Examining Data Processing Work as Part of the Scientific Data Lifecycle: Comparing Practices Across Four Scientific Research Groups

Drew Paine, Human Centered Design and Engineering, University of Washington
Erin Sy, Human Centered Design and Engineering, University of Washington
Ron Piell, Human Centered Design and Engineering, University of Washington
Charlotte P. Lee, Human Centered Design and Engineering, University of Washington

Abstract
Data processing is work that scientists must undertake in order to make data useful for analyses, and is a key component of twenty-first century scientific research. The analysis of scientific data is contingent upon the successful collection or production and then processing of data. This qualitative research study, of four data-intensive research groups, investigates scientists engaging in data processing work practices to describe and analyze three distinctive but intertwined practices: cleaning data products, selecting a subset of a data product or assembling a new data product from multiple sources, and transforming data products into a common format. These practices are necessary for researchers to transform an initial data product in to one that is ready for scientific analysis. This research finds that data processing work requires a high level of scientific and technical competence that does not merely set up analyses, but also often shapes and is shaped by iterations of research designs and research questions themselves.

Keywords: data processing, infrastructure, data-intensive science, qualitative methods
Citation: Paine, D., Sy, E., Piell, R., Lee, C.P. (2015). Examining Data Processing Work as Part of the Scientific Data Lifecycle: Comparing Practices Across Four Scientific Research Groups. In *iConference 2015 Proceedings*.

Acknowledgements: The authors thank our participants for their time, our anonymous reviewers for their feedback, and Ying-Yu Chen and Ilana Diamant for help with data collection. This material is based upon work supported by the National Science Foundation grants IIS-0954088 and ACI-1302272.
Contact: pained@uw.edu, esy5@uw.edu, ronpiell@uw.edu, cplee@uw.edu

1 Introduction

Twenty-first century scientific research is a collaborative process engaging many stakeholders, technology systems, and practices in a complex, ever evolving sociotechnical milieu. Many scholars in Computer Supported Cooperative Work (CSCW) and Information Science find that knowing and understanding the context and process of the creation of datasets is necessary and important for researchers to be able to analyze, share, or reuse data in research work (Faniel & Jacobsen, 2010; cf. Jirotka et al., 2005; Rolland & Lee, 2013). Without this knowledge and understanding, it is difficult or not possible for researchers to answer research questions. They may not know how variables were stored (i.e. in which units of measurement) or whether artifacts of an experiment are still present, thus potentially skewing analyses and findings.

Key to understanding the context of a dataset's creation is the work of processing data from its initial collected state to an understood and well-organized state that enables researchers to examine their particular research questions. This work, in between collection or production and analysis of data, is of vital importance to the research process in spite of the laborious nature of the work. This "in between" work is referred to in this paper as processing work. To further our understanding of processing work, we examine the following research question: ***How is data processing work a bridge between the collection and analysis work of scientific research?***

In this paper, we examine the work of four scientific research groups as they engage in the computational processing of their data. We examine data processing work as the extensive process of taking the data that is gathered for a project in its various forms, and preparing it for use in the analysis of the particular questions driving a project. Processing work is necessary across many stages of a research project. In this paper, we specifically focus on the processing work of four scientific research groups after they have collected or produced their "raw" data products.

The scientific research process iterates between multiple stages of work – beginning with research design and data collection leading to repeated processing and analysis work – which produces many different products. Data processing work is one of these stages. Research projects begin by defining the initial research design. This research design then acts as a guide for the collection or production of data, which results in initial *data products* (datasets output by instruments or processing and analysis pipelines). Data processing work takes these initial products and refines them so that they are usable for the ensuing analysis work. Examining data processing highlights the many practices that are involved in the production of multiple research products. These research products are not only the *data products,* but also the *software products* being developed or used and, in many fields, the *model products* (computational simulations that combine theory, data, and algorithms to produce new data) that are used to answer research questions. The data processing work sets the stage for the eventual analysis work.

In the remainder of this paper, we begin by reviewing relevant literature then introduce our research sites and methods. We then describe the processing work of all four of the research groups being studied. Finally, we examine and discuss common concerns for processing work that we see across different groups and research areas.

2 Literature Review

The Computer Supported Cooperative Work (CSCW) and Information Science communities have long studied the collaborative work of scientific research. Much of this work adopts the relational infrastructure perspective offered by Star and Ruhleder (1996). Over the last decade, the organizing concept of *cyberinfrastructure* (CI) has been used by these communities due to such projects' application of large-scale computing systems to often geographically distributed and interdisciplinary work (cf. Jirotka, Lee, & Olson, 2013; Ribes & Lee, 2010 for recent overviews). Scholars in these fields examine not only the technical challenges of developing such systems but also the complex social milieu in which the work of development takes place (Bietz, Paine, & Lee, 2013; Edwards et al., 2013; Faniel & Jacobsen, 2010). In addition, Lee et al. (2006) developed the notion of the *human infrastructure* of cyberinfrastructure to illustrate the complex, shifting ways individuals, groups, organizations, technologies, and data are aligned to enact CI.

Today, much related work has shifted from being referred to as studies of cyberinfrastructure to studies of data-intensive science or "big data." Our study of four scientific research groups examines their work with big data. In this paper, we focus on their work to process the data products that are produced or collected so that they are eventually able to do their desired analysis work. Thus, in the remainder of our literature review, we first examine work on the production, sharing, and reuse of data in scientific research. Since this work provides context for processing work, we then examine two models of the research data lifecycle.

2.1 Data Production, Sharing, and Reuse

Investigations of the production of data are necessary for understanding and supporting cyberinfrastructure use and development and big data or data-intensive science. Scholars across CSCW and Information Science foreground issues of the work that goes into data collection or production, transformation into a shareable product, and the issues of trust and understanding that hinder reuse by other scholars. Here we offer a brief overview of some of the prominent themes in this area.

Birnholtz and Bietz's (2003) work examines how data contributes to knowledge creation and to the community. They also note the importance of the context of production to data's use and reuse, and that from an economic perspective, data can be a source of rents or profit for researchers. Faniel and Jacobsen (2010) note that earthquake engineers in a cyberinfrastructure project were unable to reuse data without first understanding how it was collected or what a particular variable actually meant in the context of an experiment. Expanding upon this, Rolland and Lee (2013) illustrate the extensive practices cancer epidemiology post-doctoral researchers must use to assess whether a dataset is germane to their intended research projects, and the inquiries they must undertake to assess the production or collection process of the data provided. In addition, Borgman et al. (2012), note that "*data are the 'glue' of a collaboration, hence one lens through which to study the effectiveness of such collaborations is to assess how they produce and use data.*" Thus, following the data in scientific research is one successful strategy for inquiry to understand this collaborative work.

Finally, data in scientific research is often referred to in its initial state as "raw." This conceptualization denotes a data product's initial state where it is not yet usable for analysis for particular research questions. The volume *"Raw data" is an oxymoron,* edited by Gitelman (2013), brings together multiple pieces that emphasize that data is never "raw," always embodying the decisions of the individuals involved in its production along with the social circumstances. For example, Edwards' (2010) book length examination of meteorology and climate science research effectively illustrates the sociotechnical, sociohistorical, and sociopolitical nature of data and of research work overall. Edward's examination of the data, software, and model products that enable and drive climate science research is relevant to data-intensive research as it illustrates the complex processes that have led to modern climate science research.

2.2 Models of the Research Data Lifecycle

Scientific research is an iterative process where broadly: questions are formulated, research design is developed, data is gathered and analyzed, and knowledge is published as researchers engage in multiple rounds of different work processes to produce and refine data products to advance the state of knowledge in a field. Models of the research process often explicitly outline the lifecycle of the data involved in the work. Here, we examine two such models as background for our examination of data processing work in this paper. Each of these models recognize the iterative, back-and-forth nature of the stages of scientific research, in spite of their idealized presentation of a circular process.

The first model that we examine comes from the United Kingdom Data Archive (Archive, 2013). The United Kingdom Data Archive's model outlines a circular process of the data lifecycle, seen in Figure 1 - left. The UK Data Archive model of the data lifecycle offers lists of some particular activities that scientists engage in for each of these stages. In particular, this model defines data processing to include the following activities: entering, digitizing, transcribing, or translating data; checking, validating, and cleaning data; anonymizing data when necessary; describing the data; and managing and storing the data. Many of these activities overlap with work that is necessary and taking place in other stages. For example, managing and storing data is a constant process as scientists engage in their work with intermediate products throughout the research process.

The second model that we examine is presented by Wallis et al. (2008) from their work with an ecological sensor network cyberinfrastructure project, the CENS project. Wallis et al. identified nine stages of the general life cycle of this project, Figure 1 - right. These stages identified were: Experiment Design, Calibration and Setup, Capture or Generation, Cleaning, Integration, Derivation, Analysis, Publication, and Preservation. Wallis et al. note that the order of these stages, or steps in the research, are not absolute, with some being iterative and others parallel. Of these nine stages, three are noted as explicitly iterative and parallel with each other: Cleaning, Integration, and Derivation. These three stages are the processing work of the scientists in this ecology project. Cleaning is noted by Wallis et al. as "calibration and ground-truthing" of the information to normalize any calibration offsets from sensing equipment. Derivation is described as averaging of individual data points into new composite points for eventual integration. Integration, lastly, is noted as combining data points from multiple datasets by multiple researchers for multiple reasons.

The three steps that fall under processing work in Wallis et al.'s discussion of the data lifecycle in the CENS project and the activities offered by the UK Data Archive model illustrate the task of preparing data for analysis after it is collected.

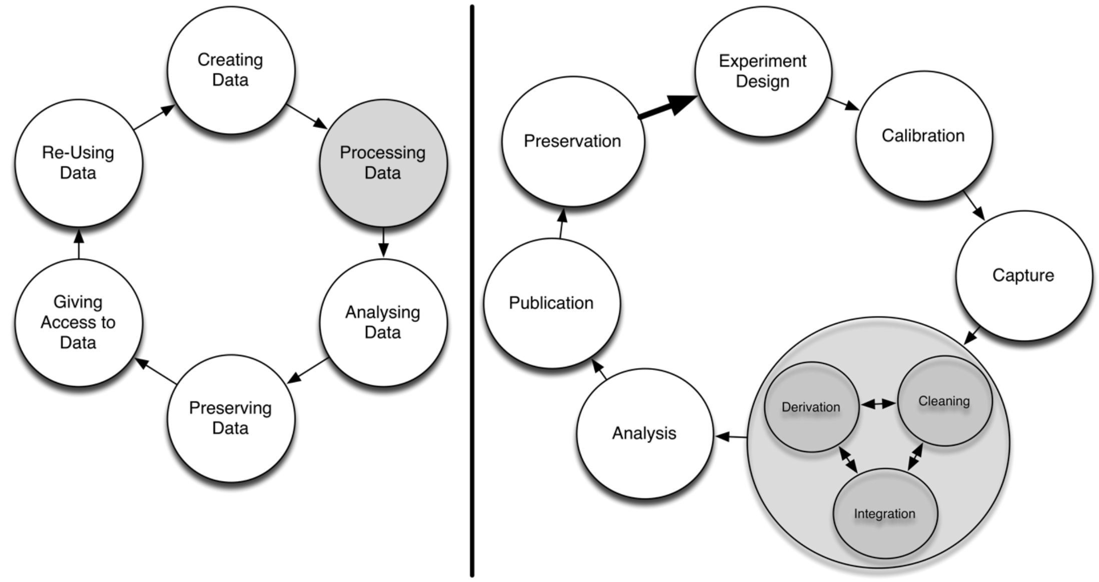

Figure 1. At left, The United Kingdom Data Archive's (2013) data lifecycle model with the processing stage highlighted in grey. At right, Wallis et al.'s (2008) lifecycle of CENS project data with processing stages highlighted in grey.

3 Research Sites and Methods

Through a multi-year qualitative study of data-intensive natural scientists taking place at the University of Washington in Seattle, WA; we further interrogate the role of data processing in scientific research. This study began in Autumn 2010 and continues to date. We are studying four research groups, one each in Atmospheric Sciences, Marine Geophysics, Microbiology, and Radio Astronomy. All four groups are led by a principal investigator (PI) who is a professor at the university. Here we introduce each research group and their work, before discussing our research methods.

3.1 **Research Sites**

Our study has four different scientific research groups enrolled. These groups were chosen for their use of large and increasing quantities of data, and pragmatically a willingness to participate over a five-year period or longer. Further details on our initial sampling and selection of groups is available in Paine et al. (2014). Studying individual research groups across four scientific disciplines allow us to begin to examine similarities and differences in work practices across diverse groups. However, we are limited in our ability to determine when such differences are in part due to variations in disciplinary practices. Developing our understanding further across and within disciplines is the subject of potential future work. Here we briefly introduce each research group, its members, and scientific field. We also note their sources of data and the specific projects we followed for each group. All groups, individuals, and projects are referred to here by pseudonyms.

The first research group in our study is led by Hank. Hank is an Atmospheric Sciences professor studying the interaction of “dynamics, radiation, and cloud processes” as they impact the global climate cycle. Hank and his group examine the effect of low-frequency changes in variables on the sensitivity of climate models to test hypotheses regarding factors changing the Earth’s climate. Hank’s research group is comprised of four Doctoral students (Anita, Bryan, Dane, and Palmer). All four PhD students in Hank’s group are responsible for work on individual dissertations. Hank and his doctoral students work with different climate models to evaluate how changing particular variables of interest affects the overall predicted state in the model. The group relies upon publicly available models and datasets with which to conduct their work.

Waldo leads the second group enrolled in our study. Waldo is a marine geophysicist in the Department of Oceanography whose research examines submarine volcanoes and mid-ocean ridge hydrothermal systems. Waldo and his research group use seismic techniques along with computational simulations to examine these phenomena. This research seeks to better understand how the Earth’s physical structure is changing by studying the focal undersea formations. Waldo’s research group is composed of three Doctoral students (Dahl, Rollin, and Megan). In addition, the group collaborates closely with PIs and students at another university in the region.

Waldo’s group works on undersea seismic projects where data is collected via an ocean-floor instrument network during a cruise to a focal area of the ocean. Waldo and his group work with data gathered from ocean expeditions to produce models of the Earth’s crust in order to develop a better understanding of its layers and structure. During the particular focal ocean expedition we are studying, the Ridge Experiment, the researchers placed ocean bottom seismometers in pre-determined locations. They proceeded to cruise across the region shooting off a series of air gun explosions in order to produce sounds waves that travel through the Earth’s crust. These waves are then reflected back into the recording equipment as seismic waves. The seismic waves are then recorded by the seismometers that were placed on the ocean bottom. Upon their return, Waldo and his group will work with this data for significant spans of time (at least a decade). They iteratively process and analyze this data to better understand the structure of each layer of the Earth’s crust.

Leading our third group is Martin, a microbiologist and virologist in the Department of Microbiology studying the Human Immunodeficiency Virus (HIV). Martin and his group engage in wet lab and computational biological research to examine the efficacy of vaccines and the evolution of different strains of HIV. The two projects we study in Martin’s group make use of the Roche 454 pyrosequencing technology to develop “deep” sequenced datasets of patient samples with the goal of obtaining better understandings of the evolution and effects of particular genes of interest. At the time of our study, the group was engaged solely in the computational analysis work for these projects, with wet lab work having taken place prior to our data collection.

Martin’s research group is composed of multiple doctoral and undergraduate students, postdoctoral researchers, and a large staff of research scientists. Across the two focal projects, work falls to one Doctoral student (Sharvani) and three research scientists (Brenton, Brenda, and Elisa). Sharvani and Elisa both work on the group’s first pyrosequencing project, PyroOne. Brenton is responsible for leading the second pyrosequencing project, PyroTwo, with help from Brenda and Elisa. Martin and his group conduct biological wet lab work to take samples from patients in HIV studies and sequence them for analysis. While the group does the work of taking a "raw" sample and producing a product that can be sequenced, we focus here solely on the computational processing work that is necessary to take the genetic sequence output from a particular hardware device and process it to be usable for different analyses.

Our fourth and final group is led by Magnus. Magnus is an empirical cosmologist in the Department of Physics studying the Epoch of Reionization through the development and application of novel radio telescopes. Empirical cosmology examines the different phases of the element Hydrogen across the electromagnetic spectrum in order to help us understand the origin and evolution of our Universe. Magnus and his research group in particular focus on the Epoch of Reionization (EoR)—a phase when Hydrogen was reionized by ultraviolet light from stars and galaxies. Studying the EoR enables cosmologists to better characterize the structure of the early Universe.

Magnus and his research group are primarily focused on the development of a data processing pipeline for a novel radio telescope. Magnus's group is composed of three post-doctoral researchers (Brianna, Igor, and Jonah), three PhD students (Abner, Nima, and Peg), and two undergraduate students. As a whole, the group's primary work at this time is the development of a data processing infrastructure. However, the undergraduate students are responsible for a smaller, local telescope project that we are not studying. Magnus's group is working with a new radio telescope, the Widefield Radio Telescope (WRT), to examine the Epoch of Reionization. Over the last year, and for the next few years, the WRT is observing the sky to gather hundreds of hours of data that will be processed. Processing the telescope observation involves executing multiple mathematical operations on multiple data products to produce a data product that allows them to measure the EoR.

3.2 Research Methods

The research groups are studied using three qualitative methods. First, when possible, observations of regularly scheduled group meetings are conducted to help us obtain and maintain an overview of each group's ongoing work. Second, two rounds of semi-structured interviews of members of each group have been conducted to help us learn about the research projects of each group and their ongoing work. Third and finally, artifacts such as project Wiki pages and websites, publications, software code, and public websites are examined, when available, to further flesh out our understanding of the projects discussed in our interviews. The findings presented in this paper are derived primarily from the second round of semi-structured interviews and fleshed out with the other data sources when necessary.

The two rounds of semi-structured interviews were designed to elicit information about the individuals and their work as members of their respective research group. The first round of interviews took place in Spring 2013 and the second round took place during Winter 2014. Each interview was recorded and professionally transcribed. The Spring 2013 interviews were focused on eliciting information about: each member's background and membership in the group; with whom they work; the projects on which they work; where they obtain data and how they analyze it; the software that they use to obtain and analyze data; and with whom the group shares its data and software. The Winter 2014 interviews were designed to go into greater depth with individuals about the different stages of their work on a given project. We discussed each individual's work to collect or produce, process, analyze, and archive data. In addition, we inquired about the software used to do so and the persons with whom they worked on these activities.

We have engaged in an iterative data analysis process to guide our ongoing data collection. We closed coded each Spring 2013 interview for the questions asked in our protocol, while also open coding for emergent themes regarding each individual and group's work (Charmaz, 2006; Emerson, Fretz, & Shaw, 1995; Weiss, 1995). We then wrote memos on each group's projects and work. This coding and memo writing informed our Winter 2014 protocol where we focused on the different stages of an individual's work and how they go about each stage.

For this paper, our Winter 2014 interviews were open coded in multiple rounds for discussions around each group's data processing work by three of the authors. Codes were merged after each round until saturation was achieved. Memos were then written describing the processing work for each group, how it is accomplished, who engages in this work, and what the data is being processed. Further memo writing was completed for other open codes as the authors discussed the findings that were emerging.

4 Findings

Here we describe the data processing work of the scientific research groups in our study by examining three common practices in such work. Once again we are examining ***data processing work***; that is the work that takes place between the collection or production of an initial data product in the research and its eventual analysis to address particular research questions and hypotheses. The three practices we identify include cleaning data products, selecting or assembling a subset of data, and transforming data into a common format. These three practices and the data processing work overall are not monolithic or one-time steps. Rather, they are components of an iterative process where initial project goals shift and clarify over time as researchers better understand and make use of their data as their research progresses. This enables them to bridge the data products that were initially produced or collected with the data products that are analyzable for specific research questions.

4.1 **Cleaning Data Products**

Verifying the quality of data is necessary in any research work and is a concern for each of the groups in our study. The cleaning of data products across all groups in our study is the work concerned with verifying the data gathered and removing bad data before it can impact the analysis process. Bad data are the elements in a data product that are undesirable to a research team and introduced through an error in an experiment process or hardware. Our participants note that bad data commonly results from errors in the experiment process or equipment in both expected and unexpected ways, i.e. radio frequency interference is expected at times in the Radio group's observation data while noise from faulty hardware is not. Cleaning work requires the individual researchers to apply their knowledge of the particular experiment design and the data production infrastructure to the resulting data products to remove undesired data.

In Martin's group, the sequencing platform chosen for the two projects we studied had design consequences that necessitated cleaning of the data that they produced. This cleaning process was completed for both of the projects over multiple months as the sequencing work was completed and software pipelines (software products where multiple components are connected via scripts and intermediate data products) were developed. This process went back and forth between testing the pipeline being developed and evaluating the intermediate data products that were produced as a particular sequence is cleaned. This iteration was necessary to help the group determine if artifacts of the experimental process were appropriately removed from the data products or whether they needed to be cleaned further. Iteration was also necessary to ensure that the operation of the different components of the software pipelines were understood and working as desired.

The genetic sequences that Martin's group used were produced using the Roche 454 pyrosequencing platform. This was a new sequencing technology for the group. In the past, sequences were produced with the well-established Sanger method. During both of our interviews, Brenton, a Research Scientist, noted that the 454 technology produces sequences that are "*riddled with errors that are … well established*" and necessitate cleaning work as group members refer to it. These "well established" errors lead to many insertions and deletions in the sequence reads. For example, there may be a region that should be read as AAAA, but is instead read by the machine as AAA or AAAAA. These insertion and deletion errors were a result of the design of the sequencing hardware and associated process. This partially led to the need to design new software pipelines to apply to this data. Automated cleaning could only take the group so far. Sequences often still had to be verified and cleaned by hand.

The volume of data from pyrosequencing is such (millions of reads for a given sequence is possible) that manually cleaning and aligning sequences for all of the reads would simply not be possible. In the case of the PyroOne project, this led to Sharvani, a PhD student, designing in a threshold in her software pipeline for throwing out bad reads or misalignments. This helped to reduce the amount of "painful" hand alignment that must be completed since fewer sequences were being output incorrectly. However, manual verification and cleaning of selected sequences was still necessary to maintain the level of data quality that the PI requires and to help the individual researchers understand the different patient's sequences. Cleaning work is a necessary bridge in Martin's group to take the "raw" data output by the 454 sequencer to produce a data product suitable for statistical analyses to be conducted.

Similarly, Magnus's group and their international collaborators must flag and remove bad data as it is gathered with the Widefield Radio Telescope. The scientific goals of the different projects using the WRT led to an instrument that is designed to be sensitive to particular radio frequencies. Interference at different radio frequencies occurs and may be captured by the telescope in spite of its intentionally remote and relatively radio "quiet" location on Earth. For example Brianna, a Post-Doc, notes that the collaboration "*know[s] that we can see significant RFI [Radio Frequency Interference] bounced off the* *moon.*" In addition, different components of the telescope may be malfunctioning during an observation, since it is composed of over one hundred different antennae and associated hardware. The processing pipeline being developed by Magnus and his group must therefore account for such malfunctions by accounting for metadata produced about the instrument during the observation process. Due to these two issues, among others, data is flagged and removed (i.e. cleaned out) at multiple stages of Magnus's group's processing software pipeline. As the pipeline has been developed, this flagging and removal has been iteratively added as the group develops a better understanding of the telescope when it is operating.

In addition to Martin and Magnus's groups, we see that cleaning is necessary in Waldo and Hank's groups to progress from data collection or production to analysis work. Waldo's group had to clean the Ridge Experiment data when it was brought back to shore from the ocean expedition before processing work could begin for Rollin and Dahl's (PhD students in the group) particular research goals. Specifically, a Masters student was tasked with verifying the waveforms collected and verifying the physical location that the project's instruments landed on the ocean bottom. This verification is necessary to enable Waldo's group to further process the data. Students in Hank's group must verify the data that is in the different data products and models they are working with. For example, Hank's students must examine the different units variables are stored in and the temporal ranges they cover, among other concerns, so that they will be able to successfully select a subset of data, as we discuss in our next subsection. This verification by members of Hank's group is necessary since they are often gathering multiple data products to use with a given climate model in their dissertations.

The cleaning activities that researchers in each of our groups engage in are necessary for ensuring that a useful and usable data product will be ready for their analysis work. Wallis et al. (2008) described cleaning activities in the CENS project as "calibration and ground-truthing" to normalize the data for the project. We also see this in the work of Waldo's group as they verify the initial data collected before they can move on to conducting other processing activities in advance of their eventual analysis work. In the work of Magnus and Martin's group, we find cleaning to be a necessary task due to artifacts of the experimental process skewing the data and to help the researchers understand the data they have collected. Cleaning in Hank's group is necessary to not only understand the variables available in a particular data or model product, but also that it is stored in units appropriate to the task at hand. All of the cleaning work requires the individual researchers apply their knowledge of the experiment's data collection or production process along with that of the instruments utilized to this processing work. The cleaning work that is part of these groups' data processing work thus repeats and evolves as the researcher's understanding grows, particularly when they are engaging in other data processing practices.

4.2 **Assembling or Selecting a Subset of Data**

The second common data processing work practice that we see with our scientific groups is the selection or assembly of a subset of data, out of a project's larger collected sample of data, to address a particular research question. The data being gathered for each group's projects is extensive and can be used to answer many different research questions. A key concern for the individuals interviewed, was assembling a new data product using data from multiple sources or selecting a subset of a project's larger dataset, specifically for the questions they are examining. This process is vital for these groups to be able to take their data products and use them for analysis, since the necessary data may be spread across multiple data products or part of a larger data product where some of the data might be irrelevant to the question at hand, slowing down the eventual analysis work. This selection or assembly work often follows an initial round of cleaning work, while in turn necessitating further cleaning as data is selected or transformed into a common format (discussed in the next subsection) due to a better understanding of the data itself.

The members of Hank's group are pulling different pieces of data from multiple data products that they purposefully collected to assemble a new data product. This newly assembled data product will be used to execute a particular climate model as they iterate between processing data products and analyzing the outputs of climate models that the data is fed into. Bryan, Palmer, and Dane as PhD students all pull data from multiple data products and use different variables to address the questions they are examining in conjunction with a particular climate model. This is necessary since the desired variables may either be unavailable in one data product (or model) or not cover the desired temporal or geographic range. In addition, a subset of these variables may end up being selected to constrain the problem space to the desired geographic or temporal range once again.

For example, processing work for Bryan's experiment involves assembling the necessary variables from different datasets and configuring the model he is executing by setting different measurements. Bryan notes that he must configure the size of grid data that data is computed with or the levels of carbon dioxide or ozone for a particular execution of the model. Bryan describes writing scripts in MATLAB to pull different pieces of data together from different data products (assembling a new data product) to be averaged, so that the resulting products can then be transformed into a common format and supplied to the actual FORTRAN-based model as starting or bounding conditions. From here the model can be executed to create the data products that can be analyzed.

In contrast to the students in Hank's group, the researchers in Waldo, Magnus, and Martin's groups select subsets of data from their experiment's overall data products to answer their particular research questions. The PhD students in Waldo's group spend significant amounts of time selecting a subset of the Ridge Experiment data points for their particular research questions. Each student describes picking tens of thousands of data points from the larger data product as part of their processing work. This selection is necessary to pull out the data relevant to the research questions that their dissertations are attempting to address.

Rollin's work offers one example of this selection process. Rollin's processing work requires that he manually pick out the arrival time of waveforms for each air gun shot. For his stage of the project, there were more than 90,000 such "picks" to be done manually. This picking takes place using a MATLAB toolset that Waldo (the PI) has built, and made freely available. According to Rollin, this provides a "*nice little GUI and then we can go through and look at the arrival of each packet of energy from those air gun shots, and you just make little clicks along in the GUI in order to identify where that arrival is.*" This manual work could potentially be automated, but Waldo and the group elected to proceed with the manual process on this project since the students were not yet familiar with the type of data produced and there was uncertainty regarding the development of algorithms to select the desired data from the full set of data. The picks that are made are then fed into the computational model of the structure of the part of the Earth's crust that Rollin is studying. Rollin's process of picking data points was completed iteratively as he developed his model of the Earth's crust; moving between development of the model, selecting subsets of data through this picking process, and preparing the data points for the model. This iteration not only improves the data being fed into his model, but also helps him develop his knowledge of the Earth's structure being studied and how his model is capturing or failing to capture that knowledge. Over time, the model's resolution of the phenomena of interest improved as more data was fed into it.

Similar to Rollin (and the other researchers in Waldo's group), once the researchers in Martin's group have cleaned their sequence data products, they must select the data points for the particular genes relevant to a given research question. The larger data products in each of these groups were produced to support multiple research questions necessitating this selection of relevant subsets for particular questions. The researchers on a given project in Martin's group will go back and forth between cleaning work, selecting a subset of a particular sequence, and engaging in the analysis for a particular research question of interest with the cleaned data product. Iterating back and forth between practices is necessary since the wet lab work to produce sequences, the data processing work, and the development of the data processing software pipeline all proceeded at different paces, and they do not always have raw data ready to be processed for a particular project.

Finally, Magnus and the Radio group (along with their collaborators) also work with a subset of their data at this time. In contrast to the above example, their use of a subset is currently motivated by their active development and debugging of their data processing software pipeline. These researchers must develop such a pipeline to be able to answer their research questions because the volume of data being collected, and consequently that has to be processed, is too great for existing solutions. In addition, this pipeline's design accounts for the particulars of their experiment design and its hardware (see Paine & Lee, 2014 for detail). The collaboration's selection of a subset of data allows the researchers to ensure that they are testing and debugging a common set of data. Over time their understanding of this subset of observations grows and this supports their iterative development process since they are better able to determine when an anomaly in their output is an artifact of their processing pipeline or the actual observational data. In addition, the collaboration uses the knowledge obtained from this subset of data to help further focus its ongoing data collection to areas of particular scientific relevance.

The selection of a subset of data from a larger data product or assembly of a new data product from multiple data products is a common concern and practice across the four groups in our study. Members of Hank's group must assemble a new data product by selecting variables from multiple data products while members of Waldo and Martin's groups each work with subsets of the group's larger data products. Magnus's group and their collaborators work with subsets of their data as they develop and test their overall processing pipeline. In all four cases, this work requires that the researchers understand the data that is available to them and germane to the question they are attempting to answer. The selection and assembly work in these groups often leads to further cleaning being necessary or the transformation of data into a common format (our third practice).

4.3 Transforming Data into a Common Format

The third and final data processing work practice that we find, is the transformation of data into a common format to be used in subsequent analyses or as input to computational models. The analyses of Hank and Waldo's group's directly take place using the outputs of computational models that are either executed or developed by the researchers in the groups and require data be transformed into a common format that can be supplied to the model. Magnus and Martin's groups are developing and applying processing pipelines. Key to these software systems is the transformation of data products into common formats to be used across components of the pipeline. In this subsection, we examine the work these researchers engage in to ensure that their data is in the appropriate format for their models or processing systems that their analyses will be conducted with. This work builds upon the first two practices we identified by completing further operations to transform the data to the appropriate units or incorporate it into a computational model.

Bryan, Dane, and Palmer in Hank's group are each using a particular climate model for their dissertation research. In addition to selecting a subset of variables from different data products, as we discussed in the previous section, these researchers must further process these variables to ensure that the data is appropriate and in a common format for use with the model. A commonly noted task was the need to average variables to produce a newly derived variable. In addition, variables must often be converted to the proper units since a given data product may store the data in different units based on the conventions that particular product uses, which often will differ from that of another climate model. Bryan describes writing scripts in MATLAB to pull different pieces of data together from different data products to be averaged so that the resulting products can then be supplied to the actual FORTRAN-based model as starting or bounding conditions. From here, the model can be executed to create the data products that can be analyzed.

In contrast, Rollin and Dahl from Waldo's group are developing computational models of the Earth's crust using the data from their project's ocean expedition. The manual picking work that we saw in our last subsection, to pull out a particular subset of their data, is necessary for of each these researchers to be able to prepare data for their individual models. As they develop and iterate upon their models, their outputs become inputs for the model of the next lower layer of the crust. Work on the different models proceeds simultaneously, so that over time, the improvement of a model of the upper layer will improve the lower layer model. Researchers in Waldo's group prepare data for their computational models through their cleaning work and their selection of a subset of data, while also implementing algorithms to incorporate this data into the model and, depending on the layer, connect to another researcher's model. Each individual must therefore ensure that their computational model, and its associated data, is prepared well enough to be used by the other researchers in the collaboration. For example, Dahl is developing a model of layers of the Earth's crust below the layer that Rollin's model covers. Dahl notes that "*[Rollin's] problem had to be solved really before I [Dahl] could move too far forward in my problem*"; otherwise, it would be difficult and problematic to try and understand the lower layers of the crust without first understanding the upper layers. Rollin emphasized this notion, commenting that what is nice "*is that what I do in my portion of it acts as a springboard into things that can be done with other pieces of the project.*"

In our third group, Martin's microbiology group, Sharvani describes how most of the different components of her processing pipeline work with FASTA sequence files. This is a common format for genetic sequence data. However, at least one component produces output files in a different format, the ACE format. Sharvani bemusedly noted that the software component "*clearly wouldn't give an output where we could just open it up [with the pipeline],*" requiring her to develop another component to transform the data product from ACE files back into the FASTA format in order to be usable in the pipeline. Similarly, the data processing pipeline Magnus's group is developing must transform data products between different formats. The first stage of the pipeline produces data products in the fairly generic UV FITS format, while other components work with IDL save files or regular FITS files. Different components of the pipeline therefore must transform the data products between these different formats. At times this is simply a matter of reading in data and writing it out in a different format—similar to Sharvani's conversion of ACE files back to FASTA files. Additionally in our fourth group, Magnus's group, different pieces of the processing pipeline operate with data in different mathematical frames of reference. For example, some operations apply algorithms to the data in 3 dimensions while others work in 2 dimensions. The pipeline is therefore designed to transform the data into the appropriate format for each stage and output the different data products.

The transformation of data for use in computational models or processing pipelines is necessary to enable their use for particular analyses. Without this processing work, the data that was produced or collected could not be analyzed. The selected data may need to be converted to the appropriate units or averaged together to be of use for the chosen model—the derivation and integration activities noted by Wallis et al. (2008). Much of the preparation work is contingent upon the researchers having engaged in cleaning work or the assembly and selection of a subset of data at least once. This will often lead to further iterations as they better understand the data they have available and the question they are trying to answer. The work across these three practices is not disconnected but rather intertwined, requiring a researcher to iteratively complete these tasks as they better understand the problem they are studying and the data products at hand to be able to successfully address their research goals.

5 Discussion and Conclusion

Across all four research groups, we see that these three practices of data processing work are integral to bridging the gap between the initial data products and the scientifically analyzable data products of a project. The production and collection of data results in data products that must be further processed in order to be usable for particular analyses, since they may contain artifacts of the experiment or fieldwork processes. In addition, the data that was collected or produced may need to be transformed into a common format when preparing it for use in a computational model, processing pipeline or a subset selected for the particular research question at hand.

The three practices we see (cleaning, selecting subsets of data, and transforming data into a common format) highlight the necessity of accounting for, and understanding error throughout, the research process. Waldo's group faces this task during their ocean expeditions when attempting to produce a dataset since the product must be scientifically usable for a decade long time span. Many scholars in CSCW and Information Studies stress the importance of knowing and understanding the process of producing data to its use for analyses while also noting the lack of planning for long-term archiving (Birnholtz & Bietz, 2003; Borgman, et al., 2012; Wallis, et al., 2008). By examining data processing work, we see the concerns and practices in which researchers engage to produce scientifically useful data products for the immediate goals at hand.

Processing work takes the initial data products and refines them for a particular scientific goal. Removing bad data and validating what is kept through cleaning is an important and necessary component of data processing work since experimental processes introduce errors, which must be systematically accounted for to enable researchers to move forward with their work. The software products used in processing work are also a potential source of additional error. The heart of much of the processing work in Magnus, Martin, and Waldo's groups is balancing what is accomplished through computational versus manual means. The volume of data in Magnus's group is so large that manual processing would never be a feasible option, necessitating the extensive work to develop and test their processing pipeline. This work is also fundamentally necessary as they work to better understand the operation of the telescope that they are working with – as a novel system – in conjunction with the particular body of scientific knowledge driving the work. In contrast, Martin and Waldo's groups – in spite of significant volumes of data – are still processing much of their data manually, since algorithms have not been found to be reliable or the individuals are still working to understand the actual outcomes of their experimental process.

Across the work of collecting, processing, and analyzing data, the line between one stage of the research and another is often blurred. The two models of the research data lifecycle introduced earlier recognize that different types of processing work are necessary to bridge data collection or production and analysis work. They also state that research work does not always progress cleanly from one stage to another, as our findings also illustrate. However, what constitutes data collection, processing, or analysis is not always perfectly delineated. The three practices presented here as data processing work might simply be the grunt work of analysis or production to some researchers. We find that the amount of time spent on processing work – and its necessity to connecting the initial data products produced and those that are analyzable – warrants foregrounding and emphasis as a particular stage of the research process that should be examined. For example, the work of Magnus's group, at this time, continually iterates between development and testing of their data processing pipeline and analysis of the plots and data products it outputs. Pre-analysis of data products that are output at different stages is necessary to validate whether the processing is being accomplished as expected and to improve the group's understanding of its data before eventual analysis work is undertaken to address particular questions and hypotheses.

The two models of data lifecycles introduced earlier both note that iteration between stages often takes place (Archive, 2013; Wallis, et al., 2008). However, neither of these models explicitly recognized that iteration takes place all the way back to a project's initial research design nor does either model actually examine the ways in which any of the particular stages iterate back and forth between each other. As data is collected or produced, processed, and analyzed, the researchers in our study constantly and iteratively check what usable and valid data they have, what the data actually supports, and what questions continue to be supported as the data is further processed and analyzed. This ongoing data processing work drives iterations *beyond* data processing work itself, and thus helps constrain the scope of answerable research questions while bridging the data collection and analysis work of these scientific researchers. We see that iterating between processing data and analyzing data is necessary for these scientists to simultaneously develop an improved understanding of the scientific research problem and also an account of the possibilities afforded by the collected data, and this depends on the design and quirks of the instruments employed along with the nuances of their actual use. Rollin, one of the PhD students in Waldo's group, likened this refinement of the data and his understanding to bringing a camera into focus:

> "And then as your data gets better and more refined then you can go through and sharpen that image by tweaking different parameters and making it more fine scale. It's kind of like bringing a picture into focus on your camera or something. You start off with something that's kind of blurry and then you can fine-tune things in order to make it sharper."

Data processing work is an extensive process in the scientific research process. The three practices we see across four scientific research groups enable each to refine their data products and better understand the data available for their research work. In spite of the involved, often manual, and perhaps arduous nature of this work, it is surely not "janitorial" (Lohr, 2014) as some might frame it. Rather, data processing work requires methodical practices that are integral to the success of research work, even if they do require significant amounts of time and effort as the quantity of data available, and in use, increases. While data processing work may not be a new task in the course of research, most examinations of research work focuses primarily on the production or collection of data or the work to analyze it. Processing work is often subsumed in these discussions. However, we find that foregrounding processing work is important for understanding much of a researcher's day-to-day work in data-intensive research.

6 References

Archive, U. D. (2013). Research Data Lifecycle Retrieved July 18, 2013, from http://data-archive.ac.uk/create-manage/life-cycle

Bietz, M. J., Paine, D., & Lee, C. P. (2013). *The work of developing cyberinfrastructure middleware projects*. Paper presented at the Proceedings of the 2013 conference on Computer supported cooperative work, San Antonio, Texas, USA.

Birnholtz, J. P., & Bietz, M. J. (2003). *Data at Work: Supporting Sharing in Science and Engineering.* Paper presented at the GROUP, Sanibel Island, Florida, USA.

Borgman, C., Wallis, J., & Mayernik, M. (2012). Who's Got the Data? Interdependencies in Science and Technology Collaborations. *Computer Supported Cooperative Work (CSCW), 21*(6), 485-523. doi: 10.1007/s10606-012-9169-z

Charmaz, K. (2006). *Constructing Grounded Theory: A Practical Guide Through Qualitative Analysis*: Sage.

Edwards, P. N. (2010). *A Vast Machine: Computer Models, Climate Data, and the Politics of Global*: MIT Press.

Edwards, P. N., Jackson, S. J., Chalmers, M. K., Bowker, G. C., Borgman, C. L., Ribes, D., Burton, M., & Calvert, S. (2013). Knowledge Infrastructures: Intellectual Frameworks and Research Challenges. Ann Arbor.

Emerson, R. M., Fretz, R. I., & Shaw, L. L. (1995). *Writing Ethnographic Fieldnotes*. Chicago, IL: The University of Chicago Press.

Faniel, I., & Jacobsen, T. (2010). Reusing Scientific Data: How Earthquake Engineering Researchers Assess the Reusability of Colleagues' Data. *Computer Supported Cooperative Work (CSCW), 19*(3), 355-375. doi: 10.1007/s10606-010-9117-8

Gitelman, L. (Ed.). (2013). *"Raw data" is an oxymoron*. Cambridge, MA: MIT Press.

Jirotka, M., Lee, C. P., & Olson, G. M. (2013). Supporting Scientific Collaboration: Methods, Tools and Concepts. *Computer Supported Cooperative Work (CSCW), 22*(4-6), 667-715. doi: 10.1007/s10606-012-9184-0

Jirotka, M., Procter, R. O. B., Hartswood, M., Slack, R., Simpson, A., Coopmans, C., Hinds, C., & Voss, A. (2005). Collaboration and Trust in Healthcare Innovation: The eDiaMoND Case Study. *Computer Supported Cooperative Work (CSCW), 14*(4), 369-398. doi: 10.1007/s10606-005-9001-0

Lee, C. P., Dourish, P., & Mark, G. (2006). *The Human Infrastructure of Cyberinfrastructure.* Paper presented at the CSCW, Banff, Alberta, Canada.

Lohr, S. (2014, August 18, 2014). For Data Scientists, 'Janitor Work' Is Hurdle to Insights, *The New York Times,* p. B4. Retrieved from http://www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html?_r=2

Paine, D., & Lee, C. P. (2014). *Producing Data, Producing Software: Developing a Radio Astronomy Research Infrastructure.* Paper presented at the eScience 2014, Guarujá, São Paulo, Brazil.

Paine, D., Sy, E., Chen, Y.-Y., & Lee, C. P. (2014). Surveying University of Washington Research Leaders Regarding Advanced Computational Needs *Computer Supported Collaboration Laboratory Technical Reports*: University of Washington.

Ribes, D., & Lee, C. (2010). Sociotechnical Studies of Cyberinfrastructure and e-Research: Current Themes and Future Trajectories. *Computer Supported Cooperative Work (CSCW), 19*(3), 231-244. doi: 10.1007/s10606-010-9120-0

Rolland, B., & Lee, C. P. (2013). *Beyond trust and reliability: reusing data in collaborative cancer epidemiology research*. Paper presented at the Proceedings of the 2013 conference on Computer supported cooperative work, San Antonio, Texas, USA.

Star, S. L., & Ruhleder, K. (1996). Steps Toward an Ecology of Infrastructure: Design and Access for Large Information Spaces. *Information Systems Research, 7*(1), 24.

Wallis, J. C., Pepe, A., Mayernik, M. S., & Borgman, C. L. (2008). *An Exploration of the Life Cycle of eScience Collaboratory Data*. Paper presented at the iConference 2008. http://hdl.handle.net/2142/15122

Weiss, R. S. (1995). *Learning From Strangers: The Art and Method of Qualitative Interview Studies* (Vol. The Free Press). New York, NY.

7 Table of Figures