Catching Up to Corporate:
A Shift towards Academic Data Governance

Jennifer L. Clark, M.S.
University of Illinois, Urbana-Champaign

Lori A. Hurley, M.S.

Andrea L. Ogier, M.S.
Data Science and Informatics Librarian
Center for Digital Research and Scholarship
University Libraries
Virginia Tech

Introduction

The recent rise in concern over the implications of big data analysis extends to issues endemic to the research process. Academic researchers, whether using large or small datasets, have always faced challenges in data processing, but the rapidly increasing volume of data has forced the issue of the quality of data management into the limelight. Inconsistent and unreproducible workflows in data transformations and data processing can affect analysis and hinder scientific progress on the big and small scale. Big data has brought these issues to the forefront, but the implications on the credibility of knowledge-based science have a more widespread reach.

As scientific research output becomes more complex, the need for data governance frameworks in addition to the growing number of data management policies and standards becomes more apparent. Corporate institutions have long benefited from models for process improvement and quality assurance, such as Six Sigma. Recently, many corporations have begun to adopt data governance frameworks, such as COBIT and ISO/IEC 38500, suggesting that the creation and implementation of data management policies are processes that should be regulated, measurable, repeatable, and transparent -- at least from within an enterprise. No data governance framework currently exists in academia for research stakeholders (including researchers, research institutions, journals, and funding agencies), meaning data management activities are not yet as standardized.

The increasing application of these frameworks in the corporate sphere highlights a dedication to cost savings through better data quality, which in turn, will allow for more robust and broader-range use of this data through analytics. Unfortunately, the current incentive structure in academia does not reward researchers for transparent workflows and high data quality. Without a framework for consistent processes and quality assurance in place, barriers to shareable, high quality data will continue to exist. In this paper we will outline challenges scientists and researchers face in the pursuit of open, reproducible science; discuss current movements or tools responding to these challenges; and suggest a way forward based on a survey of academic researchers.

Data Governance Issues In Academia

Before scientific stakeholders can move towards a more standardized, regulated environment of data management, certain systemic problems need to be addressed. Though the problems that follow do not represent an exhaustive list, they provide an introduction to the hurdles researchers face when trying to practice open science supported by high quality data management practices.

NOT ALL SCIENCE IS BIG SCIENCE

In order to develop policies for documenting data lineage and data management process improvement, data management professionals must first be able to recognize, and categorize, scientific disciplines by how a discipline's data is conceived. Data from Big Science often receives the most attention and planning, receiving more funding and better technical infrastructure. This data is typically "highly organized on the front end — researchers define it before it even starts rolling off the machines — to make it easier to handle, to understand, and to archive" (Carlson 2006). On the other hand, data from small science is "horribly heterogeneous," requiring more processing time and attention with less funding and staff (Carlson 2006). What was once thought of as big data has quickly been redefined, leaving scientists handling these smaller yet still computationally needy datasets with less support for their highly customized needs.

THE CULTURE OF CLOSED SCIENCE

Though Big Science differs from small science in data creation, science as a whole suffers from a secretive, closed-off culture which has led to an epidemic of unreproducible results. Quality assurance through checks and balances of the research workflow are supposed to be built into proper research design, the peer-review process, and trusted discoveries, but scientific reproducibility is at an all-time low. Stodden (2012) notes that "researchers today

aren't sufficiently prepared to ensure reproducibility, and after-the-fact efforts -- even heroic -- are unlikely to provide a long-term solution" (11-12). Unreproducible "landmark" studies have recently received attention in a variety of top-tier journals. Begley and Ellis (2012) presented the recreation of 53 studies by Amgen, a biotechnology firm, in which only 6 (11%) were confirmed (532). Baggerly and Combes (2009) point out that data processing and poor documentation often lead to "forensic bioinformatics" and present five case studies with errors (1309). They also warn that "the most common errors are simple (e.g., row or column offsets); conversely, it is our experience that the most simple errors are common" (1309).

These problems are widely recognized, but there has been slow progress towards the sharing and publication of important research documents and supplemental files, including custom code and lab notebooks, that could pave the path for quality assurance and reproducibility. The Yale Law School Roundtable on Data and Code Sharing (2010) notes that as central as data and supplemental code are to the growth and future of science, standards governing these measures are not yet concerned with scientific validation and reproducibility (8-12). *Nature* (2008) agrees, stating in an editorial from an issue devoted entirely to Big Data that "researchers need to be obliged to document and manage their data with as much professionalism as they devote to their experiments...Universities and funding agencies need to provide and support curation facilities, tools and training" (1).

However, the idea of increased transparency in data management and manipulation practices is highly controversial. The fight for funding and the race for new patents leave little time for either the development and upgrading of computational skills or for collecting and providing the documentation needed to disseminate code and workflows. Additionally, the promotion and tenure structure currently found in academia rewards quantity of publications and grant dollars, though there is some acknowledgement of quality (mostly via peer review for the article). With the current lack of quality controls, there is little risk attached to publishing inaccurate data, and therefore there is little incentive to sacrifice quantity in order to assure quality.

Shifting the culture of academia from quantity to quality means tying the funding dollars to reproducibility, and a correspondingly higher quality of data management processes. Though data management plans are considered an essential component of many funding agencies' grant applications, few require or even recommend the inclusion of verification standards or quality assurance for custom-created software or code. The National Science Foundation's (NSF) policy (2011) only stipulates that NSF will provide to researchers "appropriate support and incentives for data cleanup, documentation, dissemination, storage and the like" (Section D, Part 4). The National Institutes of Health's (NIH) policy (2003) has simply promised to connect investigators so that they may share expertise in "cleaning and formatting data, writing documentation, redacting data to protect subjects' identities and proprietary information, and estimating costs to prepare documentation and data for sharing."

Finally, in some areas of research, intellectual property rights are of concern, as modeling or data analysis software may be the result of long-term work that then provides an edge for future research (*The Economist* 2013). These kinds of considerations keep the movement towards openness from being more widespread.

Small science faces even greater challenges in this environment, as it is seen as "improbable and risky" and "less likely to attract large grants if they can get any grant at all" (Heidorn 2008, 282). Because open-source data transformation code/software are unavailable and pre-existing datasets do not exist or have not been shared, these scientists are left with little to no resources for processing, storage, and documentation, the lack of which may hinder the progress of their labs. In addition, smaller projects may be unable to afford to outsource their data processing needs or any verification procedures, thereby placing them at still a greater disadvantage.

The paucity of grant funding is also an issue, as universities and other institutions place a higher weight on researchers who attract grant money. Major funding organizations, such as the NIH and NSF, announced in 2013 that due to the recent sequester both organizations would be operated at a 5% decrease in funding in 2013 as compared to 2012 (Suresh 2013; NIH 2013a). These cuts, on top of higher education cuts nationally, place extra pressure on researchers to compete for fewer research positions as well as smaller pools of funding, despite rising research costs. This competitive environment is increasingly a disincentive for complying with demands for transparency, which has a high cost in time and money.

RESEARCHERS HAVE CHANGING NEEDS

In addition to the outdated publish-or-perish models, the rapid disappearance of funding, and the competitive, closed environment these problems create, the interdisciplinary nature of modern research has created an additional burden for researchers needing an ever widening range of skills to interact with datasets. Many scientific disciplines do not include training in computer programming and statistics, forcing scientists to look outside of their disciplines to learn these necessary skills. While this necessity is not always a drawback (as it may lead to more interdisciplinary collaborations), researchers may not have the time to fully explore an unrelated discipline or simply may be unaware of gaps in their skillsets. Even those who possess the knowledge and experience to programmatically script calculations or transformations of their data may not have time to keep current on theories of efficient data storage, management, and mining technologies. Researchers who have received this type of training, outside of their departments or on their own, are often more competitive and productive (Venkatraman 2013).

Reliance on pre-packaged software, such as R, SPSS, Nvivo, and even Microsoft Excel, for handling research data is not a guaranteed solution to gaps in skillsets. While availability of

such tools creates some standardization and increases accessibility to data processing for scientists lacking advanced programming skills, there are few checks in place to assure that derived data and analysis are valid. Assuming the software itself is of high quality, scientists untrained in its underlying algorithms may not supply the appropriate inputs or misconstrue the outputs, leading to incorrect conclusions or missing important findings. Joppa et al (2013) found that many scientists select software without a full understanding of its proper use or any means to validate results (814-15). A recent example is found in the case of a simple error in the use of Microsoft Excel software which radically impacted the conclusions of a famous and often-cited economics paper on austerity measures (Roose 2013).

PRESERVING DATA IS NOT SUFFICIENT

Finally, the standards for data management that do exist often treat data sets as stand-alone objects to be preserved after the project has been completed. Best practices like the Digital Curation Centre's (DCC) Curation Lifecycle Model tend to be geared towards preservationists and librarians who receive processed data outside of its original context. As more science comes online through re-usable datasets, the need to capture and preserve raw and processed data, as well as supplemental code, software, and lab documentation as representation information for designated communities, becomes tantamount. Stodden (2009) agrees, stating that "releasing of data is important... but is typically not useful without a clear understanding of what methodologies were employed in the construction of the dataset" (5).

Without these supplemental research materials, the data that will be preserved will remain out of context and the research behind them will remain unreproducible. Providing access to these objects will empower both academic researchers and citizen scientists with the necessary tools to assess the reliability and credibility of the data available.

Towards Academic Data Governance

As the academic community begins to better understand research that involves big data, the current practices and policies in place will have to shift along with the new knowledge. A culture change towards open science and high standards of data governance will occur when academic incentives are re-evaluated to reward quality over quantity. Though there is still a long road ahead, journal review boards, funding agencies, and educational institutions are beginning to value researchers who practice robust standards of data management and support researchers in this pursuit.

POLICIES FOR REPRODUCIBILITY

There is a growing consensus that researchers need to both develop workflows and provide documentation of code and testing practices for their research (Yale 2010, 8-12; Goble and De Roure 2009, 88-95). Typically, journal submissions have no provisions for a peer review process to verify the quality of data transformations and analysis; however, some journals are beginning to place greater emphasis on standards of transparency. The *Journal of Open Research Software* and *IPOL: Image Processing Online* policies are examples of the increasing awareness that reproducibility must extend to the scientific software utilized (Joppa et al. 2013, 815). *Science* announced in January 2014 that in addition to adhering to US National Institute of Neurological Disorders and Stroke (NINDS) recommendations for reporting practices (Landis et. al 2012, 187-191), a statistician would be added to their Board of Reviewing Editors in order to ensure future publications' data analysis were properly reviewed (McNutt 2014, 229). In May of 2013, *Nature* and its sister publications introduced an 18-point checklist for submissions, intended to collect crucial technical and statistical information so as to insure reproducibility (*The Economist* 2013).

Proprietary platforms are also emerging in response to the problem of reproducibility. Vendors, like The Science Exchange Network, offer fee-based services to replicate experimental results. Their validation service provides a means to replicate results, allowing for the "identification of high quality reproducible research and reagents" (2014b). Once the experiment has been replicated, they advertise that they "provide all protocols, results, raw and processed data for review" (Science Exchange 2014b). The Science Exchange is also collaborating with the Center for Open Science, to replicate 50 recent cancer biology studies. The aim of the project is to "identify best practices… that maximize reproducibility and facilitate an accurate accumulation of knowledge" (Science Exchange 2014a).

Many funding agencies now require grant recipients to share data and supplemental material as a stipulation of funding. The National Science Foundation's (NSF) data sharing policy (2013) states that recipients are expected to share "the primary data, samples, physical collections and other supporting materials created or gathered in the course of work," as well as "software and inventions created under the grant" (Section D, Part 4). The National Institutes of Health (NIH) (2003) requires final research data sharing, which is defined as "recorded factual material commonly accepted in the scientific community as necessary to document and support research findings," of recipients receiving $500,000 or more in direct costs in a year. However, NIH (2003) does not consider "laboratory notebooks, partial datasets, [and] preliminary analyses" to be part of final research data. A separate policy (2013b) is outlined to handle these types of objects, called research resources or tools. Though they encourage the sharing of these resources and recognize that restriction of access to these materials "can impede the advancement of further research," the sharing policy for these materials is more guarded, with the intention of protecting intellectual property rights.

MEETING RESEARCHERS' NEEDS

There is now a push for undergraduate and graduate programs in science to routinely include computer courses in order to encourage more reliable workflows and outcomes. The emergence of data science as a discipline, evidenced by a popular Coursera course taught by Bill Howe, Director of Research at the University of Washington, and online courses from both Harvard Extension and the University of Syracuse School of Information Studies, aim to increase training in this area. Targeted traveling programs, such as Software Carpentry, allow scientists to engage in brief but intensive workshops to gain computing skills directly relevant to their current work (Venkatraman 2013). In addition to physically connecting computing experts and researchers, Software Carpentry continues to support their "graduates" with online office hours to help with their everyday computing tasks. Free online, interactive computer programming courses, through sites like Code Academy, are also beginning to surface.

Movements aimed at developing sustained collaborations between researchers with a background in a particular subject and those with training in a methodology, such as computer science or statistics, have also been growing in universities. The recent announcement of a five-year project, funded by the Moore and Sloan Foundations in partnership with three major universities to establish a data science discipline focused on shared practices rather than the current culture of data research in domain and/or departmental isolation, shows the growing recognition of the need for established common ground (Lohr 2013). In 2009 and 2011, the "Digging Into Data" grants awarded by the National Endowment for the Humanities encouraged multi-disciplinary teams to promote interdisciplinary investigation of computational techniques that may have the capacity to change the future of humanities and social science research.

In addition to these grant-funded research opportunities, graduate level information science programs are changing to incorporate specializations in data/digital curation and data analytics. Such programs aim to produce information specialists who can team with scientists to fill gaps in data management and data analysis.

Survey Questionnaire

While the barriers to open science and academic data governance extend far beyond the boundaries of scientific labs, information managers can begin to better support researchers through a more comprehensive understanding of their workflows. In an effort to begin to document the workflows of researchers working with large datasets, we designed a study to understand the current practices in data transformations and processing as well as quality assurance. Our methods and results are presented in full.

METHODS

In the spring of 2014, a voluntary survey was sent via email to 282 graduate and faculty researchers at the University of Illinois, Urbana-Champaign. Open for three weeks, with one reminder email a week before the close date, the survey garnered 17 responses. Survey respondents were allowed to move forward in the survey without answering each question, and the data was de-identified before analysis.

The questionnaire featured four sections. The first section asked demographic questions relating to researcher roles/titles and years of experience. The second section focused on data transformations (i.e. change format, merge files, make calculations, add columns based on information from raw data columns, etc.). The third section focused on quality assurance/verification practices (i.e. steps taken to ensure processing did not create errors in the data). Finally, the fourth section focused on documentation of the final data set.

Given the response rate, the survey results are not statistically significant in relation to the sample size. We hypothesize that the low response rate may be due to the controversial nature of our topic. However, we present these results in summary form, as a first approach at a study on this topic.

DEMOGRAPHICS

Of the 17 respondents, just over half reported their role as being a faculty researcher, while one third were PhD-level graduate students. The remainder were postdoctoral researchers and masters-level graduate students. Six respondents reported 15 or more years of research experience, five reported between 9 and 15 years, four reported between 4 and 8 years, and two reported 3 or fewer years of research experience. Ten of the respondents were male and seven were female. By far, the largest demographic of respondent (n = 10) reported datasets of fewer than 50,000 records, while only three reported datasets of more than 1,000,000 records.

POINTS OF FOCUS

While we refrain from making any firm conclusions from the initial results, a few interesting points of focus for future study emerged and may suggest that the responses show the growing gap between Big and small science (Table 1). Of the ten respondents with datasets under 50,000 records (Group 1), seven respondents process their own data sets by changing format, merging files, making calculations, or adding columns based on information from raw data columns. Additionally, all three respondents with datasets of over 1,000,000 records (Group 2) process their own data sets. Seven researchers from Group 1 actively reported no standards for data processing from their industry/discipline while three refrained from answering the question; however, two of the three respondents from Group 2 answered that their industry/discipline has

requirements. Five respondents from Group 1 use proprietary software for data processing, while all three respondents from Group 2 create custom code to do so.

For quality assurance practices, defined as any steps taken to ensure data processing did not create errors in the data, four respondents from Group 1 rely on browsing, two rely on random spot checking, and one relies on histograms. In Group 2, only one of the respondents reported using histograms to assure data quality while two reported using no strategies for assuring data quality. This may suggest that the survey did not address Group 2's methods of large-scale quality assurance, or it may suggest that meeting the industry/discipline requirements for transformations are both part of data processing practices, as well as quality assurance. In contrast, researchers dealing with smaller datasets could rely more on the increased probability of random chance over a smaller dataset as exhibited in spot-checking and browsing through their data.

Finally, only one respondent from Group 1 reported the creation of separate documentation, such as lab notebooks, for data processing workflows. Two respondents from Group 2 reported the creation of separate documentation, while two respondents reported the creation of comments within the custom code created for data processing. Five respondents from Group 1 and two respondents from Group 2 reported the creation of documentation for the finalized dataset, including descriptions of the columns/attributes.

Table 1. Selected Web Survey Responses.

Questions	Researchers working with less than 50,000 records	Researchers working with more than 1,000,000 records
	Group 1 (n=10)	Group 2 (n=3)
Process Own Data Sets	7	3
Industry/Discipline Provides Requirements for Processing of Data Sets	0	2
Software Used for Processing	5 use proprietary software	3 use custom code
Spot Checking Data Set for Quality Assurance	2	~
Browsing Data Set for Quality Assurance	4	~
Histograms of Data Set for Quality Assurance	1	1

No Quality Assurance	~	2
Separate Documentation of Data Processing	1	2
Documentation within Code of Data Processing	~	2
No Documentation of Data Processing	1	~
Documentation/Description of Finalized Data Set	5	2

Conclusions

The current efforts towards reproducible science are a strong start to a better system that will reward high standards of data management and lineage, but these steps are just the beginning. In addition to developing a better understanding of researchers' data processing workflows through additional surveys and interviews, it is crucial that data management professionals extend their reach to move beyond highly-contextualized data management policies to include the creation and application of data governance frameworks in order to encourage traceable, reproducible workflows. Information Science programs should also incorporate these frameworks within data management curricula as a recognition that defining stakeholders and understanding dependencies -- i.e. academia as an enterprise -- is crucial to successful implementation of data management policies.

Professional societies and associations should collaborate with data management professionals to develop better standards and best practices for data transformations. These organizations should require that their standards be met before publication and may even decide to extend their standards to accreditation processes for academic institutions. Also, funding agencies and journal review boards need to not only support and reward researchers who follow agreed-upon standards and frameworks, but also require that supplemental research materials be shared, and provide more technical infrastructure and better tools to assist with difficult workflows. Finally, training and education opportunities need to be extended for both researchers and their collaborators. Research institutions should not only provide their research staff with training on data processing and cleansing, but they should also incorporate these necessary skills into the curriculum for graduate and undergraduate students in scientific fields. The normalization of such efforts to integrate governance into the processing of scientific data, in conjunction with an incentive structure that rewards researchers for exemplary data quality, will propel academia towards high quality, reproducible science.

ACKNOWLEDGEMENTS

The authors would like to express their immense gratitude to Dr. Melissa Cragin and Dr. Kathleen McDowell for their continuous guidance throughout this research process.

References

Baggerly, Keith A. and Kevin R. Combes. 2009. "Deriving chemosensitivity from cell lines: Forensic bioinformatics and reproducible research in high-throughput biology." *Annals of Applied Statistics* 3: 1309-1334. doi: 10.1214/09-AOAS291.

Begley, C. Glenn and Lee M. Ellis. 2012. "Drug development: Raise standards for preclinical cancer research." *Nature* 483: 531-533. doi:10.1038/483531a.

Carlson, Scott. 2006. "Lost in a Sea of Science Data." *The Chronicle of Higher Education*, June 23. https://chronicle.com/article/Lost-in-a-Sea-of-Science-Data/9136.

De Roure, G. & Goble, C. 2009. "Software Design for Empowering Scientists." *Software, IEEE* 26/1: 88-95. doi:10.1109/MS.2009.22.

Economist, The. 2013. "Trouble at the Lab." *The Economist*, October 19. http://www.economist.com/news/briefing/21588057-scientists-think-science-self-correcting-alarming-degree-it-not-trouble.

Heidorn, P. Bryan. 2008. "Shedding light on the dark data in the long tail of science." *Library Trends* 57: 280-299. doi:10.1353/lib.0.0036.

Joppa, Lucas N., Greg McInerny, Richard Harper, Lara Salido, Kenji Takeda, Kenton O'Hara, David Gavaghan, and Stephen Emmott. 2013. "Troubling Trends in Scientific Software Use." *Science* 340: 814-15. doi:10.1126/science.1231535.

Landis, Story C., Susan G. Amara, Khusru Asadullah, Chris P. Austin, Robi Blumenstein, Eileen W. Bradley, Ronald G. Crystal, Robert B. Darnell, Robert J. Ferrante, Howard Fillit, Robert Finkelstein, Marc Fisher, Howard E. Gendelman, Robert M. Golub, John L Goudreau, Robert A. Gross, Amelie K. Gubitz, Sharon E. Hesterlee, David W. Howells, John Huguenard, Katrina Kelner, Walter Koroshetz, Dimitri Krainc, Stanley E. Lazic, Michael S. Levine, Malcom R. Macleod, John M. McCall, Richard T. Moxley III, Kalyani Narasimhan, Linda J. Noble, Steve Perrin, John D. Porter, Oswald Steward, Ellis Unger, Ursula Utz and Shai D. Silberberg. 2012. "A call for transparent reporting to optimize the predictive value of preclinical research." *Nature* 490: 187-191. doi:10.1038/nature11556.

Lohr, Steve. 2013. "Program Seeks to Nurture 'Data Science Culture' at Universities." *The New York Times*, November 12. http://bits.blogs.nytimes.com/2013/11/12/program-seeks-to-nurture-data-science-culture-at-universities/.

McNutt, Marcia. 2014. "Reproducibility." *Science* 343: 229. doi: 10.1126/science.1250475.

National Institutes of Health. 2003. "NIH Data Sharing Policy and Implementation Guidance." https://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm.

—. 2013a. "NIH Fiscal Policy for Grant Awards -- FY2013." http://grants.nih.gov/grants/guide/notice-files/NOT-OD-13-064.html.

—. 2013b. "NIH Grants Policy Statement." http://grants.nih.gov/grants/policy/nihgps_2013/nihgps_ch8.htm#_Toc271264947.

National Science Foundation. 2011. "Chapter VI - Other Post Award Requirements and Considerations." http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/aag_6.jsp#VID4.

—. 2013. "Chapter II - Proposal Preparation Instructions." http://www.nsf.gov/pubs/policydocs/pappguide/nsf13001/gpg_2.jsp#dmp.

Nature. 2008. "Community Cleverness Required." *Nature* 455: 1. doi:10.1038/455001a.

Roose, Kevin. 2013. "Meet the 28-Year-Old Grad Student Who Just Shook the Global Austerity Movement." *New York Magazine*, April 18. http://nymag.com/daily/intelligencer/2013/04/grad-student-who-shook-global-austerity-movement.html.

Science Exchange Network. 2014a. "Reproducibility Project: Cancer Biology." http://validation.scienceexchange.com/#/cancer-biology.

—. 2014b. "Validation by The Science Exchange Network." http://validation.scienceexchange.com/#/home.

Stodden, Victoria. 2009. "Enabling Reproducible Research: Licensing for Scientific Innovation." *International Journal of Communications Law & Policy* 13.

—. 2012. "Reproducible Research: Tools and Strategies for Scientific Computing." Computing in Science & Engineering 14.4:11-12. doi: 10.1109/MCSE.2012.82.

Suresh, Subra. 2013. "Impact of FY 2013 Sequestration Order on NSF Awards." *National Science Foundation*, February 27. http://www.nsf.gov/pubs/2013/in133/in133.pdf.

Thaney, Kaitlin. 2014. "Code as a Research Object: Updates, Prototypes, next Steps." *Mozilla Science Lab*. http://mozillascience.org/code-as-a-research-object-updates-prototypes-next-steps/.

Venkatraman, Vijaysree. 2013. "When All Science Becomes Data Science." *Science Careers*. http://sciencecareers.sciencemag.org/career_magazine/previous_issues/articles/2013_05_13/caredit.a1300099.

Yale Law School Roundtable on Data and Code Sharing. 2010. "Reproducible Research." *Computing in Science & Eng* 12: 8-13. http://www.stanford.edu/~vcs/papers/RoundtableDeclaration2010.pdf.

Appendix

SURVEY QUESTIONS

Background Information

1. What is your role at the University of Illinois at Urbana-Champaign?
a) Graduate student, Master's level
b) Graduate student, PhD level
c) Faculty researcher
d) Research Staff (Research Scientist, Research Assistant, etc.)
e) Post Doc

2. How many years of research experience do you have?
a) 0 - 3
b) 4 - 8
c) 9 - 15
d) 15+

3. What is your gender?
a) male
b) female
c) other
d) prefer not to respond

Research Practices

4. Which of the following best describes the typical size (by number of records/rows) of your datasets?
a) less than 50,000 records
b) between 50,000 and 250,000 records
c) more than 250,000 and 1 million records
d) over 1 million records

5. In your research, do you have to process or transform data/datasets (i.e. change format, merge files, make calculations, add columns based on information from raw data columns, etc.) or do you not do that?
a) yes, I process data/datasets
b) no, I do not process data/datasets

Data Transformations

6. Does your department or industry provide requirements (processes you must follow) or best practices (general high-level guidelines) for data transformations/processing (i.e. changing formats, merging files, making calculations, adding columns based on information from existing columns, etc.) or do they not exist? (Select all that apply.)
a) Yes, my department provides requirements
b) Yes, my industry provides requirements
c) Yes, my department provides best practices
d) Yes, my industry provides best practices
e) No, neither requirements nor best practices exist

7. What method is used for data transformations/processing? (Select all that apply)
a) Existing internal custom code (i.e. code from another project created specifically for your lab/research)
b) Existing external custom code (i.e. code from another project created by another researcher/lab)
c) Pre-packaged spreadsheet software (i.e. Microsoft Excel, Corel Quattro Pro, or Numbers for Mac)
d) Pre-packaged statistical analysis software (i.e. R, SPSS, or Matlab)
e) New custom code (i.e. code newly created specifically for your lab/research)
f) Other (Please Specify)

Custom Code

8. Would you be willing to make custom data transformations/processing code available with the

final dataset (provided that you maintained appropriate control and copyright) or would you not be willing?
a) Yes, I would be willing
b) No, I would not be willing

9. Which of the following benefits do you see to making custom data transformations/processing code available with the final dataset? (Select all that apply.)
a) Reproducibility
b) Knowledge sharing
c) Enhancement of scholarly impact
d) Enhancement of professional reputation
e) Other (Please Specify)
f) I see no benefit

10. Which of the following people write custom code for your lab/research? (Select all that apply.)
a) I write my own code
b) Graduate Assistant (Master's level)
c) Graduate Assistant (PhD level)
d) Post Doc
e) Research Staff (i.e. Research Scientist or Research Assistant, not a student)
f) Outsourced programmer/developer outside of lab but provided by department
g) Outsourced programmer/developer outside of lab and outside of department
f) Other (Please Specify)

11. Which of the following describes documentation created for the custom code? (Select all that apply.)
a) Custom code includes comments within the code
b) Document(s) separate from the code (lab notebooks, shared docs, etc.)
c) Other (Please Specify)
d) None

Quality Assurance

12. After data transformations/processing have been applied, which of the following quality assurance/verification practices (i.e. steps taken to ensure processing did not create errors in the data) are completed? (Select all that apply.)
a) Random spot checking of transformed data
b) Quality Assurance/Data Verification using a pre-packaged proprietary software tool
c) Quality Assurance/Data Verification using a pre-packaged open source software tool
d) Browsing of records for accuracy ("eyeballing" the dataset)

e) Quality Assurance/Data Verification of the code using a sample dataset created for testing purposes
f) Histograms or visualizations to seek anomalies in the data
g) Other (Please Specify)
h) None

(If b or c is selected in Question 12) 13. Please specify which pre-packaged software tool you use for Quality Assurance/Data Verification: _____

Quality Assurance Workflow

14. If your research involves a project plan, are quality assurance/verification practices (i.e. steps taken to ensure processing did not create errors in the data) included in that project plan or are they not included?
a) Yes, quality assurance is always included in the project plan.
b) No, quality assurance is never included in the project plan.
c) Sometimes, quality assurance is occasionally included in the project plan.

15. Who performs the quality assurance/verification practices (i.e. steps taken to ensure processing did not create errors in the data)?
a) Person who performed the data transformations/processing
b) Person who was not responsible for the data transformations/processing
c) Multiple team members

16. What is the role of the person who performs quality assurance/verification practices (i.e. steps taken to ensure processing did not create errors in the data)?
a) I perform my own quality assurance
b) Graduate Assistant (Master's level)
c) Graduate Assistant (PhD level)
d) Research Staff (i.e. Research Scientist or Research Assistant, not a student)
e) Post Doc
f) Other (Please Specify)

Final Documentation

17. Once the final data set(s) is created (i.e. after all data transformations/processing and quality assurance/verification practices are performed), is documentation created to describe the dataset (i.e. a document outlining the definitions of attribute/column in the dataset) or is it not created?
a) Yes, Documentation is created
b) No, Documentation is not created