and common-pool resources).

3. A Design Framework for Open Data Surveys

The surveys in our sample identify phenomena of interest, collect relevant data and report descriptive statistics. We infer that survey designers have two motivations. One motivation is to describe the perceptions and practices of a given population at a point in time, compare differences across populations and/or solicit feedback from researchers on policies or practices researchers would value. A second motivation is to understand the changing practices of researchers. Several survey sponsors in our sample have reissued their surveys to track changes over time (e.g., Wiley 2014, 2016).

Within our sample, the survey designers are silent whether their surveys are informed by a deductive model of behavioural change. In less formal terms, they do not outline a theory of change that provides a rationale for the selection of variables and how those variables connect actions to outcomes. Had this approach been articulated, we could assume a third motivation is present – a desire to explain or predict changes in data sharing practices.

As Fecher, Friesike and Hebing (2015) report, there are many potential explanations for why some researchers share and others do not, yet there are only emergent efforts to situate such variables in analytical framework and explore their causal relationships. For survey data to inform policy or support data sharing initiatives, it is highly desirable to understand these relationships.

In this section, we introduce an analytical framework and map existing survey questions to the component parts of this framework. This mapping exercise demonstrates how questions within our survey sample are compatible with this analytical framework and how the component parts are logically connected.

Our starting point is Elinor Ostrom's (2005) Institutional Analysis and Development (IAD) framework, though we recognize there are several frameworks that could be utilized for such purposes (see Ostrom 2009 for a discussion of related models). Simply stated, the IAD framework posits that if we are to explain observed outcomes or account for changes in and differences across settings, we need to understand the context, identify physical and invisible institutions (e.g. legal rules and social norms) and examine how actors' interactions / learning influence their choices.

The IAD framework has been used as an analytical tool across a wide range of contexts to diagnose public policy outcomes or outcomes of social cooperation or conflict, and the consequences of those policies or choices (Ostrom 2009, Blomquist and deLeon 2011). The framework emerged from efforts to explain the influence and development of institutions that enable (or hinder) collective action problems relating to common-pool resources and public goods. Open research data is a public good. Once data is made available in the public domain, users have unlimited access and using the data does reduce its availability to others.

When the IAD framework is applied to relatively controlled situations, variables can be formally modelled and empirically tested. The utility of formal modelling and the ability make causal inferences declines as the complexity of the situation increases. As researchers employing the IAD framework move from controlled social experiments to large-scale observational studies where individual or group decisions/actions are influenced by a wide range of factors or it is difficult to control for explanatory variables, quantitative methods tend to give way to qualitative methods of analysis.

Figure 5 is a visual presentation of the main categories and their connections.

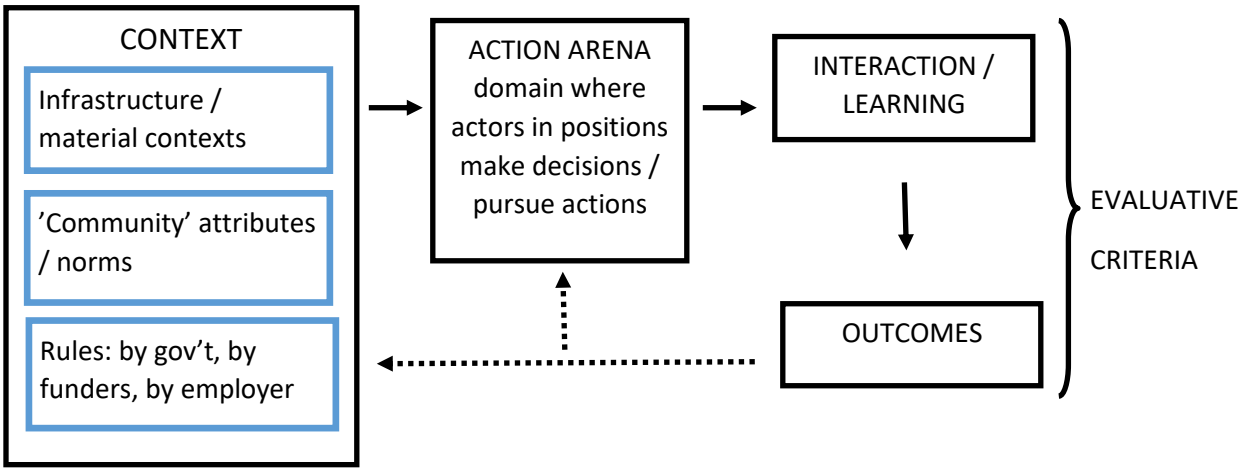

Figure 5: Visualization of the IAD framework (Ostrom 2005:15)

The IAD components are logically linked. At the centre of the framework is the 'action arena'. This is the immediate environment where actors take decisions in light of information available to them, the control they exert and the outcomes they seek.

The action arena is nested in a broader context shaped, for example, by available research infrastructure, norms of research communities and rules by government or funding agencies. This context influences the decisions actors make at any one point in time.

Formal rules encompass laws and regulations as well as contractual obligations which funders, publishers or employers may create relating to the management and sharing of research data. Formal rules may be multi-tiered: a national privacy or public data law, a funder's contractual obligation to archive data, and an employer's guideline on data sharing could all influence a researcher's decision to share or not share. By contrast, informal rules are not codified but nonetheless create norms that shape behaviour. The attributes of a researcher's peer group or scholarly community may value and routinely publish their research data. This would likely have a strong influence on an individual's own choices.

Both formal and informal institutions are socially constructed but they have important differences. Social norms tend to be routinely reproduced and may be slow to change, whereas formal rules are introduced at points in time and compliance with them depends on their legitimacy, compatibility with social norms and compliance mechanisms (Ostrom 1990, 2005, Jepperson 1991).

The final category in the 'context' box is the physical environment. For our purposes, infrastructure (data storage capacity), presence/absence of data archives, internet access and associated data costs, for example, would all be relevant variables.

The right side of IAD framework identifies the outcomes and the criteria by which an outcome or possible outcomes are assessed. Outcomes of primary interest in the surveys relate to researchers' actual data management and sharing practices.

When the IAD framework is employed, the evaluative criteria associated with possible outcomes are typically known. If increasing access to data is a desired outcome, this evaluative criterion would inform the selection of outcome indicators and place an emphasis on understanding how the context, the action arena and how learning and interactions shape the decisions of researchers and organizations contribute to achieving this outcome. Criteria such as sustainability, efficiency or accountability could be applied to the governance of open data.

The interaction and learning component of the IAD framework focusses attention on how changing conditions or feedback loops influence decisions and the context over time. A researcher may have a positive or negative experience when publishing their data and feedback from that experience may affect their willingness to share future datasets. Alternatively, a newly introduced data curation support program or data sharing policy may alter an individual's or an entire research community's interest or ability to publish their data. Surveys that capture panel or time-series data could be used to assess the impact of these changing conditions (interventions) or feedback loops over time.

Mapping Survey Questions

Having outlined the framework, we will now illustrate how survey questions align with the IAD components. The IAD framework has been criticized for being difficult to operationalize and standardize, with ensuing efforts by proponents to provide clarity (McGinnis 2016). The guidance Ostrom provides to operationalize the IAD framework tends to assume the analyst observes a process, relationships and chains of events. By contrast, a survey designer captures numerous perspectives (sometimes thousands) at a point or points in time. Both approaches generate valuable data that can inform policy analysis.

Below, we state Ostrom's guidance to operationalizing the framework and map illustrative survey questions to each component.

Action Arena

Ostrom (2011: 11-12) identifies seven variables that define the action arena. These variables illuminate the actors, what choices they can make and what they perceive to be the benefits/costs in light of information available to them. These seven variables are listed below with illustrative surveys questions**.**

(i) Set of actors (sex, age, geographic location)

In which of the following settings do you work? Select all that apply. (Wiley 2017)

(ii) Positions of actors (discipline and professional designation).

Please select the description that most closely reflects your research role. (Wiley 2017)

(iii) Set of allowable or expected actions and their linkage to outcomes:

Who is responsible for the execution a research data management plan? (CWTS & Elseiver)
Who is responsible for monitoring compliance of a research data management plan? (CWTS & Elsevier)
Which of the following requirements has your institution made of your research? Select all that apply. (Wiley 2017)

(iv) Level of control an actor has over their decisions:

What resources are required to format your research data for sharing?
Does your institute allocate funds to cover the costs of managing and/or archiving research data? (CWTS & Elsevier)
How challenging do you find it to comply with institutional/funder requirements of your research? (Wiley 2017)

(v) Information available to actors that informs their decisions

How much time do you focus on making your research reproducible? (Wiley 2017)
Who do you believe 'owns' the research data that you have made or will make available to others as part of your last research project? (CWTS & Elsevier)

(vi) Cost/benefit, which serve as incentives and deterrents – assigned to actions and outcomes

What do you think the benefits are (if any) to you as an author for sharing your research data alongside your research article? (CWTS & Elsevier)
Why did you choose to share your data publicly? (Wiley 2017)
What value do you think sharing data has?
To what extent do you feel there is a relationship between your research outputs and the ability to get funding? (Wiley 2017)

(vii) Potential outcomes that are linked to individual sequence of actions[2]

Do you take steps to manage your research data and/or archive it for potential re-use by yourself and/or others? (CWTS & Elsevier)
What types of data do you typically store? Select all that apply. (Wiley 2017)

Contextual Categories

Variables in this category identify the enabling conditions that influence choices within the Action Arena. These variables may change over time but are largely thought of immutable to an individual's influence.

(i) Physical environment (infrastructure; repository capabilities, financial resources)

Which of the following locations do you use to archive your research data (options: repository provided by my funder, publisher, my institute, my department, other)? (CWTS & Elsevier)
Where did you or would you access someone else's data?
Where have (or would) you accessed others' research data? Please list specific websites, for example. (CWTS & Elsevier)

(ii) Community Attributes & Norms (expectations of peers, research community practices, cultural values, social and ethical norms)

Is sharing research data associated with credit or reward in your field? (CWTS & Elsevier)

Which of the following impacts your decision to archive research data (options: my department, colleagues/collaborators, my wider research field (e.g. documented code of conduct or informal practice) (CWTS & Elsevier)

(iii) Formal rules and regulations (government, funder or employer policies)

Which of the following requirements has your institution made of your research [list]? (Wiley)
Which of the following requirements has your primary funder made for your research [list]? (Wiley)
The financial and regulatory aspects of the research environment refer to the regulations that govern your research. Please rate the influence of the following factors

[2] There are two locations in the IAD framework where outcomes are identified: once in Action Arena and again in the 'Outcome' category. In the Action Arena, Ostrom use the label 'potential outcomes' which we understand as the range of possible choices an individual could make whereas the latter is the aggregated outcome of individual choices (e.g., proportion of researchers in a university or country who share the research data).

on your ability to conduct research (selection options: my institution does not have regulations about sharing data / does not support sharing data activities financially; national regulations do not support data sharing activities financially / do not exist for data sharing). (INASP)

Learning / Interaction

Questions of interest are those that help understand how an individual's actions are impacted by learning from past interactions within the action arena. For example, questions that assess whether a researcher holds a positive or negative assessment of previous data-sharing efforts would be a predictor of future actions. Hypothetical questions that probe an individual's likely response to a new service or incentive is another line of questioning.

Some illustrative questions are:

> How interested would you be in a service that helps you deposit your data in a repository? (SpringerNature)
> How interested would you be in a service that helps improves the discoverability of your data? (SpringerNature)
> Are there other ways in which you would like help dealing with your research data? (SpringerNature)

Outcomes

As discussed above, all the surveys are interest in whether researchers share their data. Table 2 provides a list of relevant questions from the sample surveys. These questions result in supply-side outcomes. There are also examples of demand-side outcomes as evidenced by respondents use of, or reporting the benefits of using, archived data. In addition to Table 2, illustrative questions are:

> Have you published the research data that you used or created as part of your last research project in any of the following ways? (CWTS & Elsevier)
> Have you made use of another researcher's publicly available data to answer your own research question(s)? (Wiley 2017)

This mapping exercise illustrates how questions from the selected surveys align with Ostrom's framework. Should a survey sponsor seek to design a survey using this framework, the question bank we drew from should provide suitable guidance to help characterize, and potentially explain, data sharing practices.

The rationale for building a survey using Ostrom's framework (or other models for analysing institutional change) is to account for variables that are conceptually linked to the outcome. The SpringerNature survey, for example, revealed sharp differences in data-sharing practices

between natural and social scientists. Without an explanatory framework and data, analysts are poorly equipped to explain why those differences exist.

The IAD framework can also provide insight into the governance of public goods. One of Ostrom's significant contributions was to demonstrate that the generation and protection of public or common-resource goods is likely dependent on a multi-tiered governance arrangement where formal and informal institutions interact. The role of rules (funder / government requirements), formal organizations (universities), and norms (peer expectations) operate at different levels, have different qualities and ways of influencing data sharing. Some variables may be costly and slow to change whereas others may be inexpensive and quick to change. Knowing these qualities and how these 'institutions' interact would be imminently helpful to those seeking the shape the enabling environment and the incentives for data sharing.

Further thinking along these lines would also assist with the development of more focused or modular surveys. As suggested above, some survey sponsors are interested in understanding the barriers or incentives to data sharing and have no intention of using the responses to explain how those barriers or incentives interact with other factors that influence data sharing practices. Irrespective of the motive and whether modular or holistic surveys are contemplated, designing surveys that generate interoperable and comparative data, and are informed by a theory of change would be a significant contribution.

4. Use of Surveys

This final section provides case examples of how survey sponsors have used survey findings to inform their practices and policies. The case examples illustrate government-led initiatives arising from national-level surveys. In each case, survey data provided insights or the impetus to influence data sharing practices by using resources and influence available to them.

Case 1: Japan[3]

In 2015, the Cabinet Office of Japan released its first official document on Open Science called "Promoting Open Science in Japan - Opening up a new era for the advancement of science"(Government of Japan, 2015). Among other recommendations, the document articulated a policy directive for promoting open data.

The National Institute for Science and Technology Policy (NISTEP) recognized the potential contribution of a national survey to support ongoing analysis of the government's new policy direction. In 2016, NISTEP developed a survey with the intent to understand the perceptions,

[3] This case is based on a presentation by Kazuhiro Hayashi and subsequent comments. Presentation available at: https://www.rd-alliance.org/sites/default/files/RDA%20Botswana%20Kaz%20Hayashi%20_Japan%20Survey.pptx

practices and obstacles that Japanese researchers encountered relating to sharing research data and supporting open access to publications.

Their baseline survey was implemented late in 2016. Over 2000 academic, government research organizations and private companies belonging to NISTEP's Science and Technology Expert Group received the survey and 1,398 responded (~70.5% response rate). NISTEP shared their analysis with the Cabinet Office and the Ministry of Education Culture Sports Science and Technology (MEXT). Government officials were interested to know where Japanese researchers archived research data and what obstacles / disincentives they reported. To overcome one challenge, the government funded the National Institute of Informatics (NII) to develop a cloud-based research data infrastructure called "NII Research Cloud."

The 2016 survey created a baseline profile of Japanese researchers but NISTEP was unable to benchmark their results to other survey findings. Recognizing the utility of comparing results with other surveys, NISTEP developed a second survey in consultation with Springer Nature. For the 2018 follow-up survey, NISTEP identified, translated and incorporated comparable questions from the Springer Nature survey (Allagnat et al 2019).

The NISTEP experience illustrates two points relating to the use and usefulness of its survey. On its use, policy makers supporting the government's open science directive welcomed and utilized NISTEP's survey findings. The Cabinet Office's Integrated Innovation Strategy (2018) cited NISTEP's survey and stated that the government would implement follow-up surveys to monitor progress of the Open Science agenda. On its usefulness, NISTEP recognize that to benchmark the direction of change over time and with other countries, such efforts would be greatly aided by adopting common questions. To make the survey comparable with others, NISTEP collaborated with SpringerNature and modified the 2016 survey (Allagnat et al 2019).

Case 2: Austria[4]

In 2015, *e-Infrastructure Austria* launched the Austrian National Research Data Survey. The survey sought to determine how Austrian researchers manage their data and more generally, to raise awareness within the research community on how to support open science (Bauer 2015).

The survey incorporated the following components: data types and formats; data archiving practices, backup and loss procedures; ethical and legal aspects; accessibility to and subsequent use of data; and, infrastructure and services. The survey reached an estimated population of

[4] This case is based on a presentation by Paulo Budroni available at: https://www.rd-alliance.org/sites/default/files/RDA%20Botswana%20P%20Budroni%20Austria_learning%20from%20surveys.pptx

36000 researchers working in the public research system. There were 3026 respondents which equated to a 9% response rate.

The report, *Researchers and their Data*, served as a platform for policy development at multiple levels. In addition to publishing the report, numerous dialogues and workshops were organized to present the findings and identify actions to support open data. These consultations were held across the country at medical, technical, academic and artistic centres involving a range of stakeholders, including academic administrators, librarians, ICT officers, legal advisers, representatives of trade unions, scientists, engineers and artists.

As one observer of the process commented, these processes enabled a reimagining of how to reshape and create a new ecosystem of services (Budroni 2018). Among the proposals to emerge were project ideas for building shared data infrastructures, designing data management workflows, catalysing data re-use scenarios, and enabling service and support for open science.

The survey and ensuing dialogues, both structured and informal, are credited with changing the discourse in Austria and catalysing concrete outcomes. In terms of shaping the discourse, the survey results increased awareness among researchers and government. Both the benefits and challenges are better understood within the research community and in government. This awareness raising was critical to ensuing changes in policy and practice. For example, three universities have created research data management policies and a further nine are under development. In November 2018, a new data librarian programme was initiated and focal points for data curation began emerging across Austria's universities. The national research funder now explicitly supports open data within its funding programs and supports data infrastructure.

Broad interest in promoting capacities in research data management led to the creation of a national chapter of the Research Data Alliance network. The Austrian experience was also watched from abroad. Those involved in designing the survey and its dissemination were invited by government agencies and academic associations in Turkey, Lebanon, Germany, Spain, Hungary, the Netherlands to share their experience and subsequent efforts in those countries can be traced back to the Austrian survey. For example, Italy and Turkey launched an open data survey modelled on the Austrian precedent.

5. Conclusion

Surveys designers and sponsors examining the practices and perceptions of researchers have contributed significantly to our evolving understanding of research data as a global public good. The data from thousands of researchers across the globe indicate what countries and disciplines lead the way in creating this open data commons and what obstacles they are encountering.

The composite picture from the sample of surveys we examined, however, does not result in a higher resolution worldview of this changing landscape. Rather, we have a montage of separate pictures as each survey focuses in on different questions with different sample populations. There are advantages to this inductive approach. For one, there is a great deal of diversity in terms of what, where and how questions are asked. An in-depth analysis of the context (infrastructure, policy environment) may be extremely valuable in one jurisdiction but unnecessary in another. A montage approach is useful in these regards but this approach is not well suited for other purposes.

When researchers or policy analysts seek to benchmark change over time and compare with other countries/populations, or to explain different outcomes, we need another approach. The mapping of survey questions to Ostrom's IAD framework was introduced for these two purposes – comparability and explanatory power.

Further work on standardizing survey questions would permit researchers to compare findings across surveys. The difficulties we encountered in comparing the most basic indicators makes the case for the adoption of a core set of questions that define outcomes and other predictive variables.

The study of how and why communities succeed or fail in building and sustaining common pool and public goods is a foundational question in policy analysis. Ostrom was awarded the Nobel Prize for her contributions to helping us understand the variables that influence individual and collective choices and how to explain change over time. The case for research data as a public good has been made by others and governments world-wide who fund public science have turned their attention to supporting and mandating open research data. Open data surveys provide empirical data that analysts should be drawing on to monitor and explain the effect of these interventions. The case studies from Japan and Austria illustrate that efforts are underway in these regards. We have argued that framing questions survey designers ask within a theory of change framework can only help advance our creative use of data that is being collected on the practices and perceptions of researchers toward open research data.

6. Bibliography

Allagnat, L., Allin, K., Baynes, G., Hrynaszkiewicz, I., Lucraft, M. 2019.Challenges and Opportunities for Data Sharing in Japan. figshare. Online resource. https://doi.org/10.6084/m9.figshare.7999451.v1

Astell, M., Hrynaszkiewicz, I., Allin, K., Penny, D., Lucraft, M., Baynes, G., et al. 2018. Practical challenges for researchers in data sharing - Springer Nature survey data (anonymised). figshare. Dataset. https://doi.org/10.6084/m9.figshare.5971387.v2

Bauer, B., Ferus, A., Gorraiz, J., Gründhammer, V., Gumpenberger, C., Maly, N., Mühlegger, J.M., Preza, J.L., Sánchez Solís, B., Schmidt, N. & Steineder, C. 2015. **Researchers and their data. Results of an Austria survey – Report 2015**. Version 1.2. DOI: 10.5281/zenodo.34005.

Berghmans, S., Cousijn, H., Deakin, G., Meijer, I., Mulligan, A., Plume, A., de Rijcke, S., Rushforth, A., Tatum, C., van Leeuwen, T., Waltman, L. 2017. **Open Data: the researcher perspective - survey and case studies**. Mendeley Data, v1 http://dx.doi.org/10.17632/bwrnfb4bvh.1

Bezuidenhout, L., & Chakauya, E. (2017). INASP survey final.pdf (Version 3). figshare. https://doi.org/10.6084/m9.figshare.4818043.v3

Bezuidenhout, L., & Chaukaya, E. 2018. Hidden concerns of sharing research data by low/middle income country scientists. *Global Bioethics*, 29 (1): 39-54. https://doi.org/10.1080/11287462.2018.1441780

Blomquist, W. and deLeon, P. 2011. The Design and Promise of the Institutional Analysis and Development Framework. *Policy Studies Journal*, 39: 1-6. doi:10.1111/j.1541-0072.2011.00402.x

Braunschweig, Katrin, Julian Eberius, Maik Thiele and Wolfgang Lehner. 2012. **The State of Open Data: Limits of Current Open Data Platforms**, https://wwwdb.inf.tu-dresden.de/opendatasurvey/www2012_short.pdf

Budroni, P. 2018. Learning from Surveys. Research Data Alliance 12th Plenary meeting presentation. Gaborone, Botswana.

Danish National Research Foundation 2017. **Open access to data – it's not that simple**. https://dg.dk/wp-content/uploads/2017/11/Open-Access-to-Data-—-It's-not-that-Simple.compressed.pdf.

Enwald, H., Kortelainen, T., & Huotari, M. 2017. **Open research data: Experiences and opinions of scholars in Finland.** presentation doi.org/10.6084/m9.figshare.5624779.v1

Elsayed, A.M., & Saleh, E.I. 2018. Research data management and sharing among researchers in Arab universities: An exploratory study. *International Federation of Library Associations and Institutions*, pp. 1-19, https://doi.org/10.1177/0340035218785196

CWTS and Elsevier 2017. **Open Data Research –a researcher perspective**. https://www.elsevier.com/__data/assets/pdf_file/0004/281920/Open-data-report.pdf

Eynden, V., Knight, G., Vlad, A., Radler, B., Tenopir, C., Leon, D., Manista, F., Whitworth, J., Corti, L. 2016. **Survey of Wellcome researchers and their attitudes to open research**. figshare. Journal Contribution. https://doi.org/10.6084/m9.figshare.4055448.v1

Fecher B, Friesike S, Hebing M 2015. What Drives Academic Data Sharing?. *PLoS ONE* 10(2): e0118053. doi:10.1371/journal.pone.0118053

Figshare, Digital Science Report 2016. The State of Open Data: A Selection of Analyses and Articles About Open Data, curated by Figshare www.dx.doi.org/10.6084/m9.figshare.4036398

Jepperson, R. 1991. 'Institutions, Institutional Effects, and Institutionalism', in **The New Institutionalism in Organizational Analysis**, Walter Powell and Paul DiMaggio (eds.). Chicago: University of Chicago Press.

Government of Japan. 2015. **Promoting Open Science in Japan — Opening up a new era for the advancement of science**. The Expert Panel on Open Science, Cabinet office. https://www8.cao.go.jp/cstp/sonota/openscience/150330_openscience_en1.pdf

McGinnis, M. 2016. Updated Guide to IAD and the Language of the Ostrom Workshop: A Simplified Overview of a Complex Framework for the Analysis of Institutions and their Development. Available at: http://php.indiana.edu/~mcginnis/iad_guide.pdf

Nature Research. 2016. Open Data Survey. figshare. Dataset. https://doi.org/10.6084/m9.figshare.4010541.v4

Nature Research, Astell, M., Penny, D., Treadway, J., Fane, B. 2017. State of Open Data survey 2017. figshare. Dataset. https://doi.org/10.6084/m9.figshare.5480710.v3

Nature Research. 2018. State of Open Data 2018. figshare. Dataset. https://doi.org/10.6084/m9.figshare.7234985.v1

Ostrom, E. 1990. **Governing the Commons: The Evolution of Institutions for Collective Action**. New York: Cambridge University Press.

Ostrom, E. 2005. **Understanding Institutional Diversity**. Princeton, NJ: Princeton University Press.

Ostrom, E. 2009. Beyond Markets and States: Polycentric Governance of Complex Economic Systems. Nobel Prize Lecture, December 8, 2009. https://www.nobelprize.org/uploads/2018/06/ostrom_lecture.pdf

Ostrom, E. 2011. Background on the institutional Analysis and Development Framework. *Policy Studies Journal*, 39(1):7-27.

Schmidt, B., Gemeinholzer, B., Treloar, A., Hodge, J., Santanchè, A., & Oakley, K. 2015. Belmont Forum Open Data Survey 2014. Zenodo. Data set. http://doi.org/10.5281/zenodo.1172960

Schmidt B, Gemeinholzer B, Treloar A. 2016. Open Data in Global Environmental Research: The Belmont Forum's Open Data Survey. *PLoS ONE* 11(1): e0146695. https://doi.org/10.1371/journal.pone.0146695

Serwadda, D., Ndebele, P., Grabowski, M.K., Bajunirwe, F., & Wanyenze, R.K. 2018. Open data sharing and the Global South – Who benefits? *Science*, 359(6376): 642 – 643. http://science.sciencemag.org/content/359/6376/642

SPARC and DCC 2017. An Analysis of Open Data and Open Science Policies in Europe, (May 2017).

SPARC and DCC 2018. An Analysis of Open Data and Open Science Policies in Europe, v2.1 (January 2018).

SpringerNature 2018. **Practical challenges for researchers in sharing data. White Paper**. https://doi. org/10.6084/m9.figshare.5975011

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., Dalton ED., Allard S., Frame M., Pjesivac I., Birch B., Pollock D., Dorsett K. 2015a. Data from: Changes in data sharing and data reuse practices and perceptions among scientists worldwide. Dryad Digital Repository. https://doi.org/10.5061/dryad.1ph92

UNESCO 2015. **UNESCO Science Report: towards 2030**. Paris: UNESCO.

Wiley 2016. Wiley Data Sharing Survey. figshare. Dataset. https://doi.org/10.6084/m9.figshare.3468368.v2

Wiley 2017. Wiley Open Science Researcher Survey 2016. figshare. Dataset. https://doi.org/10.6084/m9.figshare.4748332

Wiley 2017a. Wiley Open Science Researcher Survey 2016 Infographic. https://doi.org/10.6084/m9.figshare.4910714.v1

Woolfrey, L. 2015. An Open African Data approach to improving data quality, A DPRU Policy brief commissioned for the World Bank Group by DataFirst Data Service, University of Capetown. www.dpru.uct.ac.za/sites/default/files/image_tool/.../DPRU%20PB%2014-42.pdf.

Wolff-Eisenberg, C., Rod, A., Schonfeld, R. 2016. UK survey of academics - 2015. IthakaS+R, Jisc, RLUK https://doi.org/10.18665/sr.282736.

Appendix A: Geographic distribution of responses data tables and heat maps

Table A1: Total responses from each country for each survey used to generate heat maps

ABREV	COUNTRY	CONTINENT	Belmont_ 2016	Elsevier_ 2017	SOD_ 2016	SOD_ 2017	SOD_ 2018	Nature_ 2018	Tenopir_ 2009_2010	Tenopir_ 2014_2015	Wiley_ 2014	Wiley_ 2016	TOTAL
AFG	Afghanistan	Asia	2	1	0	0	0	0	0	1	0	0	4
ALB	Albania	Europe	1	0	1	1	1	6	0	0	1	2	13
DZA	Algeria	Africa	1	1	1	8	5	4	0	0	4	4	28
AND	Andorra	Europe	0	0	2	1	1	0	0	0	0	0	4
AGO	Angola	Africa	0	0	0	0	0	1	0	0	0	0	1
AIA	Anguilla	North_America	0	0	1	0	1	0	0	0	0	0	2
ARG	Argentina	South_America	7	4	7	35	12	18	7	6	32	50	178
ARM	Armenia	Other	0	0	0	2	0	1	0	0	1	2	6
AUS	Australia	Oceania	50	31	38	108	37	191	18	16	87	201	777
AUT	Austria	Europe	19	4	6	8	16	66	3	1	3	28	154
AZE	Azerbaijan	Other	0	1	0	0	0	0	0	1	0	0	2
BHR	Bahrain	Asia	0	0	0	0	1	1	0	1	0	0	3
BGD	Bangladesh	Asia	1	1	2	15	4	4	0	0	3	15	45
BLR	Belarus	Europe	1	0	1	2	0	2	0	0	0	2	8
BEL	Belgium	Europe	29	4	13	13	5	110	4	1	3	37	219
BEN	Benin	Africa	0	0	0	0	1	1	0	0	0	0	2
BMU	Bermuda	North_America	0	0	0	2	0	1	0	0	0	0	3
BOL	Bolivia	South_America	0	0	0	1	0	2	0	1	0	0	4
BIH	Bosnia and Herzegovina	Europe	0	1	0	1	1	3	0	0	1	0	7
BWA	Botswana	Africa	0	0	0	1	0	2	2	1	0	2	8
BRA	Brazil	South_America	12	34	51	123	43	73	15	31	158	163	703
BGR	Bulgaria	Europe	2	0	3	6	2	31	2	2	8	4	60
BFA	Burkina Faso	Africa	1	0	0	1	0	1	0	2	0	1	6
KHM	Cambodia	Asia	0	0	0	2	0	0	0	1	1	2	6
CMR	Cameroon	Africa	0	1	1	5	2	2	1	3	0	4	19
CAN	Canada	North_America	34	23	52	78	40	412	47	103	19	129	937
CPV	Cape verde	Africa	0	0	0	1	0	1	1	0	0	0	3
TCD	Chad	Africa	0	0	0	1	0	0	1	0	0	0	2
CHL	Chile	South_America	4	7	3	8	1	13	2	1	24	43	106
CHN	China	Asia	43	111	45	116	28	62	33	34	90	387	949
COL	Colombia	South_America	5	3	6	12	8	12	3	9	15	15	88
COD	Congo Democratic Republic of	Africa	1	0	0	1	1	0	1	1	0	0	5
CRI	Costa Rica	North_America	0	0	2	2	3	1	2	1	3	2	16
CIV	Cote d'Ivoire	Africa	0	1	0	0	0	0	0	0	0	0	1
HRV	Croatia	Europe	2	3	3	7	1	45	2	0	12	13	88
CUB	Cuba	North_America	2	0	0	3	0	0	0	0	0	2	7
CYP	Cyprus	Europe	2	3	0	2	1	14	2	0	6	7	37
CZE	Czech Republic	Europe	2	6	8	7	4	89	6	1	22	29	174
DNK	Denmark	Europe	9	4	10	10	5	157	5	4	2	33	239
DOM	Dominican Republic	North_America	0	0	0	0	0	1	1	0	0	0	2
ECU	Ecuador	South_America	1	1	1	1	1	3	2	2	2	6	20
EGY	Egypt	Africa	1	5	2	19	8	10	2	2	6	64	119
SLV	El Salvador	North_America	0	0	2	0	0	0	0	0	0	0	2
EST	Estonia	Europe	2	2	1	2	1	17	2	0	7	8	42
ETH	Ethiopia	Africa	1	1	0	16	2	4	1	1	0	8	34

FLK	Falkland Islands	South_America	0	0	1	0	0	0	0	0	0	0	1
FJI	Fiji	Oceania	1	0	0	0	0	1	0	0	0	1	3
FIN	Finland	Europe	15	4	3	15	2	87	7	6	3	28	170
FRA	France	Europe	72	20	36	17	21	254	11	7	50	80	568
GUF	French Guiana	South_America	1	0	0	0	0	0	0	0	0	0	1
PYF	French Polynesia	Oceania	0	0	0	0	1	1	0	0	0	0	2
GAB	Gabon	Africa	0	0	0	0	0	1	0	0	0	0	1
GMB	Gambia	Africa	0	0	0	0	0	1	0	0	0	0	1
GEO	Georgia	Other	2	1	0	0	1	0	0	0	1	0	5
DEU	Germany	Europe	212	66	82	65	59	435	28	21	22	141	1131
GHA	Ghana	Africa	1	3	1	2	1	2	0	2	2	6	20
GRC	Greece	Europe	23	5	14	26	12	153	4	3	8	45	293
GRL	Greenland	North_America	0	0	0	0	0	1	0	0	0	0	1
GRD	Grenada	North_America	0	0	0	0	0	0	0	0	0	2	2
GTM	Guatemala	North_America	0	0	1	0	0	0	0	1	0	1	3
GUY	Guyana	South_America	0	0	1	0	0	0	0	0	0	0	1
HTI	Haiti	North_America	0	0	0	0	1	0	0	0	0	0	1
HKG	Hong Kong	Asia	1	5	1	6	1	8	0	0	16	16	54
HUN	Hungary	Europe	2	3	2	4	3	56	2	2	3	10	87
ISL	Iceland	Europe	0	0	1	1	0	7	0	0	0	5	14
IND	India	Asia	24	54	92	182	90	99	16	20	98	320	995
IDN	Indonesia	Asia	1	5	4	12	7	2	3	1	2	8	45
IRN	Iran	Asia	2	7	16	60	21	28	2	6	51	156	349
IRQ	Iraq	Asia	1	1	2	3	4	3	0	0	2	5	21
IRL	Ireland	Europe	4	4	9	6	3	52	2	0	4	20	104
ISR	Israel	Asia	7	5	9	30	6	9	0	0	31	29	126
ITA	Italy	Europe	119	48	60	33	49	735	21	10	36	328	1439
JAM	Jamaica	North_America	0	0	0	0	0	2	0	0	0	4	6
JPN	Japan	Asia	30	42	41	30	19	32	8	7	62	199	470
JOR	Jordan	Asia	0	3	2	5	5	3	0	0	5	21	44
KAZ	Kazakhstan	Other	0	2	1	0	1	0	0	1	1	2	8
KEN	Kenya	Africa	2	3	3	6	1	3	2	4	4	7	35
KIR	Kiribati	Oceania	0	1	0	0	0	1	0	0	0	0	2
PRK	Korea North	Asia	0	9	0	0	0	1	0	0	0	0	10
KOR	Korea South	Asia	8	30	17	31	6	11	1	2	28	47	181
KOS	Kosovo	Europe	0	0	0	0	0	1	0	0	0	0	1
KWT	Kuwait	Asia	1	1	0	0	1	0	0	0	4	1	8
LVA	Latvia	Europe	0	2	1	0	2	8	0	0	2	1	16
LBN	Lebanon	Asia	1	1	2	2	1	2	0	0	1	5	15
LSO	Lesotho	Africa	0	0	0	1	0	0	0	0	0	0	1
LIE	Liechtenstein	Europe	0	0	0	0	0	1	0	0	0	0	1
LBY	Libya	Africa	0	0	0	0	1	0	1	1	0	1	4
LTU	Lithuania	Europe	0	0	2	1	3	17	0	0	13	4	40
LUX	Luxembourg	Europe	1	0	0	0	0	9	0	0	0	1	11
MKD	Macedonia	Europe	0	0	0	0	0	10	0	1	3	0	14
MWI	Malawi	Africa	0	0	0	1	2	0	0	0	0	1	4
MYS	Malaysia	Asia	2	8	5	29	5	6	1	0	15	48	119
MLI	Mali	Africa	0	1	0	1	1	0	0	0	0	0	3
MLT	Malta	Europe	0	0	0	1	1	3	0	0	0	5	10
MUS	Mauritius	Africa	0	0	0	0	1	0	0	0	0	0	1
MEX	Mexico	North_America	5	12	17	59	28	46	9	3	20	64	263

MDA	Moldova	Europe	0	0	1	0	0	0	0	0	0	0	1
MCO	Monaco	Europe	0	0	0	0	0	1	0	0	0	0	1
MNG	Mongolia	Asia	0	1	0	0	0	1	0	0	0	0	2
MSR	Montserrat	North_America	0	0	0	0	0	1	0	0	0	0	1
MAR	Morocco	Africa	0	1	0	9	2	2	0	1	1	4	20
MOZ	Mozambique	Africa	0	0	0	0	0	0	1	0	0	1	2
MMR	Myanmar	Asia	1	0	0	1	0	0	0	0	0	1	3
NAM	Namibia	Africa	0	1	0	2	0	0	0	0	0	0	3
NPL	Nepal	Asia	1	1	1	7	0	0	2	0	0	8	20
NLD	Netherlands	Europe	33	14	23	30	18	219	10	7	33	49	436
NCL	New Caledonia	Oceania	0	0	0	0	1	1	0	0	0	0	2
NZL	New Zealand	Oceania	8	3	12	18	15	36	4	4	21	21	142
NIC	Nicaragua	North_America	0	0	1	0	2	0	0	0	0	0	3
NGA	Nigeria	Africa	5	7	4	24	11	10	2	3	5	18	89
NOR	Norway	Europe	29	3	12	10	2	116	2	1	3	44	222
OMN	Oman	Asia	0	0	0	1	0	0	0	0	0	4	5
PAK	Pakistan	Asia	1	2	9	18	9	15	1	2	11	33	101
PLW	Palau	Oceania	0	0	0	0	0	1	0	0	0	0	1
PSE	Palestinian Territory	Asia	0	1	0	0	0	0	0	0	0	0	1
PAN	Panama	North_America	0	0	0	2	0	0	1	2	0	1	6
PNG	Papua New Guinea	Oceania	1	0	0	1	0	0	0	0	0	0	2
PRY	Paraguay	South_America	0	0	1	0	0	0	0	0	0	0	1
PER	Peru	South_America	3	2	4	6	1	5	1	3	4	4	33
PHL	Philippines	Asia	0	1	0	13	1	5	1	0	2	10	33
POL	Poland	Europe	7	7	13	12	7	131	3	3	8	60	251
PRT	Portugal	Europe	10	13	18	27	17	165	1	7	33	72	363
PRI	Puerto Rico	North_America	0	0	0	0	0	4	0	0	0	0	4
QAT	Qatar	Asia	1	0	2	3	1	2	0	1	1	4	15
ROU	Romania	Europe	14	6	5	12	12	63	2	3	33	18	168
RUS	Russia	Other	0	96	11	58	11	81	5	5	30	37	334
LCA	Saint Lucia	North_America	0	0	0	1	0	0	0	0	0	0	1
WSM	Samoa	Oceania	0	0	0	1	0	0	0	0	0	0	1
SAU	Saudi Arabia	Asia	0	2	5	6	3	7	0	0	8	20	51
SEN	Senegal	Africa	1	0	0	0	0	0	0	0	0	1	2
SRB	Serbia	Europe	0	5	1	5	1	30	0	1	27	25	95
SYC	Seychelles	Africa	0	0	0	0	0	0	0	1	0	0	1
SGP	Singapore	Asia	1	5	7	13	4	4	1	3	8	21	67
SVK	Slovakia	Europe	1	1	3	5	2	28	1	1	9	4	55
SVN	Slovenia	Europe	4	2	1	1	8	22	1	0	3	9	51
SLB	Solomon Islands	Oceania	0	0	0	0	13	0	0	0	0	1	14
ZAF	South Africa	Africa	2	5	7	24	0	9	2	41	12	38	140
ESP	Spain	Europe	47	35	42	81	46	558	15	18	35	198	1075
LKA	Sri Lanka	Asia	1	0	1	7	1	2	0	0	5	4	21
SDN	Sudan	Africa	0	0	0	0	1	1	0	1	0	2	5
SWZ	Swaziland	Africa	0	0	0	1	0	0	1	0	1	0	3
SWE	Sweden	Europe	24	9	18	21	22	165	3	7	20	57	346
CHE	Switzerland	Europe	29	6	23	9	12	120	11	4	24	33	271
SYR	Syria	Asia	1	0	0	2	1	0	0	0	0	1	5
TWN	Taiwan	Asia	7	24	7	21	3	14	2	0	27	62	167
TZA	Tanzania	Africa	0	1	0	3	1	4	2	2	0	5	18
THA	Thailand	Asia	1	5	2	11	10	6	2	4	10	45	96

TGO	Togo	Africa	0	0	0	0	0	2	0	0	0	1	3
TTO	Trinidad and Tobago	South_America	0	1	0	1	2	1	0	0	0	0	5
TUN	Tunisia	Africa	0	1	1	7	0	1	2	1	4	14	31
TUR	Turkey	Other	7	7	9	28	17	16	4	5	43	121	257
TCA	Turks and Caicos Islands	North_America	0	0	0	1	0	0	0	0	0	0	1
UGA	Uganda	Africa	0	1	0	5	0	1	0	2	0	4	13
UKR	Ukraine	Europe	1	2	2	11	9	18	1	2	3	8	57
ARE	United Arab Emirates	Asia	0	3	2	2	4	5	0	0	4	6	26
GBR	United Kingdom	Europe	90	72	141	98	78	615	36	22	109	182	1443
USA	United States	North_America	185	242	342	464	222	1753	899	482	1018	515	6122
URY	Uruguay	South_America	2	2	3	1	2	8	2	1	3	7	31
UZB	Uzbekistan	Asia	0	1	0	1	0	1	0	1	0	0	4
VUT	Vanuatu	Oceania	0	0	1	0	0	0	0	0	0	0	1
VEN	Venezuela	South_America	0	3	3	6	4	3	0	1	1	2	23
VNM	Vietnam	Asia	0	2	3	3	1	6	0	1	0	12	28
PSE	West Bank	Asia	0	0	0	0	0	1	0	0	0	0	1
YEM	Yemen	Asia	0	0	0	0	1	1	0	0	1	1	4
ZMB	Zambia	Africa	1	0	0	3	0	1	0	1	0	1	7
ZWE	Zimbabwe	Africa	0	0	0	1	0	1	1	2	0	2	7

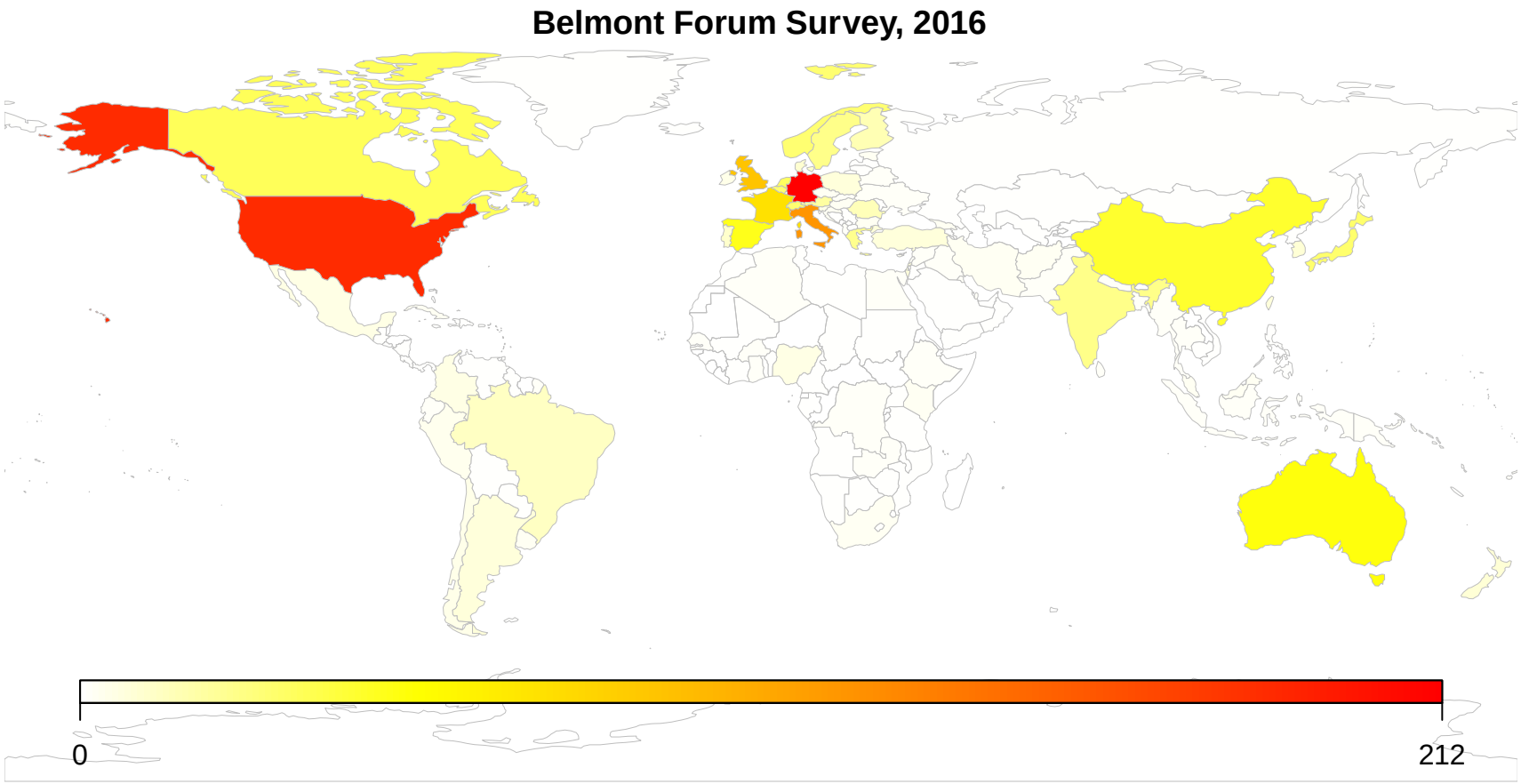

Figure A1: Heat-map distribution of survey respondents by country for the Belmont Forum survey conducted in 2016

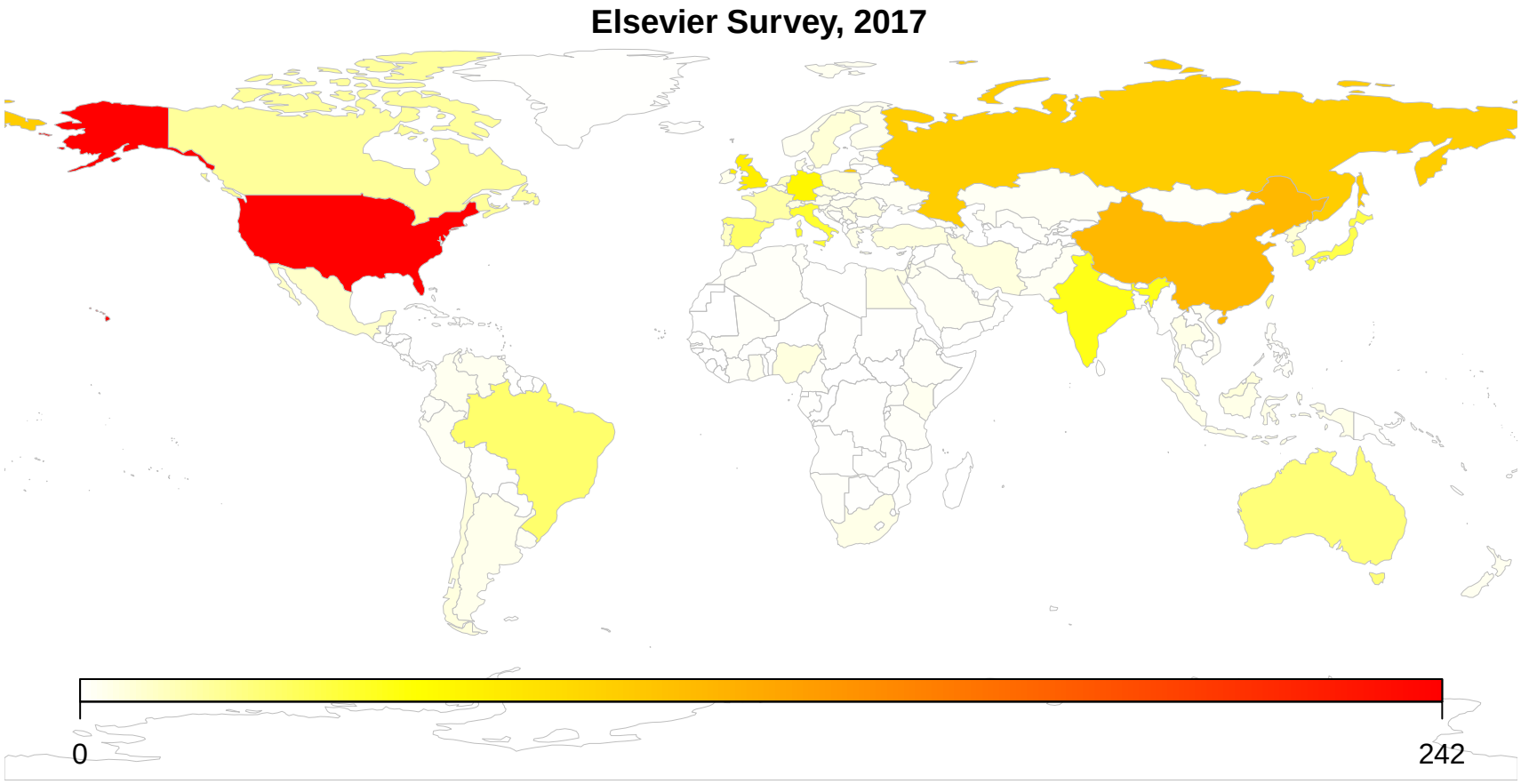

Figure A2: Heat-map distribution of survey respondents by country for the Elsevier& CWTS survey conducted in 2017

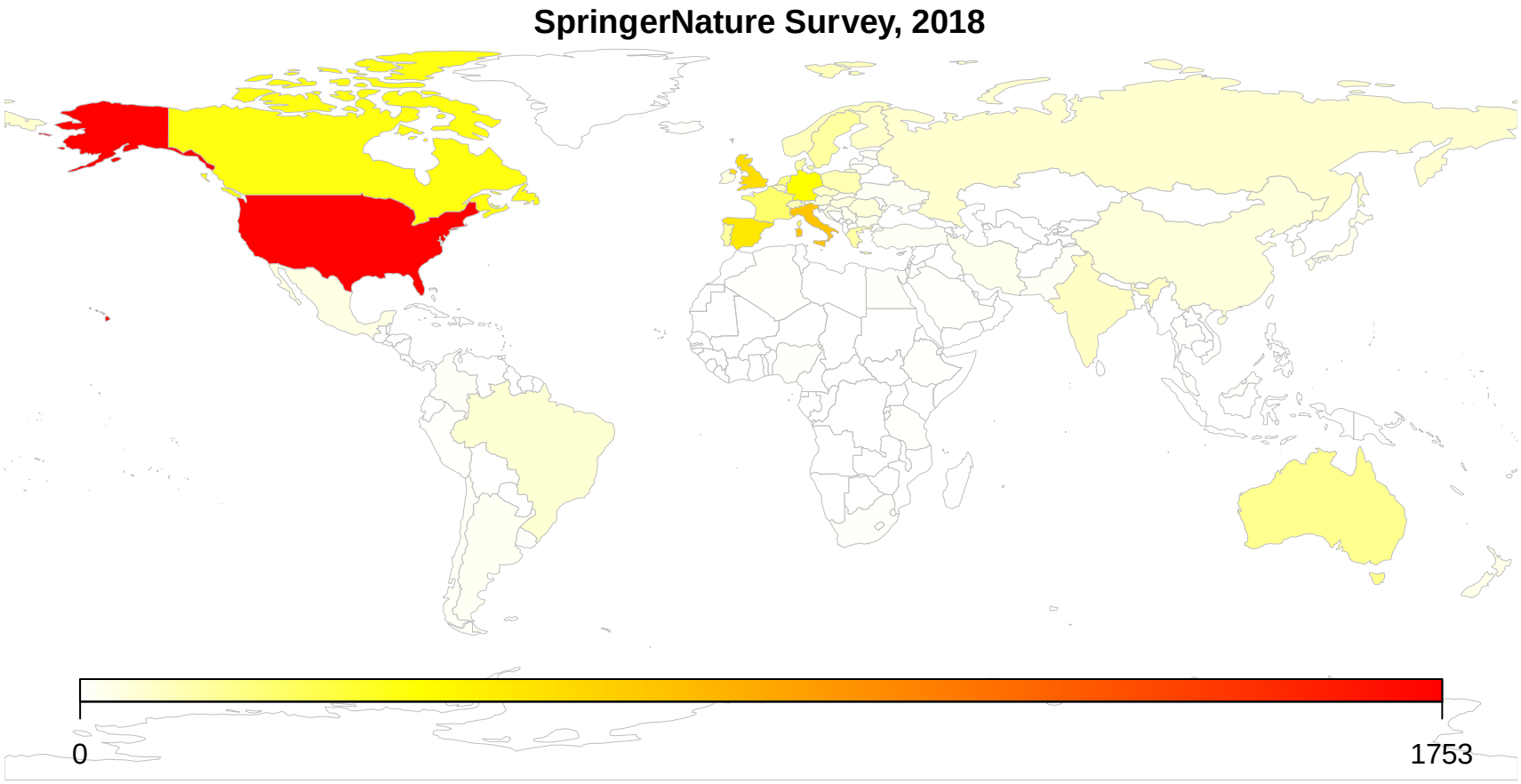

Figure A3: Heat-map distribution of survey respondents by country for the SpringerNature survey conducted in 2018

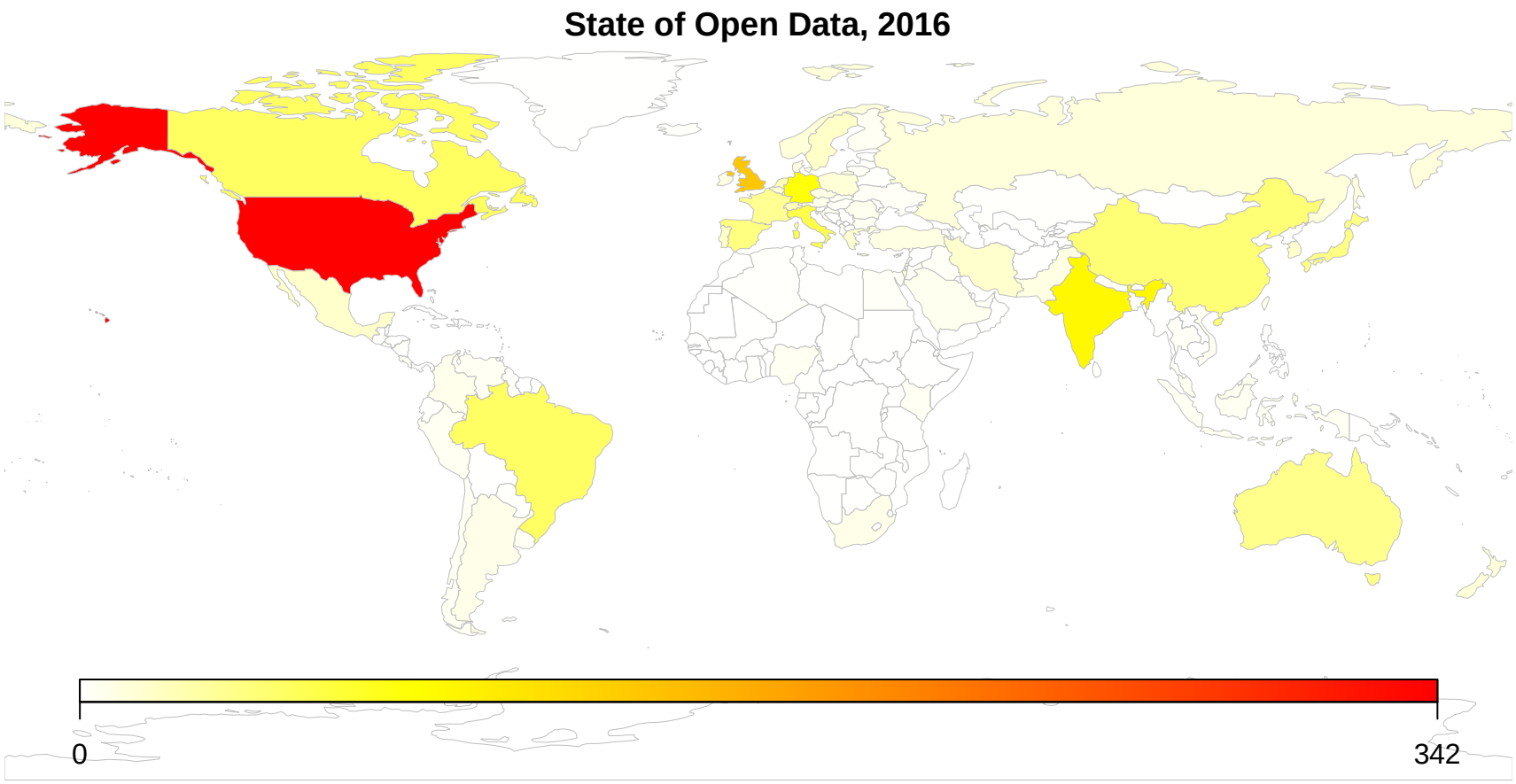

Figure A4: Heat-map distribution of survey respondents by country for State of Open Data survey conducted in 2016

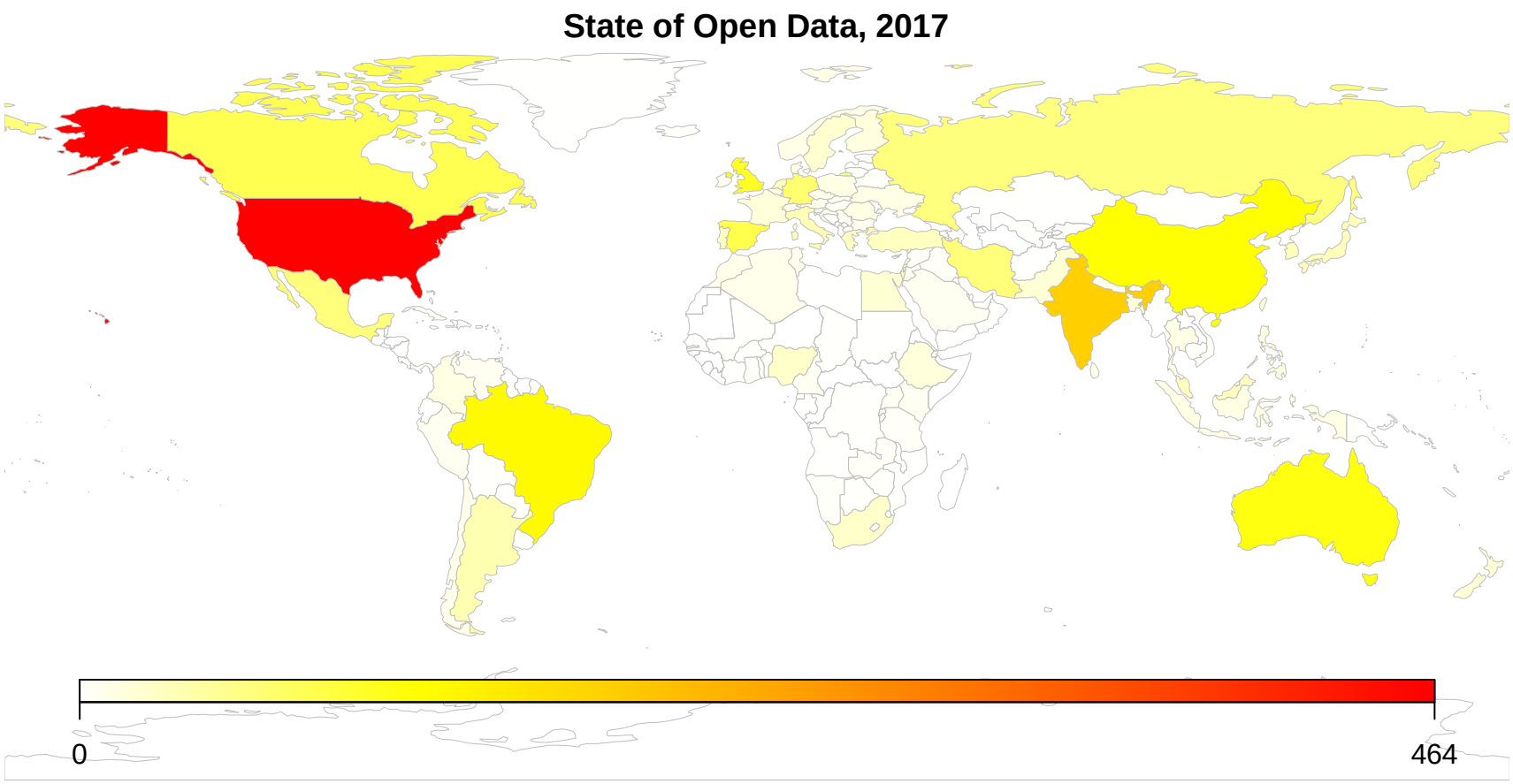

Figure A5: Heat-map distribution of survey respondents by country for the State of Open Data survey conducted in 2017

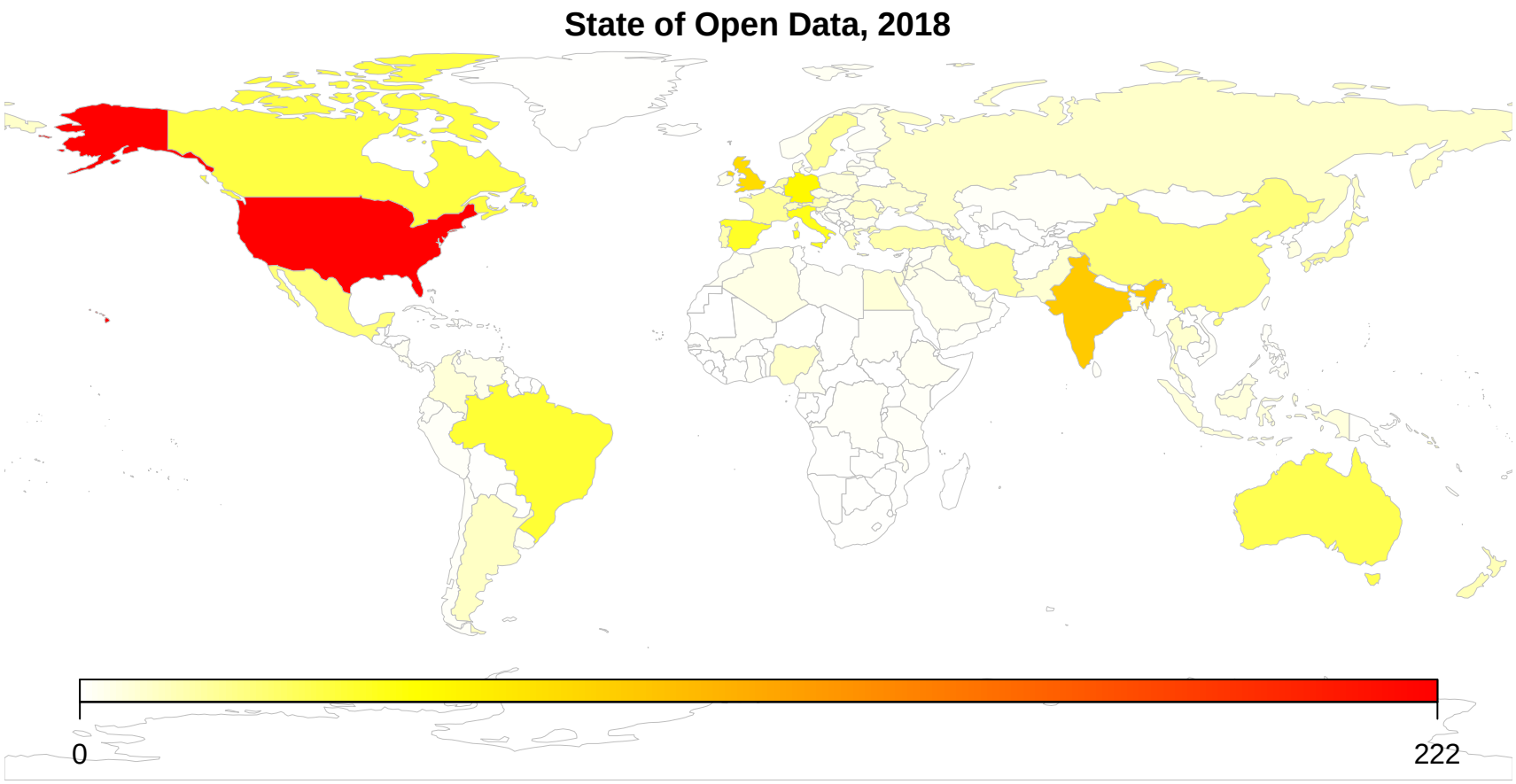

Figure A6: Heat-map distribution of survey respondents by country for the State of Open Data survey conducted in 2018

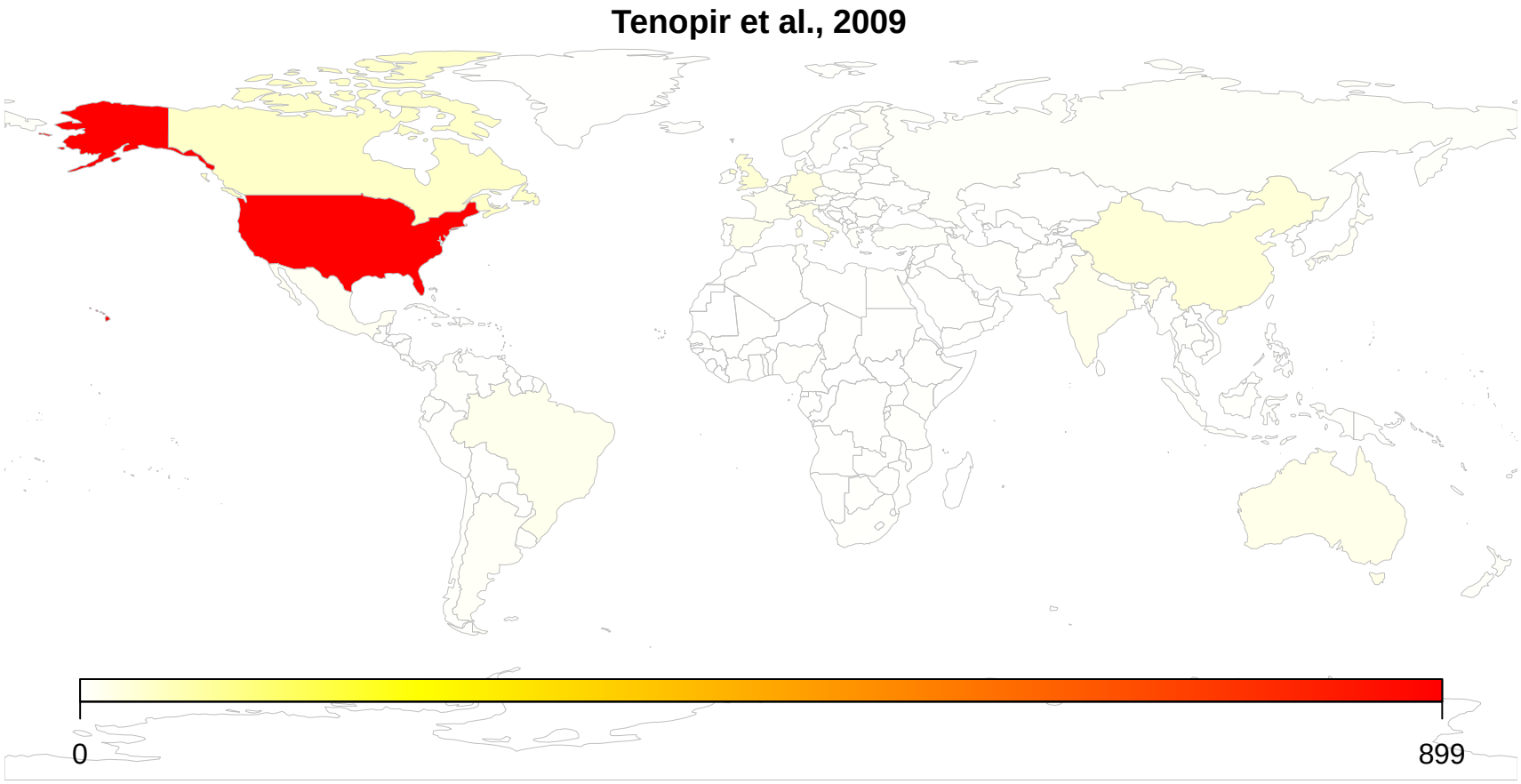

Figure A7: Heat-map distribution of survey respondents by country for the survey conducted by Tenopir et al. in 2009-2010

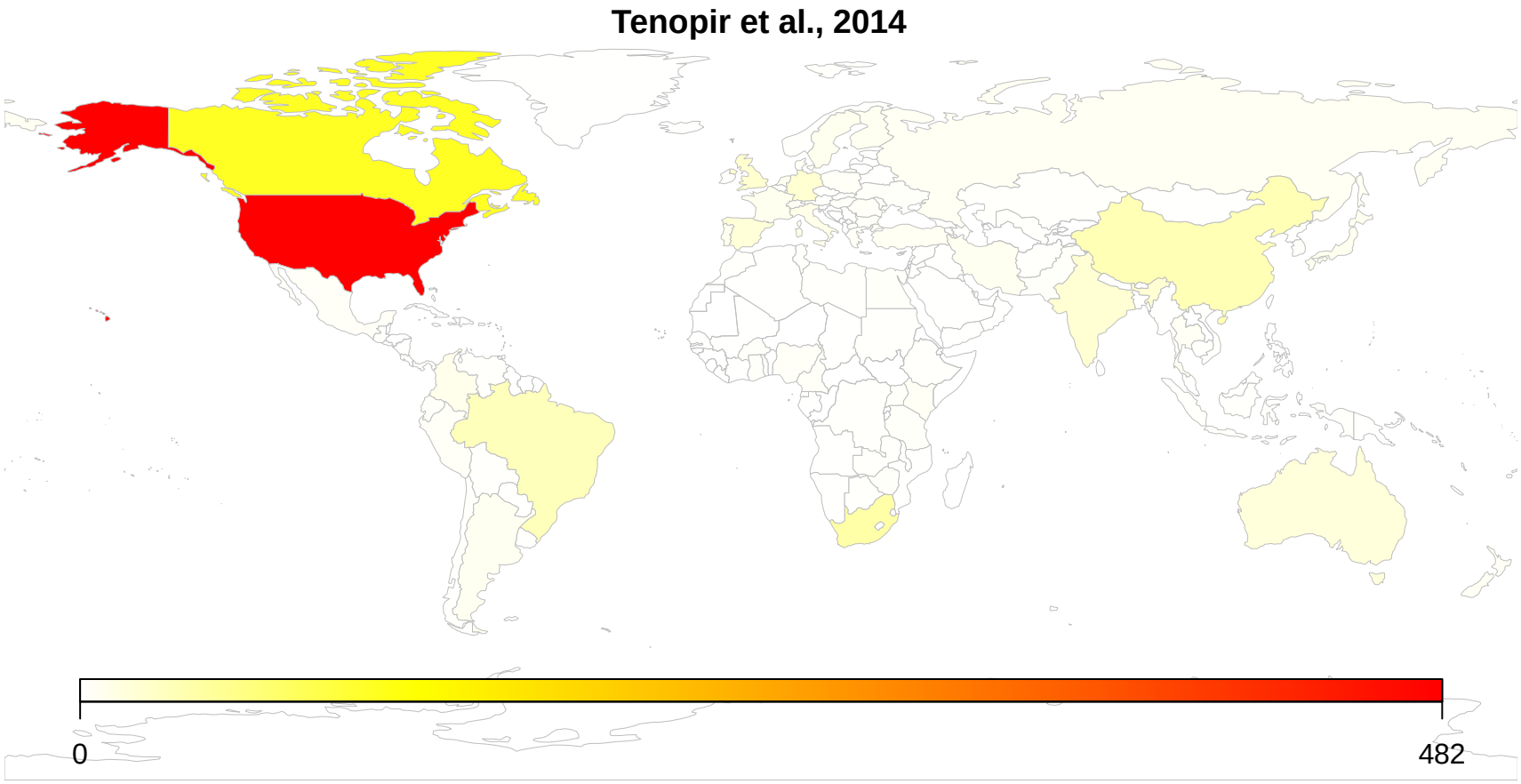

Figure A8: Heat-map distribution of survey respondents by country for the survey conducted by Tenopir et al. in 2014-2015

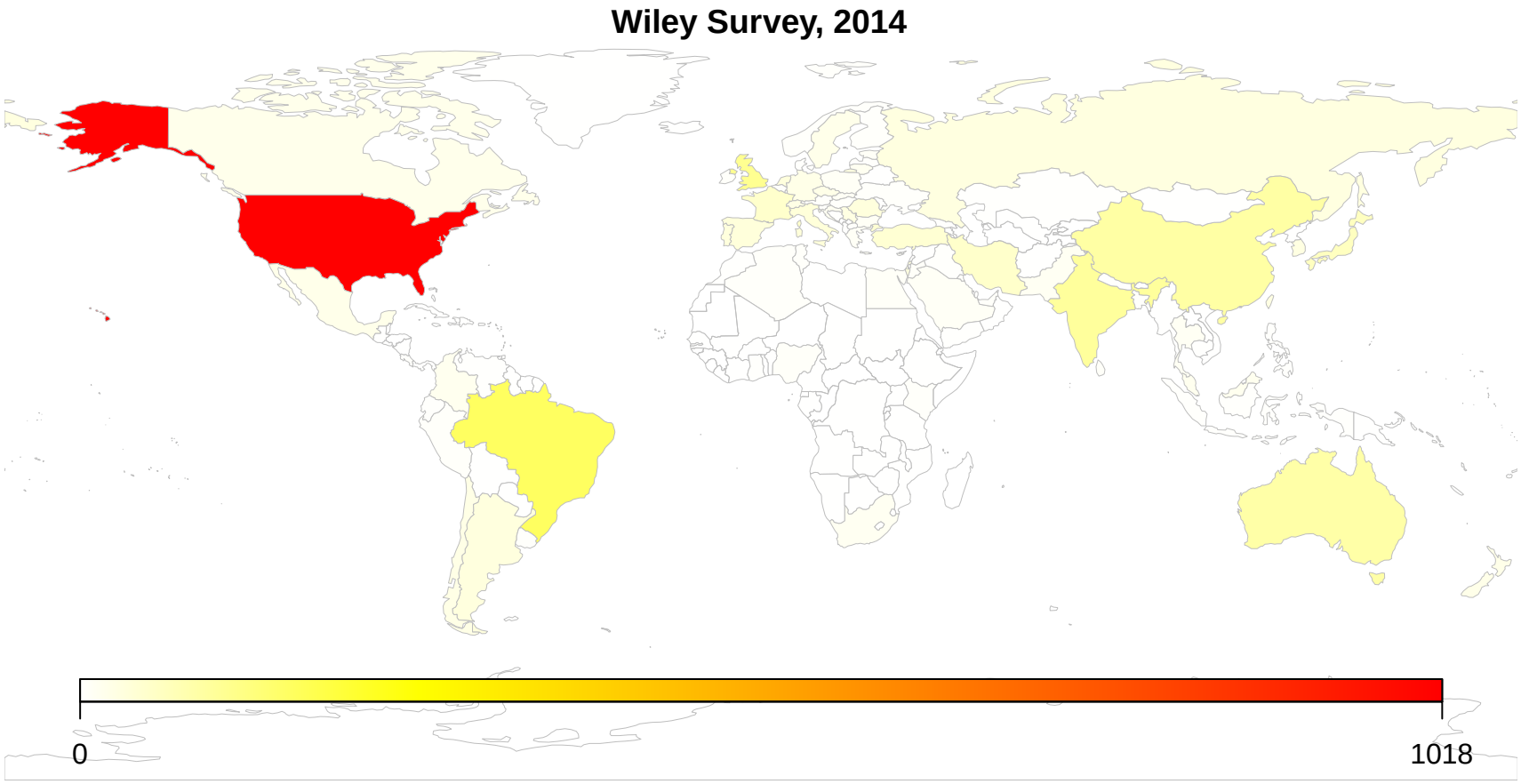

Figure A9: Heat-map distribution of survey respondents by country for the Wiley survey conducted in 2014

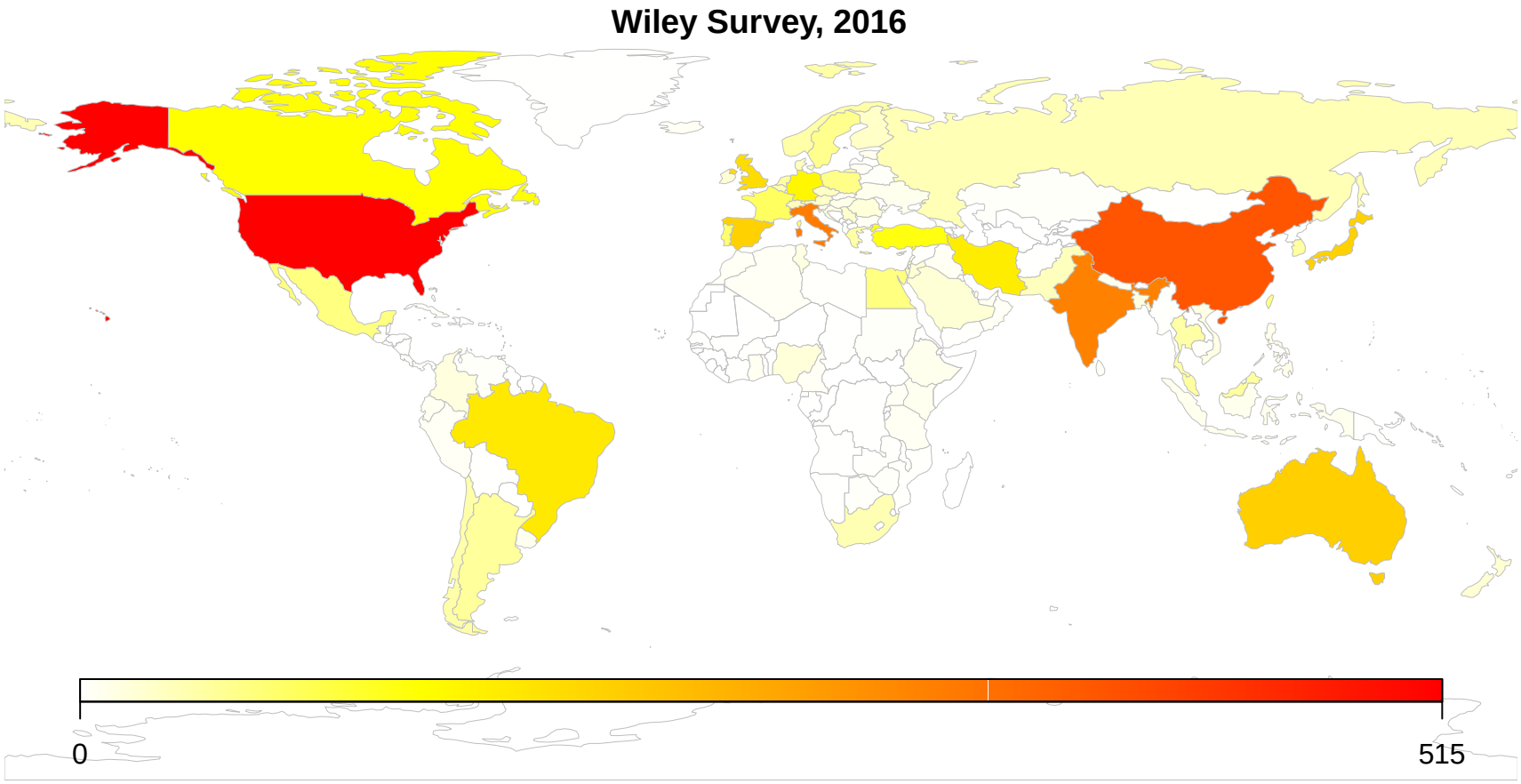

Figure A10: Heat-map distribution of survey respondents by country for the Wiley survey conducted in 2016

Example code to obtain number of responses per country and produce heat map

Code used to extract the number of responses per country. Here we use the data from the Elsevier survey as an example of how the data was compiled for Table A1.

```
### 1.1 clear memory

rm(list=ls())

ls()


### 1.2 set directory and data frame

### replace “Name” and “Folder_name” with your user name and the ### folder from which you are ###working on the desktop

### here we have loaded the Elsevier data as the file

### “Elsevier_2017.csv”

### note, that the name of the column with the country data ### needs to be named or re-named “Country” for the code to work

getwd()

setwd("/Users/Name/Desktop/Folder_name ")

mydata<-read.csv("Elsevier_2017.csv", header = TRUE)


###1.3 check data is loaded properly

mydata


###1.4 set data up to analyze number of responses per country using the “plyr” package

library(plyr)

country_table<-count(mydata, "Country")


###1.5 export country table to the directory to compile in Excel

write.table(country_table, "/Users/Daniel/Desktop/Open_data/Elsevier_country.txt", sep="\t")
```

This second snippet of code is what is used to generate the heat maps from the master file included in Table A1. Here we will use the Elsevier dataset again to demonstrate.

```
### 2.1 clear memory

rm(list=ls())

ls()


### 2.2 set directory and data frame

### replace “Name” and “Folder_name” with your user name and the folder from which you
### are working on the desktop

### in this case we are loading the data file containing all responses provided in Table ### A1,
note that thie file was edited in advance to verify for countries that were not ### compatible
with the coding used in rworldmap

### Note that the number of responses per country for the Elsevier survey have been coded
### as “Elsevier_2017”

getwd()

setwd("/Users/Name/Desktop/Folder_name")

mydata<-read.csv("country_compiled_edited.csv", header = TRUE)


### 2.3 run map package using 3 digit ISO code

library(rworldmap)

sPDF<-joinCountryData2Map(mydata,joinCode ="ISO3", nameJoinColum= "ABREV")



### 2.4 generate heat map

### parameters in this script can be changed to fit your needs

### for several maps, copy and paste this script and chains mapTitle and numCats

### also change nameColumnToPlot to match desired dataset

### for highest resolution on legend, numCats should equal max sample size + 1
```

```
mapDevice()
mapParams<-mapCountryData(sPDF, nameColumnToPlot="Elsevier_2017", addLegend=FALSE,
                          catMethod = "fixedWidth", numCats = 243, colourPalette = "heat",
                          mapTitle="Elsevier Survey, 2017")
do.call(addMapLegend, c(mapParams, legendWidth = 0.5,
                        legendLabels = "limits", legendMar = 2))
```

A Methodological Note on Interoperability for Appendices B and C

To analyze geographic and disciplinary trends in data sharing, our first step was to obtain all the datasets and questionnaires and begin by recoding results in the open source software R. While importing the data, issues arose in terms of how the files were formatted, which affected their readability upon conversion to a comma delimited format often used to avoid formatting issues. In most cases and to avoid further issues in analyzing the data, all data filtering in the original MS Excel files and text formatting were removed. To simplify analyses further, column names for the variables of interest were recoded to simple words to facilitate coding in R and made consistent across all datasets (e.g. “Country” for country of respondent, “Discipline” for columns asking about field of study/research, and “Sharing” for the column containing the recoded data for the questions shown in Table 2).

In some cases, such as the survey sponsored by the Belmont Forum, respondents could respond with free text making it laborious to apply a consistent coding scheme. In others, such as the Elsevier survey, responses to the various options were coded logistically (1 for a “yes” to one of the options, 0 for a “no”) whereas others, like the State of Open Data, included the actual text options. This required that logical arguments be developed in R that were specific to each data set to generalize these responses to a yes or no according to the rules in **Table 2**. Furthermore, in many cases columns needed to be combined, scanned for blank entries, which then had to be removed. As such, the .csv files that were eventually analyzed were heavily edited. Below, we have included examples of codes employed in RX.

Appendix B: Country bar graphs

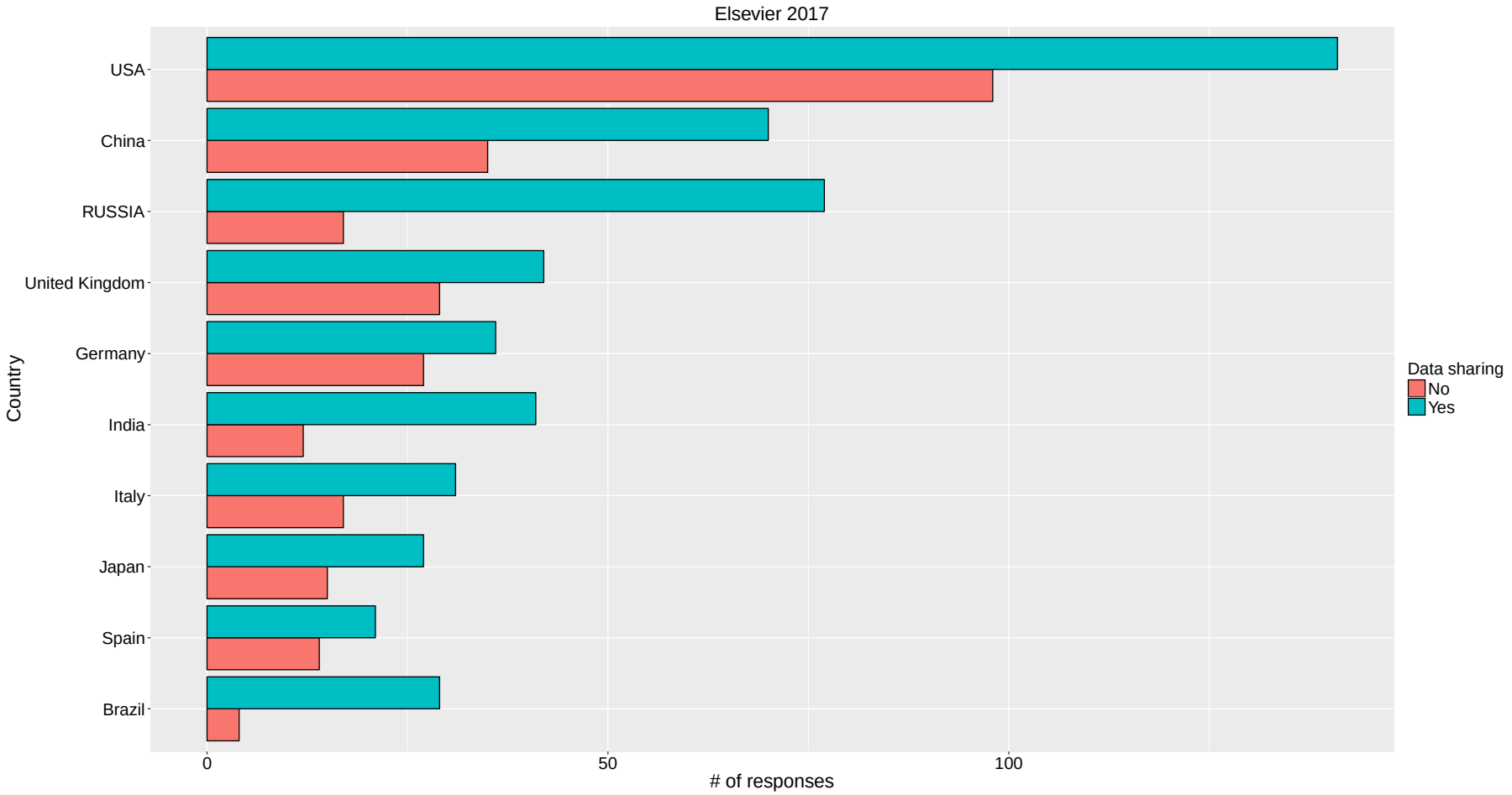

Figure B1: Data sharing across disciplines and through time for Elsevier CWTS survey conducted in 2017

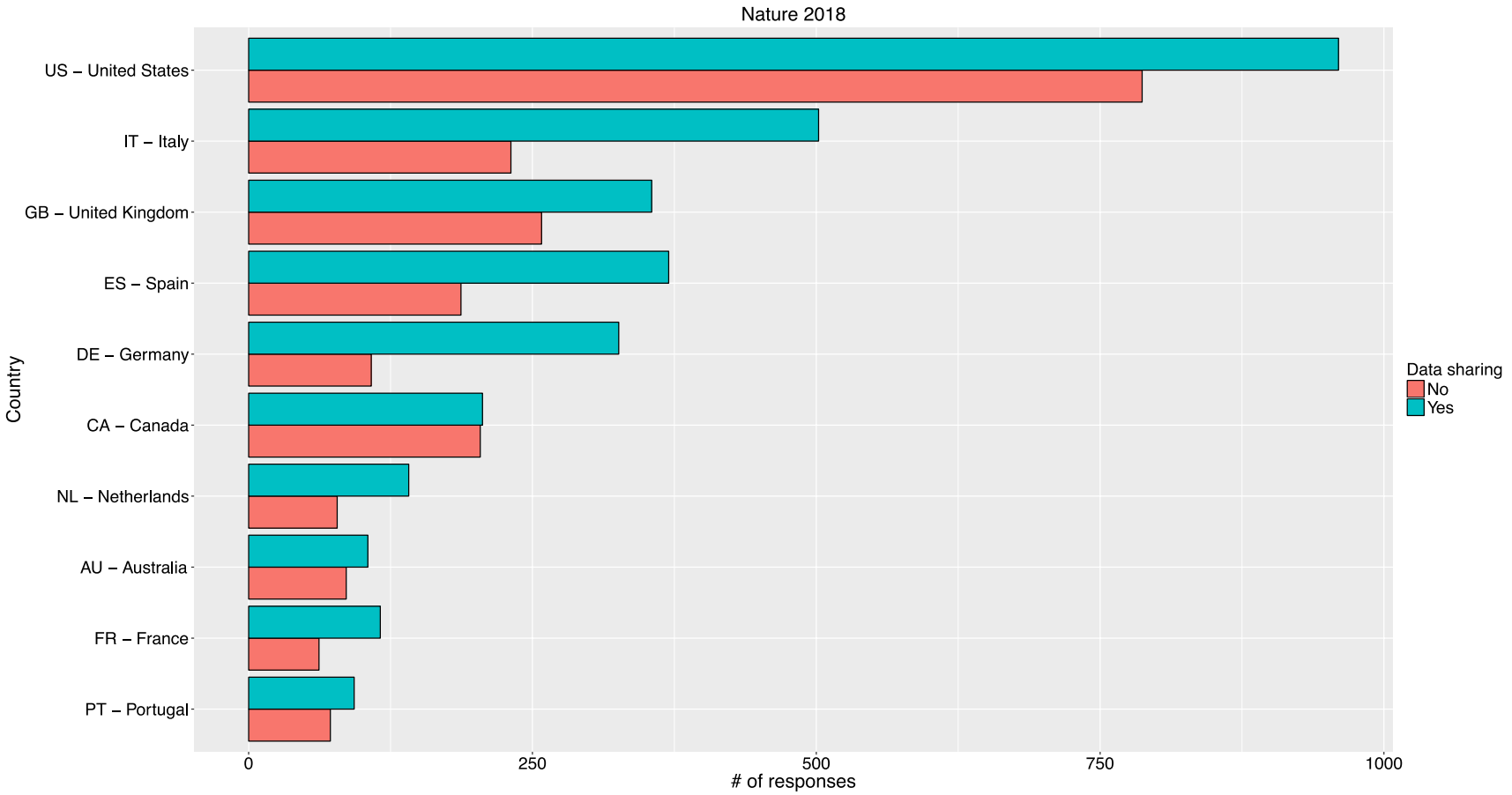

Figure B2: Data sharing across disciplines and through time for the SpringerNature survey conducted in 2018

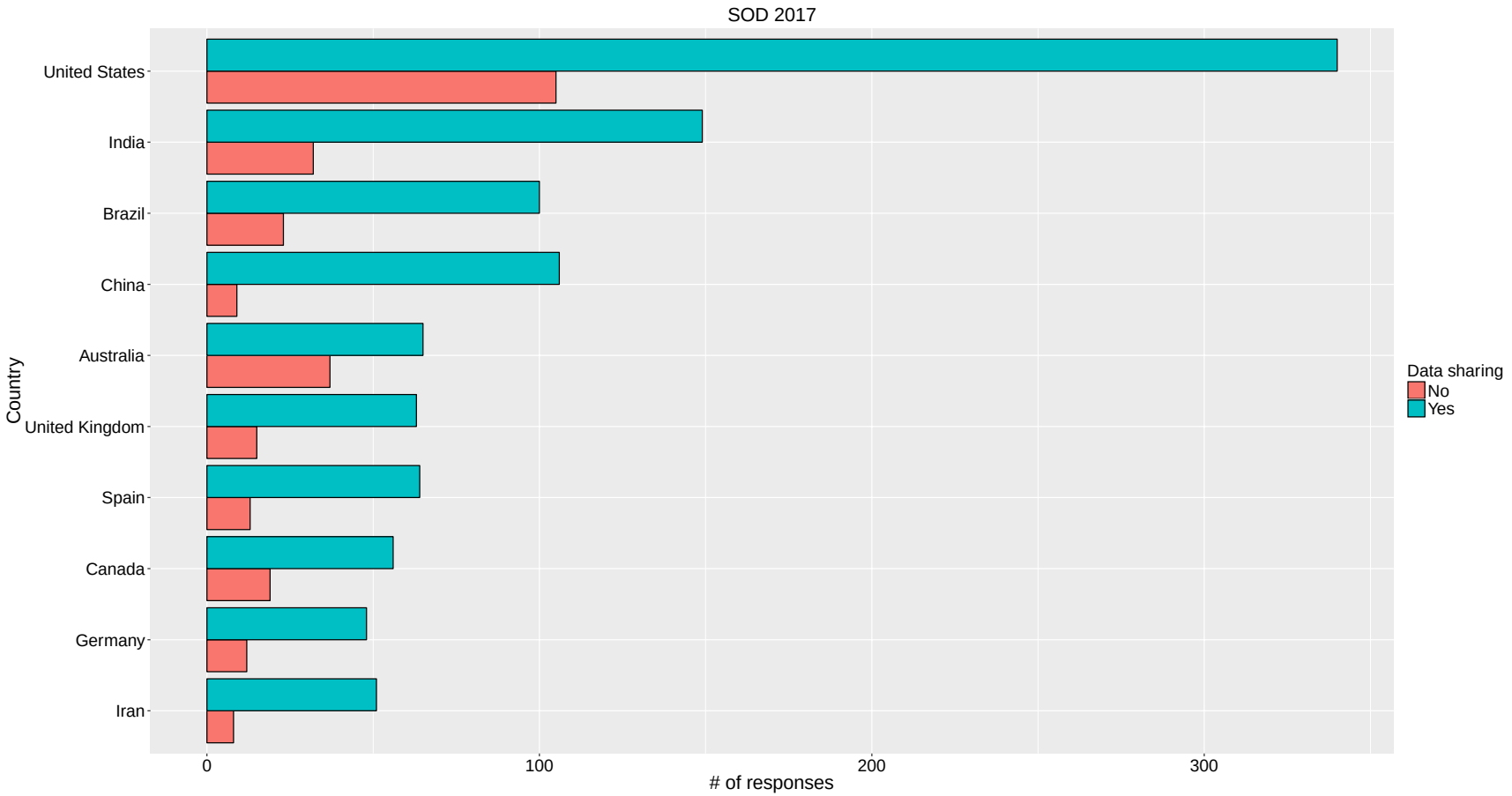

Figure B3: Data sharing across disciplines and through time for State of Open Data survey conducted in 2017

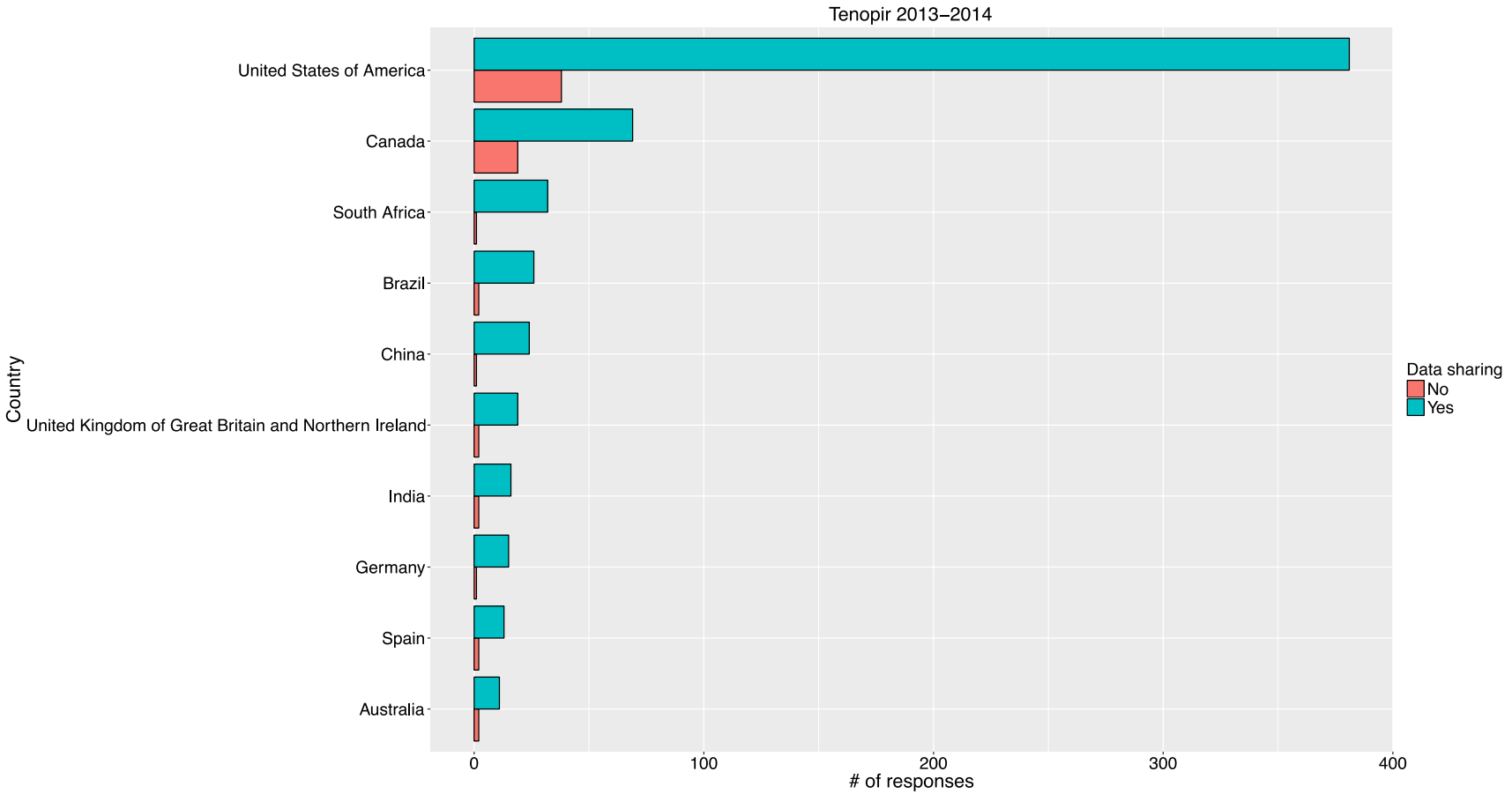

Figure B4: Data sharing across disciplines and through time for survey conducted by Tenopir et al. in 2013-2014

Example code for country bar graphs

This code shows how the bar graphs were created to highlight country differences in data sharing. Note that given issues with interoperability between surveys, some manual editing of variables names was done in Excel prior to working in R for convenience. These edits are noted in the script where pertinent. This code snippet employs packages associated with R's tidyverse to facilitate data manipulation and visualization.

```
### clear memory

rm(list=ls())


### three packages are required to modify the data and plot the ### bar graphs showing data sharing across disciplines

library(plyr)

library(dplyr)

library(ggplot2)


###define directory to work from

getwd()

setwd("/Users/Name/Desktop/Folder_name")


### Here we use the Elsevier 2017 dataset as an example

### note that the data column for the responses to whether data ### is shared or note need to be called "sharing" and the column ### for country needs to be called "Country"

Elsevier_2017<-read.csv("/Users/Daniel/Desktop/Open_data/Elsevier_2017.csv",

                        header = TRUE)


### remove blanks prior to encoding new variable "Data_public"

Elsevier_2017<- subset(Elsevier_2017, sharing != "")
```

```
Elsevier_2017<-droplevels(Elsevier_2017)
Elsevier_2017<- subset(Elsevier_2017, Country != "")
Elsevier_2017<-droplevels(Elsevier_2017)

### convert answers to Yes/No based on conditions set prior
Elsevier_2017$Data_public<-ifelse(Elsevier_2017$sharing == 1, c("No"), c("Yes"))

### count the frequency of yes vs no responses by country
Elsevier_2017_counts<-count_(Elsevier_2017, c('Country', "Data_public"))

#### use tidyverse to create variable for total country count
#### create new data frame take top 20 based on total country ### count
Elsevier_2017_top20 <- Elsevier_2017_counts %>%
  group_by(Country) %>%
  mutate(total_count = sum(n)) %>%
  ungroup() %>%                    ## remove grouping for subsequent graphing
  arrange(desc(total_count)) %>%
  ungroup() %>%
  slice(1:20)

### graph new data reordering by total count and using yes/no for fill
ggplot(Elsevier_2017_top20,
       aes(x = reorder(Country, total_count), y = n, fill=Data_public)) +
  geom_bar(position="dodge", stat="identity", color= "black") +
  coord_flip() +
  ggtitle("Elsevier 2017") +
  xlab("Country") + ylab ("# of responses") + labs (fill ="Data sharing") +
```

```
theme(plot.title = element_text(color = "black", size = 28),
      axis.title.x = element_text(color = "black", size = 26),
      axis.text.x = element_text(color = "black", size = 26),
      axis.title.y = element_text(color = "black", size = 26),
      axis.text.y = element_text(color= "black", size = 26),
      legend.title = element_text(color = "black", size = 26),
      legend.text = element_text(color = "black", size = 26))
```

Appendix C: Discipline bar graphs

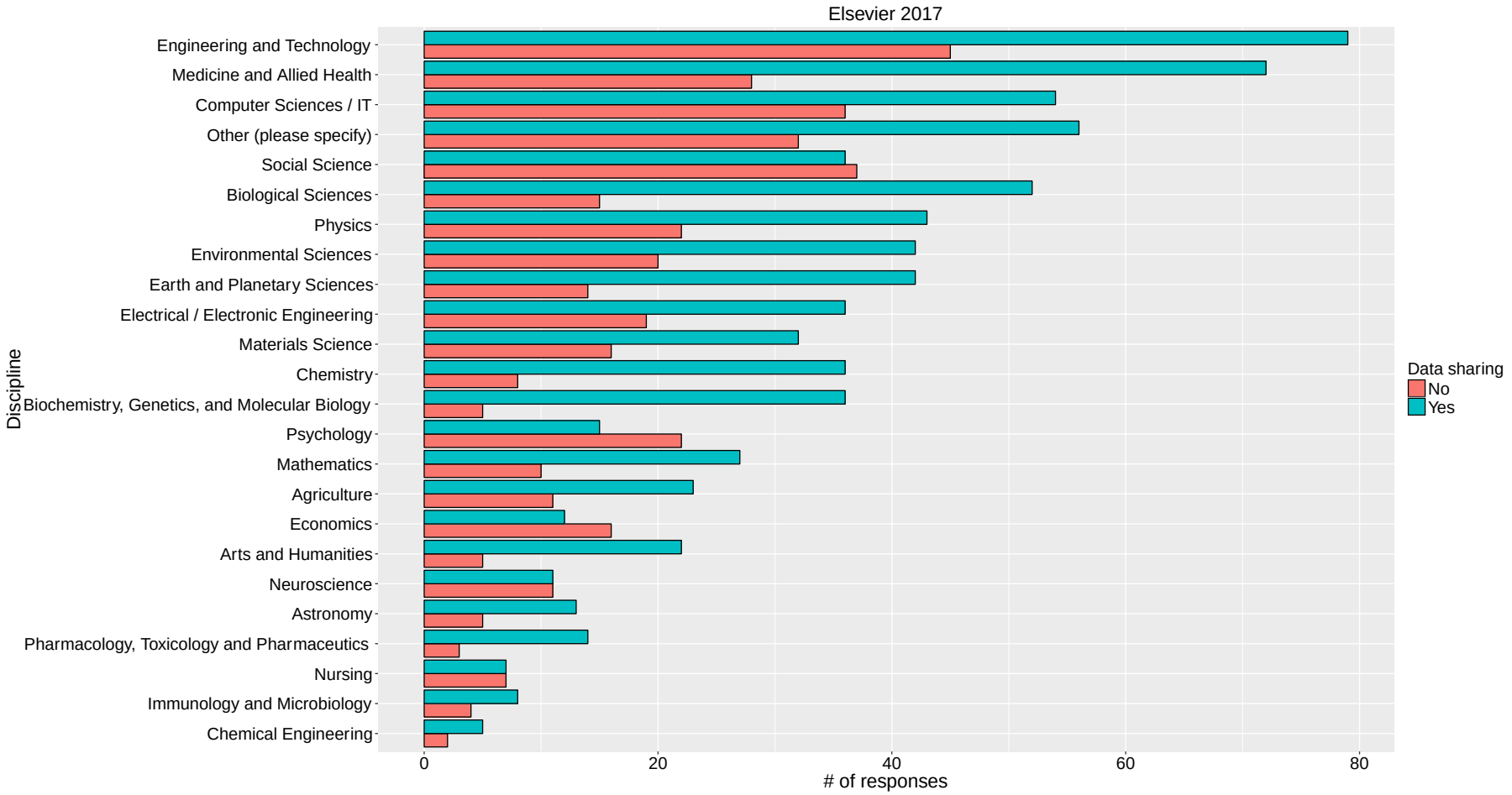

Figure C1: Data sharing across disciplines and through time for Elsevier CWTS survey conducted in 2017

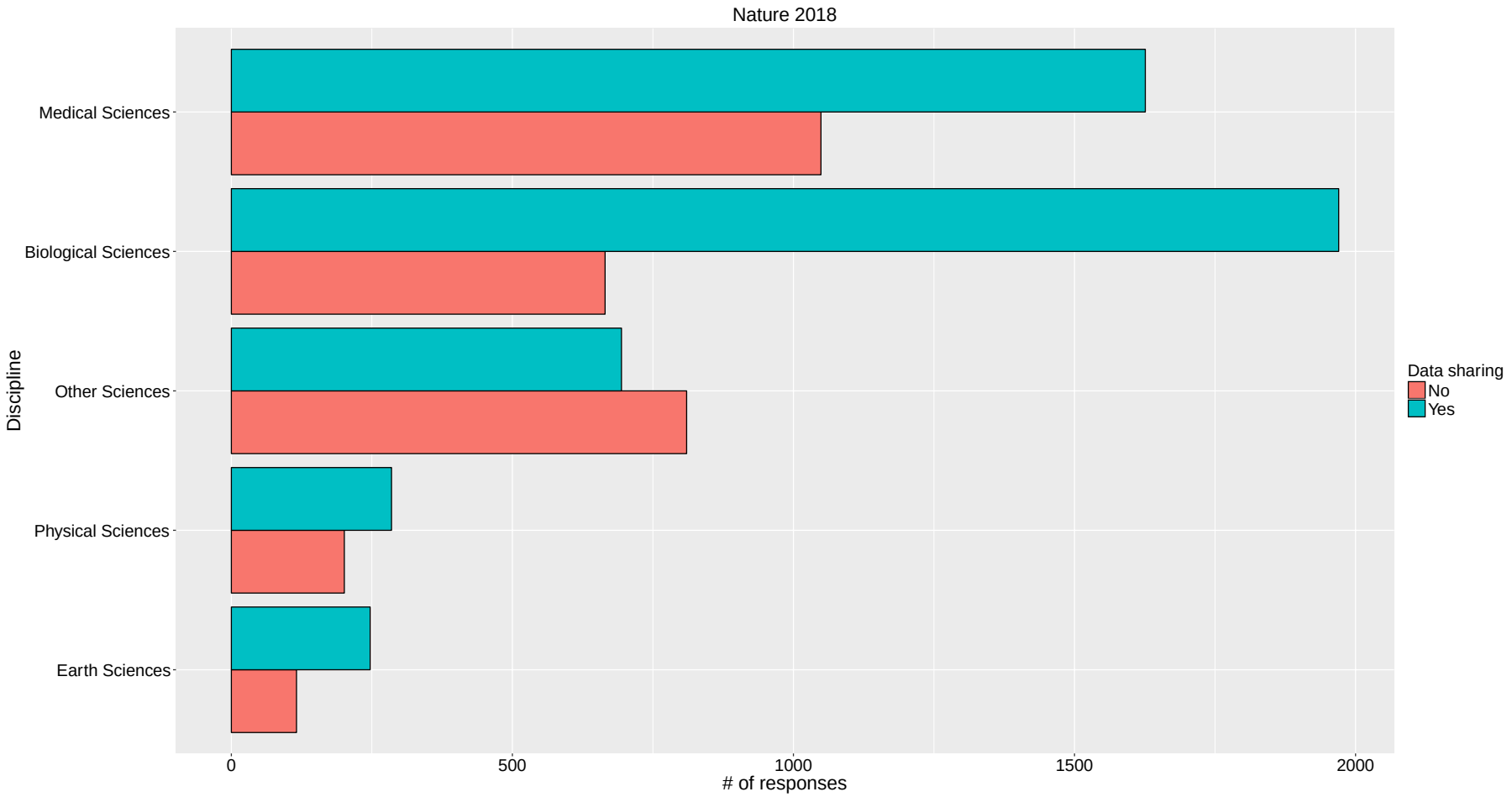

Figure C2: Data sharing across disciplines and through time for SpringerNature survey conducted in 2018

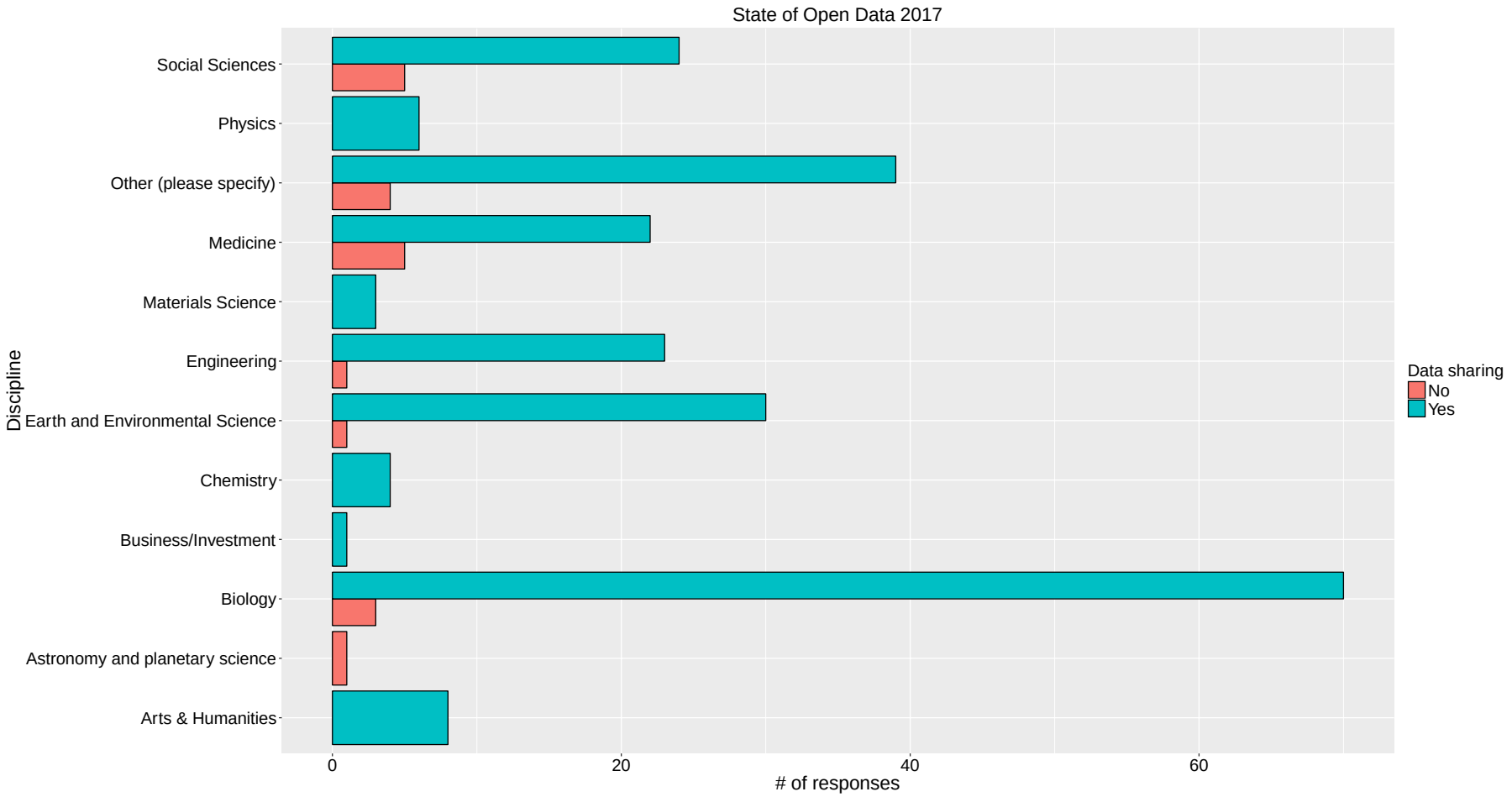

Figure C3: Data sharing across disciplines and through time for State of Open Data survey conducted in 2017

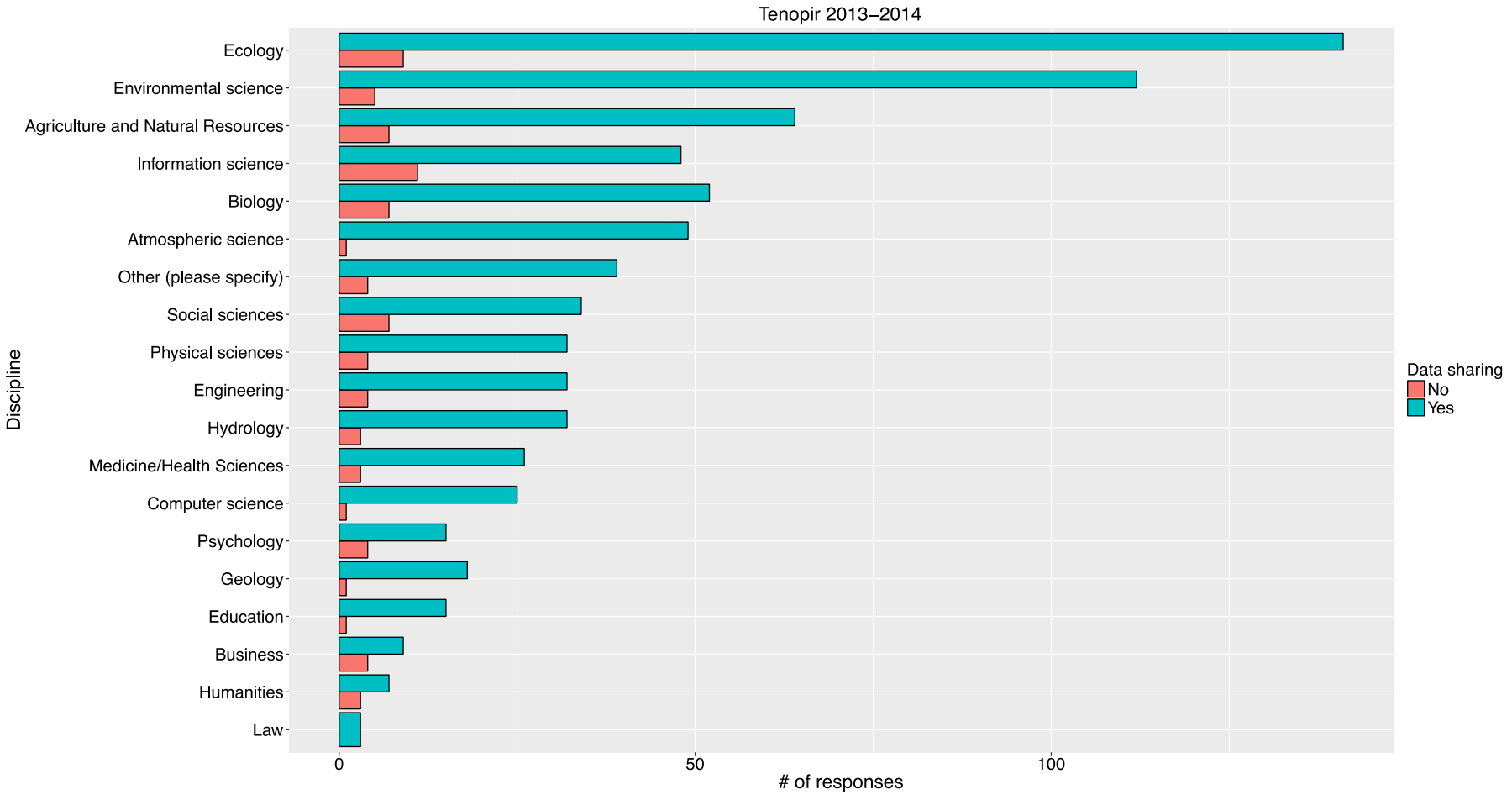

Figure C4: Data sharing across disciplines and through time for survey conducted by Tenopir et al. in 2013-2014

Example code for discipline bar graphs

This code shows how the bar graphs were created to highlight differences in data sharing across discipline. Note that given issues with interoperability between surveys, some manual editing of variables names was done in Excel prior to working in R for convenience. These edits are noted in the script where pertinent.

```
### clear memory

rm(list=ls())


### three packages are required to modify the data and plot the ### bar graphs showing data sharing across disciplines

library(plyr)

library(dplyr)

library(ggplot2)


### define directory to work from

### change “Name” and “Folder_name” to desired titles to work ### from desktop

getwd()

setwd("/Users/Name/Desktop/Folder_name")


####Elsevier 2017

### here we will use the Elsevier dataset again as it needs to ### be modified based to highlight what constitutes a “Yes” vs ### “No”

### the column containing responses to the question pertaining ### to data sharing needs to be called “sharing”

### the column showing country information needs to be called ### “Discipline”

Elsevier_2017<-read.csv("/Users/Daniel/Desktop/Open_data/Elsevier_2017.csv",

                        header = TRUE)
```

```
### remove blanks prior to encoding new variable

Elsevier_2017<- subset(Elsevier_2017, sharing != "")

Elsevier_2017<-droplevels(Elsevier_2017)

Elsevier_2017<- subset(Elsevier_2017, Discipline != "")

Elsevier_2017<-droplevels(Elsevier_2017)


### need to convert answer options for “sharing” to Yes/No based ### on the conditions established prior

### in this case, results are encoded as a 1 or 0 (logistic) to ### indicate whether respondent chose an option

### all answers compiled into one column prior to changing

### variable name to “Data_public”


Elsevier_2017$Data_public<-ifelse(Elsevier_2017$sharing == 1, c("No"), c("Yes"))


### use count_ function to get frequency of Yes vs No for each ### discipline

Elsevier_2017_counts<-count_(Elsevier_2017, c('Discipline', "Data_public"))


### graph using ggplot2 package

ggplot(Elsevier_2017_counts,

      aes(x = reorder(Discipline, n), y = n, fill=Data_public)) +

  geom_bar(position="dodge", stat="identity", color= "black") +

  coord_flip() +

  ggtitle("Elsevier 2017") +

  xlab("Discipline") + ylab ("# of responses") + labs (fill ="Data sharing") +

  theme(plot.title = element_text(color = "black", size = 28),

       axis.title.x = element_text(color = "black", size = 26),

       axis.text.x = element_text(color = "black", size = 26),
```

```
axis.title.y = element_text(color = "black", size = 26),
axis.text.y = element_text(color= "black", size = 26),
legend.title = element_text(color = "black", size = 26),
legend.text = element_text(color = "black", size = 26))
```