Surveying Scientists About Their Software

Drew Paine
Human Centered Design & Engineering, University of Washington
pained@uw.edu

Justin M. Woodum
Human Centered Design & Engineering, University of Washington
jwoodum@uw.edu

Charlotte P. Lee
Human Centered Design & Engineering, University of Washington
cplee@uw.edu

June 24, 2014

HUMAN CENTERED DESIGN & ENGINEERING TECHNICAL REPORT
HCDETRS_2017_2

Surveying Scientists About Their Software

A CSC Laboratory technical report by

Drew Paine, Justin M. Woodum, Charlotte P. Lee

2014

Computer Supported Collaboration (CSC) Laboratory

https://depts.washington.edu/csclab/techreports/

Department of Human Centered Design & Engineering

University of Washington

*This material is based upon work supported by the National Science Foundation under **Grant Number ACI-1302272**. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation, the Department of Human Centered Design & Engineering, or the University of Washington.*

Abstract

Software is an important infrastructural component of scientific research practice. The work of research often requires scientists to develop, use, and share software in order to address their research questions. This report presents findings from a survey of researchers at the University of Washington in three broad areas: Oceanography, Biology, and Physics. This survey is part of the National Science Foundation funded study *Scientists and their Software: A Sociotechnical Investigation of Scientific Software Development and Sharing* (ACI-1302272). We inquired about each respondent's research area and data use along with their use, development, and sharing of software. Finally, we asked about challenges researchers face with and about concerns regarding software's effect on study replicability. These findings are part of ongoing efforts to develop deeper characterizations of the role of software in twenty-first century scientific research.

Recommended Citation:
Paine, D., Woodum, J.M. and Lee, C.P. (2014). Surveying Scientists About Their Software. (CSC-2014-02) Computer Supported Collaboration Laboratory Technical Reports: University of Washington.

Table of Contents

1 Introduction

Software is a major component of the underlying infrastructure of scientific research practice today. In this report we examine the findings of our 2013 survey of 55 University of Washington researchers about the role of software in their research practice. The University of Washington, located in Seattle, WA, is one of the United States' leading research institutions. It is regularly ranked as a top public university in the US, as well as among the top 20 internationally. Furthermore, the University of Washington is ranked number one among public universities in the US in the receipt of federal research and training funding [1].

The survey we report on here was specifically designed to examine the role of software in the research work of University of Washington scientists in three broad areas: Oceanography, Biology, and Physics. We sent our survey to Principal Investigators to obtain descriptive data about their research group composition, data use, and principally, their use, development, and sharing practices with software. In addition, we inquired about challenges they and their groups face with software and any concerns they have about software affecting the replicability of studies. Before discussing our findings, we briefly discuss the larger context that this study and the survey's findings fit within.

2 Survey Context

Scientific research and software have advanced together since the advent of computers. The growth of Big Science [2] and, today, Big Data or Data-Intensive Science [3] is ever more reliant upon software as a means of collecting, processing, analyzing, and disseminating data and scientific findings. Today, much research and development of software for scientific research in the United States takes place under the banners of Cyberinfrastructure (CI) or eScience. Funding agencies such as the National Science Foundation and the National Institutes of Health, along with private entities such as the Gordon and Betty Moore Foundation, have spent the past decade and a half expending significant effort on the development and sharing of software to support scientists' research work [4,5].

Scholars in the Computer Science and Software Engineering (CSSE), Computer Supported Cooperative Work (CSCW), and Information Science fields examine the development and use of software by scientific researchers. Hannay et al. [6] surveyed approximately 2,000 scientists, researchers, and software developers about their software work as a part of research practice. Hannay et al. determined that 91.2% of their respondents find using scientific software to be an important or very important piece of their research. Furthermore, they found that 84.3% of their respondents found *developing* software to be an important part of their research. Examining the processes of software development in scientific research, Segal [7] notes the iterative and ad-hoc nature of the work. Finally, Howison and Herbsleb [8] examine the incentives scientists do or do not have to develop software as part of their research work. Crucially, Howison and Herbsleb note that *we do not have a clear understanding of how scientific software development work fits into everyday scientific research practice*.

Our National Science Foundation funded study *Scientists and their Software: A Sociotechnical Investigation of Scientific Software Development and Sharing*[1] aims to examine the adoption, use, development, and sharing of software as an integral part of research practice. Our study aims to help close the gap in our understanding of software's role in scientific research pointed to by Howison and Herbsleb, among many others. This study is a multi-year investigation of the

[1] See https://depts.washington.edu/csclab/projects/scientists-software/ for more detail.

research and software practices of scientists in three focal areas. We aim to investigate how researchers are engaging in these software related activities in the course of their day-to-day research practice. The survey that we report on here is only one step in our larger investigation, setting the stage for later ethnographic work.

3 Overview of the Survey

Our survey consisted of 43 questions, including two administrative questions. This survey was designed to help us answer our overall study's research questions while beginning to characterize the landscape of software in the research practice of our three focal fields. We focused on inquiring about each respondent's software use, development, and sharing in their research. In addition, we inquired as to challenges they may face with software in their work and concerns they may have about software impacting the replicability of studies. The specific areas of a scholar's work that we focused on are[2]:

- *Background* – questions about their role in the UW community
- *Research Group* – questions about their group's size and composition, data use, and level of computational training of group members
- *Software Use* - questions about their group's use of software for data collection/creation, processing, and analysis along with criteria for selecting any given piece
- *Software Development* – questions about their group's development, or customization if not developing, of software
- *Software Sharing* – questions about their software sharing practices, if applicable
- *Wrap-Up* - questions about challenges with software in their research and any concerns about study replicability

These focal areas provide us with a broad overview of each researcher's group and work while delving into their software use, development, and sharing practices. For the majority of our questions we used free response text entries to allow our respondents to share as much detail as possible. Where appropriate multiple-choice list or Likert-scale questions were used. For example, when asking about the computational training of group members we used Likert-scales. Our inquiries regarding sources of data being used and entities software is obtained from or shared with used multiple choice lists.

We developed a sample of 208 researchers from the University of Washington School of Oceanography, as well as the Departments of Microbiology, Genome Sciences, Physics, and Astronomy. We included all "active" researchers in our sample. We defined active researchers to include professors - both tenure track and research track - of all ranks. Researchers who were emeritus - or who were indicated as not taking students - were not included. In addition, affiliate or adjunct faculty, visiting faculty, research scientists, and post-doctoral researchers were all excluded since the aims of our grant were to sample Principal Investigators leading their own lab whenever possible. Four researchers were removed from the survey sample once it was distributed - two were no longer University of Washington members and two specifically requested that they be removed. This left us with a sample of 204 potential respondents.

Once both the sample and survey were developed, it was distributed using a University of Washington Catalyst WebQ survey (a web-based survey that is distributed via e-mail). The survey was available between November 18, 2013 and December 18, 2013. A link to the survey was distributed via an automated email message and reminders were sent three times. A total

[2] A full list of the survey questions is available in the Appendix.

of 56 responses were captured during this time period. One of the respondents did not however take the survey, leaving us with data from 55 individuals - a 27% response rate (55 of 204). Each participant was informed of their rights as a research subject in accordance with University of Washington Human Subjects Division rules. Furthermore, each researcher was offered compensation for their time in the form of a $10 coffee gift card.

The findings discussed here offer commentary regarding the 55 responses that we received. We further examine the demographics of those who did respond in our first findings sub-section. Questions that were asked using free response text entry boxes were qualitatively coded to group similar responses into categories [9]. When applicable, quantitative information regarding responses to our questions is provided. We note that this survey was not designed for statistical analyses to be performed.

4 Findings

Below we present a picture of the state of research in three broad areas at the University of Washington in 2013. In our findings sections, we examine the demographics of our respondents before discussing selected findings regarding the use, development, and sharing of software in the research examined. We omit some findings that would either compromise the confidentiality of our informants or require further inquiry before being reported.

4.1 Background of our Respondents

To develop a characterization of the researchers who responded, we asked a set of questions about their research and group overall (see Q2-Q18 in the Appendix). Here, we examine the overall demographics of our respondents, a bit about their data use, and finally some findings on the computational training needs of members of their group. We do not report findings regarding the specific research areas or compositions of groups here to protect the confidentiality of our informants.

4.1.1 Demographics

Our survey sample encompassed members of three broad research areas: Oceanography, Biology, and Physics. Examining the home department each researcher indicated, the respondents were spread as seen in Table 1 across the three areas. Within Oceanography and Physics we received responses from at least 30% of the potential respondents. The percentage of responses from researchers in Biology was less than 20% due to the high number of potential respondents in that area at the University of Washington; this group had the second most responses to the survey in total. We do not note the names of the specific researchers, research groups, or their specific research areas here to protect the confidentiality of our informants.

Table 1. Number and percentage of respondents for each group (Q4).

	Total Number	Percentage
Oceanography	15	27.27% (15/55)
Biology	18	32.72% (18/55)
Physics	22	40.00% (22/55)

4.1.2 Data Use

In addition to background information about a researcher's discipline, work, and group, we inquired as to their data use. With one question, we asked the respondent to simply list the types of data being used in their work. In a second question, researchers were asked to select the sources of data that they use from a multiple-choice list. These sources include: Experiments, Fieldwork, Sensor Systems, Community Databanks, Simulations, Publications,

and Other where an additional response could be entered. These categories emerge from our prior studies of scientific research [10].

The answers that researchers listed for the data that they produce or collect varied by research area and among the researchers in a particular area themselves. Examples include seawater concentrations, biochemical assays, DNA/RNA sequences, sensor outputs from photodiodes, and astronomical catalogs. Examining the responses to our second question about data, we see in Table 2 that all categories are well represented. Fifty-four of our 55 respondents selected at least one category. The one respondent who did not answer this question is a theoretician and does not use empirical data in their research. Among those who answered "Other", one researcher noted "Native Communities" as a source of data, another: student test scores, another: patients from clinical situations, and two in Physics noted telescopes (although we would have categorized such instruments as Sensor Systems ourselves in the categorization scheme provided).

Table 2. Categories of data that are used in researchers work (Q9).

Source of Data Categories	# of Respondents
Experiments	35
Fieldwork	15
Sensor Systems	17
Community Databanks	27
Simulations	17
Publications	17
Other	8

4.1.3 Formal Computational Training of Group Members

The final aspect of each researcher's group that we inquired about was the computational training that they felt that the members of their group require to do their work. We asked two questions about the likelihood of a member of a given type already having formal computational training when they join the group and the likelihood that they will need to obtain formal computational training as a member of the group. In addition, we asked researchers to list the resources members might use to obtain such formal training if it is needed.

Looking across the answers to these two questions, we see that Post-Doctoral Researchers, Research Scientists, and Doctoral Students are most likely to already have formal computational training as a part of their work. Respondents also indicated that researchers of these experience levels would be most likely to need to obtain formal training if they did not already have it. At least 50% of the responses for each of these categories for either question were for Somewhat Likely or Very Likely. In contrast, respondents more often indicated less likelihood for Undergraduate Students already having or needing to obtain formal computational training to participate in research work. We found this to be an interesting point to note and further examination of the types of work undergraduates are engaged in would be worthwhile. It might be expected that the tasks assigned to undergraduate researchers in some fields are less computationally intensive, requiring less formal training. Ascertaining this and understanding how this is evolving would be of relevance to examining the training of future researchers. Of note here is the low number of responses regarding Masters students overall when compared with the number of responses for other categories of group members. While not confirmable using these responses alone, it appears likely that the researchers who responded to our survey do not have many members who are Masters students.

Table 3. Likelihood that a member will *already have* formal computational training when they join the group (Q16).

	Post-Doctoral Researchers (49 responses)	Research Scientists (47 responses)	Doctoral Students (50 responses)	Masters Students (38 responses)	Undergraduate Students (47 responses)
Very Likely	22	18	11	4	4
Somewhat Likely	14	11	17	16	14
Somewhat Unlikely	11	14	17	12	18
Very Unlikely	2	4	5	7	12

Table 4. Likelihood that a member will *need to obtain* formal computational training while they part of a group (Q17).

	Post-Doctoral Researchers (50 responses)	Research Scientists (45 responses)	Doctoral Students (52 responses)	Masters Students (37 responses)	Undergraduate Students (47 responses)
Very Likely	15	14	18	13	10
Somewhat Likely	10	9	19	7	14
Somewhat Unlikely	14	12	11	11	13
Very Unlikely	11	10	5	7	11

In addition, we asked respondents to list the resources members of their group might use to obtain formal computational training. Fifty of the 55 respondents provided answers. The common resources that were provided include online, coursework as part of a specific discipline's training, formal Computer Science courses, distinct workshops, other group members, and one-on-one mentoring from the Principal Investigator. Other resources that were mentioned less frequently include books or manuals, webinars, just going ahead and using software while learning on the job, or no formal training.

4.2 Software Use

The next section of our survey inquired about each respondent's use of software in their research. We were interested in finding out the various pieces of software used and developed for data collection and/or production, processing, and analysis work. In addition, we inquired about each researcher's criteria for selecting software for each type of research activity. From our findings we see a diverse ecosystem of software being used and developed in the course of scientific research.

The details regarding the software that is used to collect and/or produce, process, and analyze data are highly variable depending on the type of research being done and the specific goals of a researcher and their group. Software involved in data collection and/or production is commonly a set of scripts or hardware-specific software to pull data from databases or instruments. Many respondents did also indicate using more extensively developed software that is built using programming languages such as C/C++ and Python, or other more specialized software development frameworks such as Interactive Data Language (IDL) and MATLAB.

Similar software environments are used in the processing and analysis of our respondents' data. Additional languages and programs such as R, Excel, gnuPlot, and PRISM are also mentioned, however, for processing and analysis work.

We find four prevalent themes when examining the criteria respondents offer for their selection of software to use or develop. These themes include:

1. How widely adopted a piece of software or toolset is within the specific scientific community
2. Whether or not the features of the software meet their research needs
3. Whether or not the software fits within their budget
4. The expediency with which the team can learn and use it.

These criteria are broad methods for selecting and using software in work practice. Researchers expressed a need for reliable and readily usable software that will not require a huge learning curve for them and their group. Cost is a factor since funding is limited across the research world. In addition, based on the responses we received, the adoption of software by a community appears to be a mechanism many researchers rely upon to easily find and adopt software in their work. Developing a deeper and more comprehensive understanding of the criteria of selection and other decisions researchers make when developing, adopting, using, and sharing their software is the primary aim of our project overall.

Finally, we inquired about the sources researchers obtain software their from. We wished to know if they were obtaining software from four different entities that they might collaborate with. The four entities were: 1) Local Research Group, 2) Project Collaborators, 3) Community and/or Discipline, and 4) the Public. These entities were chosen to account for different types of potential collaborations.

Overall the majority of responses for each category were Yes, see Table 5 below. It is interesting to note that 18 respondents indicated No or Not Applicable to using software from their own group. This mostly aligns with the 16 responses in our next section of the survey indicating that the respondent's group does not develop software. It is also interesting to note the high percentage (66.67%) of respondents answering Yes to using software from project collaborators.

Table 5. Whether software is obtained from each source or not, with 54 of 55 answering (Q25).

	Local Research Group	Project Collaborators	Community and/or Discipline	Public
Yes	36	36	44	38
No	15	17	8	10
Unsure	0	0	1	4
Not Applicable	3	1	1	2

4.3 Software Development

A key goal of this project is studying researchers and their groups who are developing and in turn sharing their own software. We therefore asked whether each respondent and their group is developing software or not. Thirty-nine of the 55 respondents answered Yes, distributed as seen in Table 6. We followed this question up by inquiring as to what software is being developed,

why the researcher and their group found it necessary to develop such software, and whether or not the researcher was aware of comparable software that the group could use in place of what they are developing.

Table 6. Whether a respondent's group is developing software or not (Q26).

	Oceanography	Biology	Physics
Yes	11	11	17
No	4	7	5

The software that is being developed by the scientists who responded to our survey varies from ad hoc scripts for one-off data processing to full data processing pipelines or frameworks for modeling phenomena. Commonly mentioned programming languages include C/C++, Python, Perl, R, Mathematica, and MATLAB.

Three themes arise when examining the free responses regarding why the group develops the software that they do. The first theme that arises is the *goal of the research necessitating the development of the software*. Responses grouped in this theme emphasized the research necessitating the implementation of a new algorithm or other concept in software to advance the research goal. The second common theme is the *lack of capabilities in existing software for the research task at hand*. This theme connects with the first since both capture a core motivation for scientist's to develop software as a part of their research practice. Finally, the third common theme foregrounded *many scientists' need and desire to have insight into, and control of, the operation of the software that they use*. Respondents noted a need to be able to control how data is processed by software and, as a result, the requirement that they be able to interrogate its operation. All three of these commonly expressed themes highlight a need and desire to understand the operation of software being used in research practice.

Beyond the above three common themes, additional reasons offered for a group's development of software were varied. They included:

- the cost of existing software being prohibitive,
- the desire to teach students to process data via having them develop software,
- being able to scale to meet the size of the datasets in use,
- increasing the speed of computation to be adequate,
- that their work "broke" the assumptions built into existing software.

Each of these responses illustrates a variety of reasons motivating scientific research groups' development of software.

Finally, we inquired as to whether researchers were aware of any comparable software to that which their group is developing; see Table 7. If they answered yes, we did not however directly ask why they were developing comparable software. Of the 39 potential respondents, 38 answered the question; one person abstained because they felt the question was too general, having many different responses for each piece of software. Looking at the 38 responses 29 researchers said they were not aware of comparable software, 8 indicated that they were, and one researcher was unsure. Examining the Yes responses, we find that 4 came from researchers in the Biology focal field, 3 in Physics, and 1 in Oceanography. In future work, it would be of interest to inquire further as to different software that these researchers are aware of that could be comparable to what they develop.

Table 7. Researchers awareness of comparable software being available given what they and their group are developing (Q29).

	Oceanography	Biology	Physics
Yes	1	4	3
No	8	7	14
Unsure	1	0	0

4.4 Software Sharing

In addition to the goal of studying researchers whose groups develop software, we also primarily aim to examine when such software is and is not shared outside of local scientific groups. Each respondent who indicated that their group develops software was, in turn, asked whether or not their group shares the software that they develop. Thirty-nine respondents were asked this question with 32 answering Yes their group shares the software it develops; see Table 8.

We asked follow-up questions for the respondents who answered Yes to ascertain who they share their software with, what motivates them to share it, whether their group has had to share software to publish a paper, and whether a funding agency has required them to share the software that they develop. For the respondents who answered No, we asked why their group hasn't shared the software that they develop and what would need to happen for them to do so.

Table 8. Whether or not a researcher and their group shares the software that they develop (Q30).

	Oceanography	Biology	Physics
Yes	9	10	13
No	2	1	4

4.4.1 Inquiring About Researcher's Sharing Activities

We asked the 32 respondents who answered that they do share their group's software which of four types of entities they share their software products with; see Table 9. The four entities were the same as our earlier question regarding where they obtain software from. Once again the entities are: 1) local research group, 2) their project collaborators, 3) their community and/or others in their discipline, and 4) the public in general.

As was expected, all 32 answered Yes to sharing with their local research group and 31 indicated that they share with their project collaborators. The one respondent who indicated that they do not share with their project collaborators is an interesting outlier, especially since they answered yes for the three other entities. It may be that they mistakenly selected this response, however without further investigation we cannot be sure.

For the two entities that encompass much larger groups of stakeholders, we begin to see that sharing drops off. Five respondents indicated that they do not share with their community and/or discipline, with four of these five responses coming from the Oceanography focal field and one from Physics. When sharing with the Public at large, only half of the respondents (16) answered yes. Five respondents were unsure whether or not their group shares with the public and 5 indicated it was not applicable. These responses support a belief found across the answers to multiple questions in our survey that the software being developed in the course of research is too focused for widespread use.

Table 9. The entities our respondents share the software that their group develops with (Q31).

	Local Research Group	Project Collaborators	Community and/or Discipline	Public
Yes	32	31	27	16
No	0	1	5	7
Unsure	0	0	0	5
Not Applicable	0	0	0	4

We next asked our respondents how they share the software that their group develops. The most common responses include their public websites, through a general software repository such as GitHub or Google Code, version control systems, through publications, and through email. A few less common but very specific responses include *Bioconductor.org* (an effort of the Fred Hutchinson Cancer Research Center in Seattle, WA), distribution in a programming language library, and through deployment on one of their collaboration's computing systems.

The prevalence of general websites and publications as mechanisms for sharing software are not surprising in the realm of scientific research since these are well established avenues for such activities [8]. The use of software version control and repository sites points to scientists adopting wider collaborative software engineering tools in to their research practice. Furthermore, the emergence and mention of domain-specific sharing mechanisms, i.e. *Bioconductor.org*, points to a potential model for sharing software within domain-specific realms.

In addition, we asked three questions to ascertain some of the motives researchers have for sharing their software. Examining the responses researchers offered for their motivations to share, we find four themes that are prevalent across our respondents:

- Membership as part of a community
- Supporting the replication of research
- Supporting a specific collaboration they are members of
- A desire to bolster their research program

First, many respondents stated that sharing the software that their group develops is part of being a good member of a given community.(i.e. supporting other researchers). For example, one common response was that their group shares what it develops so that other researchers do not have to replicate this work and to obtain potential improvements to their own code. Second, respondents noted that sharing software supports the replication of research since other scholars are using it and pointing out errors while applying the software to new or larger datasets. Third, developing and sharing software was often noted as necessary when participating in a given collaboration. Respondents noting this responded that at times it was simply service work as a part of membership. However, for others sharing their software with a collaboration is a way to share a novel processing or analysis technique that can advance the larger group's goals. Finally, multiple respondents noted sharing their software to advance their own research program. This was most frequently tied to obtaining citations to a publication associated with the piece of software. A few respondents also noted a desire to provide

computational methods for other researchers in their field. Less commonly, one additional motivation for sharing includes a few mentions of sharing due to funding requirements.

Finally, we inquired about researchers sharing their group's software as a result of publication venues and funding agencies asking them to; see Table 10 and Table 11. In both cases the majority of responses indicated No, they were not asked to share the software that they develop by a publication venue or funding agency. Some respondents, however, did find such entities asking them to share their software.

Table 10. Whether a research group is being asked to share their software as a requirement of publishing work (Q34).

	Oceanography	Biology	Physics
Yes	1	5	2
No	8	5	11
Unsure	0	0	0

Table 11. Whether a research group is being asked to share their software as a requirement of receiving funding (Q35).

	Oceanography	Biology	Physics
Yes	2	6	2
No	6	3	9
Unsure	1	1	2

4.4.2 Briefly Examining Why Researchers Don't Share Software

If a respondent indicated that their research group does not share the software that it develops, we asked two brief follow-up questions. The first asks why their group hasn't shared the software that it develops and the second asks what would need to happen for their group to be willing to do so.

Responses provided for why a researcher's group has not shared the software it develops expressed a couple of sentiments. The first was a belief that since they are researchers and not software developers, their software products are not good enough to shared. The second common response was that the software developed is far too specific to be applicable to other researcher's work. In addition, one researcher indicated they were not sure of a venue where they would make their software available; another did not want their group to be responsible for future maintenance of anything that they do share.

When asked what needs to happen for their group to share the software that it develops, three respondents indicated that they would be willing to share if asked directly by someone. One of these three did state that they would require someone who was very excited about the particular piece of software and potentially be willing to take over its maintenance. The couple of remaining responses indicated that they just would not be willing to - or saw no reason to - due to the specificity of the software.

4.5 Reflecting on Software in Scientific Research

At the end of our survey, we asked three wrap-up questions which stepped back to look at the larger picture of software's role in scientific research work. These three questions inquired as to the largest challenge a researcher's group faces with using software in its work, whether their group discusses the impact of software on their datasets, and what concerns, if any, the researcher might have about software affecting the replicability of studies in their field.

The answers regarding researchers' biggest challenges faced with using software in their work cover a wide spectrum of concerns. Prevalent themes include:

- Ensuring that they and their group have sufficient understanding of the software that is being used in their work
- The software operating "correctly" for the task at hand
- Working with large datasets
- A lack of resources (time, money, skilled people) to develop software
- Poor usability of software

For example, researchers professed to not understanding how some of the software they or their group members use impacts the data they have. This concern - and the challenge of ensuring software being used operate "correctly" - highlight more and more readily recognized issues with modern scientific research practice [cf. 11, 12]. Respondents also expressed that it is often challenging for them to ensure that students understand the importance of software as a tool in research practice today.

We also inquired as to whether researchers are discussing the impact of software on their datasets within their groups, of which 51 respondents answered; see Table 12. Of note is that 37 of the 51 responses (73%) indicated that they and their group have talked about this impact a little or a lot. We find it interesting as well that six of the Oceanography respondents indicated their group has not talked about this issue. This may point to a disciplinary difference that could be examined in further research.

Table 12. Whether respondents and their groups discuss the impact of software on their datasets (Q41).

	Oceanography	Biology	Physics
Yes, we have talked about this a lot	4	6	7
Yes, we have talked about this a little	3	7	10
No, we have not talked about this	6	2	2
I am not sure if we have talked about this	1	2	1

Finally we inquired as to each researcher's concerns about software affecting the replicability of studies in their field. A total of 46 respondents provided an answer. Of immediate note is that 14 researchers indicated little to no concern about software impacting the replicability of studies in their field. These researchers were primarily from the Oceanography and Physics focal areas, with only one researcher in the Biology area indicating such a belief. While a lack of concern on

the part of some respondents was expected, we were surprised that almost a third of the responses to this question expressed such a belief.

Beyond the responses that expressed no concern, common concerns were of two themes. The first is that of access to - and transparency of - the software at use in a research project. The second is related to the how the software was produced, pointing out a lack of testing and documentation. Both of these general themes express the belief that access to - and understanding of - the software underlying a research project is necessary to support replication of the methods used to produce the findings. Multiple respondents directly noted in their answers that much software being used is based on potentially incorrect assumptions or full of bugs that may impact the results produced.

Multiple respondents expressed concern that they perceive a lack of testing being applied to software produced as part of the typical research process. Many pieces of software that are developed are noted as being “one-off tools” that are developed rapidly and only to the point that they function as the researcher expects them to. This leaves many potential bugs in place that might be impacting the data.

The sentiments expressed by many respondents regarding testing and the software impacting replicability of studies echo larger conversations being brought up by computer scientists, software engineers, domain scientists, and policy makers regarding the impact of software on the scientific research process [cf. 6,7, 8, 11, 12]. Our long-term study aims to develop deeper characterizations of the processes of developing, using, and sharing such software.

5 Summary

Software is a necessary component of the infrastructure that supports scientific advancement, as it is utilized for data collection, production, analysis, and dissemination. Examining the research of the areas of Oceanography, Biology, and Physics at the University of Washington, we see many interesting facets to the adoption, use, development and sharing of software by scientists.

Many of our survey respondents indicate that they develop their own software because it is necessary for research. If they did not engage in such development activities as part of their research practice they would not be able to seek answers to their questions. While many researchers express the need to develop software, they do also call out the limitations of their experience with this activity and the need for researchers in general to think about the operation of the software in the course of their research work. We note that most participants at the graduate-level and above are seen as needing computational skills in the eyes of Principal Investigators. However, it was interesting that this need was less expressed for undergraduate students participating in research. Continuing to examine the roles undergraduates are taking on as participants in research will help scientific communities better understand how undergraduate education and training is - or is not - sufficient as needed skills evolve.

The replication of research findings is a central tenet of the scientific method. Multiple scholars in the research community at large note the importance of sharing software [11, 12] developed and used in the course of producing research findings. Multiple respondents to our survey echoed such sentiments, with some emphatically commenting that if the software used in the research process is not shared and openly available then the work is not replicable. The handful of respondents who indicate that they do not share with their community or discipline to maintain a competitive advantage offers an interesting counterpoint to an otherwise general expression

of a willingness to openly share. The low number of respondents indicating that publication venues or funding agencies are requiring them to share the software that they produce in the course of their research work offers an area for further inquiry. Understanding whether such entities are mandating software products be shared and are simply not *enforcing* such requirements- or whether requirements of this type do not exist at all - is of relevance when trying to understand the policy climate scientists are working within. This may be an avenue, along with changing educational practices, where the importance of software to research may be further emphasized.

6 Acknowledgments

The authors would like to thank all of the researchers who took our survey.

This material is based upon work supported by the National Science Foundation under Grant Number ACI-1302272. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

7 References

[1] University of Washington. (2012). Academics and Research. http://www.washington.edu/discover/academics. Last accessed Feb. 26, 2014.

[2] Galison, P., & Hevly, B. W. (1992). Big science: The growth of large-scale research. Stanford University Press.

[3] The Royal Society Science Policy Centre. (2012). Science as an Open Enterprise (DES24782): UK Royal Society.

[4] Atkins, D. E., Droegemeier, K. K., Feldman, S. I., Garcia-Molina, H., Klein, M. L., et al. (2003). Revolutionizing Science and Engineering Through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure. Washington D.C.: National Science Foundation.

[5] Edwards, P. N., Jackson, S. J., Bowker, G. C., & Knobel, C. P. (2007). Understanding Infrastructure: Dynamics, Tensions, and Design (Workshop Report).

[6] Hannay, J. E., Langtangen, H. P., MacLeod, C., Pfahl, D., Singer, J., et al. (2009). How do scientists develop and use scientific software? In Proc. Software Engineering for Computational Science and Engineering, SECSE '09. 1-8.

[7] Segal, J. (2009). Software Development Cultures and Cooperation Problems: A Field Study of the Early Stages of Development of Software for a Scientific Community. Computer Supported Cooperative Work (CSCW), 18, 5, 581-606. DOI= http://dx.doi.org/10.1007/s10606-009-9096-9.

[8] Howison, J., & Herbsleb, J. D. (2011) Scientific Software Production: Incentives and Collaboration. In *Proceedings of the ACM Computer Supported Cooperative Work (CSCW) Conference.* ACM. p. 513-522.

[9] Miles, M. B., & Huberman, A. M. (1994) *Qualitative data analysis: An expanded sourcebook*. Sage Publications, Incorporated.

[10] Paine, D., & Lee, C.P. (in process). How Did I Get This Data? Examining Sources of Data Across Scientific Disciplines.

[11] Ince, D. C., Hatton, L., & Graham-Cumming, J. (2012). The case for open computer programs. *Nature, 482*, 7386, 485-488.

[12] Stodden, V. (2012). Reproducible Research: Tools and Strategies for Scientific Computing. *Computing in Science & Engineering*, 11-12.

8 Appendix: Survey Questions

The questions from the survey are presented below along with the question type and any constrained answer lists. The pages of the survey were divided into sections. The headings are presented in bold between questions. Some section names are repeated as there were logical jumps, depending on the answer to a question.

Note: If not otherwise noted, the question was a free response text entry.

Q1.IRB Acknowledgment

Background Information

Q2. What is your preferred name should we contact you again?
Q3.What is your current title?
Q4.Which department are you in?
Q5.What other departmental or institutional affiliations do you have, if any?

Research Group Information

Q6.What is the name of your research group?
Q7.Please describe in 3-5 sentences the research that your group undertakes.
Q8.What kinds of data are used in your group's research?
Q9.Where does your research group obtain data from?

- This was asked with a multiple-choice question.
- Choices:
  - *Experiments*
  - *Fieldwork*
  - *Sensor Systems (i.e. Satellites, telescopes)*
  - *Community Databanks (i.e. GenBank, NASA or NOAA datacenters, etc.)*
  - *Simulations (i.e. Climate Models, MATLAB models, etc.)*
  - *Publications*
  - *Other (free response entry)*

Q10. How many active research projects does your group currently have?
Q11. How many post-doctoral researchers are members of your research group?
Q12. How many research scientists are members of your research group?
Q13. How many doctoral students are members of your research group?
Q14. How many masters students are members of your research group?
Q15. How many undergraduate students are members of your research group?
Q16. For each of the following categories of researcher in your group: What is the likelihood that a researcher will ***already have*** formal computational training before joining in your group?

- This was asked with a multiple-response matrix question.
- Categories were:
  - *Post-Doctoral Researchers*
  - *Research Scientists*
  - *Doctoral Students*
  - *Masters Students*
  - *Undergraduate Students*
- Likert-scale answer choices were:
  - *Very Likely*
  - *Somewhat Likely*
  - *Somewhat Unlikely*
  - *Very Unlikely*

Q17. For each of the following categories of researcher in your group: What is the likelihood that a researcher will ***obtain*** formal computational training while working in your group? (assuming they do not already have such training)

- This was asked with a multiple-response matrix question.
- Categories were:
  - *Post-Doctoral Researchers*
  - *Research Scientists*
  - *Doctoral Students*
  - *Masters Students*
  - *Undergraduate Students*
- Likert-scale answer choices were:
  - *Very Likely*
  - *Somewhat Likely*
  - *Somewhat Unlikely*
  - *Very Unlikely*

Q18. What resources are used by individuals in your group to obtain formal computational training?

- *Examples of formal computational training might include:*
  - *Computer science course(s)*
  - *Programming workshop(s) Online learning experience(s)*

Software Use

The remainder of the survey consists of questions about the software that is a part of your research work.

Examples of software include:

- *Scripts or macros for working with data developed in MATLAB, Perl, Python, Excel, etc.*
- *Commercial office applications*
- *Collaboration tools such as Wikis, e-Mail, Skype, etc.*
- *Database systems*
- *Software to use specific hardware, i.e. sequencing machine or telescope control software*
- *Command line tools*

Q19. What software does your group use to create and/or collect data?

- *(If a piece of software is also used with another stage of the research process we are asking about please feel free to copy/paste between answers)*
- *Creating or collecting data may include activities such as*:*

- *designing the research*
- *conducting experiments or observations*
- *simulating a phenomena*
- *accessing community databanks or systems*
- *accessing archived data from within your group*
- **Some examples from http://data-archive.ac.uk/create-manage/life-cycle*

Q20. What particular criteria do your group use when determining which software to use for data creation and/or collection?

- *(If the criteria for selecting software for another stage of the research process we are asking about are identical, please feel free to copy/paste between answers)*

Q21. What software does your group use to process data?

- (If a piece of software is also used with another stage of the research process we are asking about please feel free to copy/paste between answers)
- Processing data may include activities such as*:
  - digitizing data from lab experiments
  - validating or cleaning datasets
  - anonymizing data when necessary
  - describing a dataset, i.e. producing metadata
  - organizing multiple datasets into one dataset for analysis
  - *Some examples from http://data-archive.ac.uk/create-manage/life-cycle

Q22. What particular criteria do your group use when determining which software to use for data processing?

- (If the criteria for selecting software for another stage of the research process we are asking about are identical, please feel free to copy/paste between answers)

Q23. What software does your group use to analyze data?

- (If a piece of software is also used with another stage of the research process we are asking about please feel free to copy/paste between answers)
- Analysis of data may include activities such as*:
  - interpreting data by running statistical tests or through visual analysis
  - producing research outputs such as charts or tables
  - producing publications
  - *Some examples from http://data-archive.ac.uk/create-manage/life-cycle

Q24. What particular criteria do your group use when determining which software to use for data analysis?

- (If the criteria for selecting software for another stage of the research process we are asking about are identical, please feel free to copy/paste between answers)

Q25. Does your research group use software developed by any of the following groups?

- This was asked with a multiple choice matrix question.
- For example:
  - Data cleaning scripts developed by a project collaborator at another institution
  - Open source libraries such a Biopython, PyML, PyVO, etc. developed by a community of practice
  - Libraries or applications such as PyMath, gnuplot, etc. developed by a member of the public at large
- Category choices were:
  - *Local Research Group*
  - *Project Collaborators*
  - *Community and/or Discipline*
  - *Public*

- Response choices were:
  - *Yes*
  - *No*
  - *Unsure*
  - *Not Applicable*

Software Development

Q26. Does your research group develop software?

- This was asked with a Yes/No radio button question.
- **This question results in a logic jump in the survey.**

Software Development

Questions for those who answered Yes to developing software.

Q27. What software does your research group develop?

- Please list the names (if applicable).

Q28. Why did your research group find it necessary to develop this software?

Q29. Are you aware of any comparable software that your group could use in place of what you are developing?

- This was asked with a radio button question.
- Choices:
  - *Yes*
  - *No*
  - *Unsure*

Software Sharing

This section is only present for those who answered Yes to developing software.

Q30. Does your research group share the software that it develops with anyone outside of your local research group?

- This was asked with a Yes/No radio button question.
- **This question results in a logic jump in the survey.**

Software Sharing

*This section is for those who answered Yes to sharing **and** developing software.*

Q31. Who does your research group share the software that it develops with?

- This was asked with a multiple-choice matrix question.
- Category choices were:
  - *Local Research Group*
  - *Project Collaborators*
  - *Community and/or Discipline*
  - *Public*
- Response choices were:
  - *Yes*
  - *No*
  - *Unsure*
  - *Not Applicable*

Q32. How does your research group share the software that it develops?

- I.e. public website, publications, GitHub, etc.

Q33. What motivates your research group to share the software that it develops?

Q34. Has your research group been asked to share the software that it develops as a requirement for publication?

- This was asked with a radio-button question.
- Choices were:
  - *Yes*

- *No*
- *Unsure*

Q35. Has your research group been required by a funding agency to share the software that it develops?

- This was asked with a radio-button question.
- Choices were:
  - *Yes*
  - *No*
  - *Unsure*

Software Sharing

This section is for those who answered No to sharing software and Yes to developing software.

Q36. Why hasn't your research group shared the software that it develops?

Q37. What needs to happen for your group to be willing to share the software that it develops?

Software Development

This section is for those who answered No to developing software. They were not asking questions regarding software sharing.

Q38. Is your research group customizing any of the software that you previously mentioned?

- This was asked with a Yes/No radio-button question.

Q39. If yes, what previously mentioned software is your group customizing?

Wrap-Up

Q40. What is the largest challenge that your group faces with software in its work?

Q41. Does your research group talk about the impact of software on its datasets?

- This was asked with a Likert-scale question.
- Choices were:
  - Yes, we have talked about this a lot
  - Yes, we have talked about this a little
  - No, we have not talked about this
  - I am not sure if we have talked about this

Q42. What concerns do you have about software affecting the replicability of studies in your field, if any?

Summary

Q43. To receive your $10 coffee gift card please provide your University of Washington box number. If you do not provide a box number we will assume you do not wish to receive a gift card.