Producing Data, Producing Software: Developing a Radio Astronomy Research Infrastructure

Drew Paine
Human Centered Design & Engineering
University of Washington
Seattle, WA 98195 USA
pained@uw.edu

Charlotte P. Lee
Human Centered Design & Engineering
University of Washington
Seattle, WA 98195 USA
cplee@uw.edu

***Abstract*—The production and use of software pipelines is a key component of much modern scientific research. We present emerging findings from our qualitative, social science study of a radio astronomy group developing software pipelines as they produce a data processing infrastructure. This paper examines how these researchers co-produce data products and software pipelines to enact their research infrastructure. We investigate the work of co-producing data and software to illustrate that to better support data-intensive science, we need to understand the practices that enable and produce data products, software, and ultimately infrastructures.**

Keywords—data production; scientific software pipelines; computer supported cooperative work (CSCW); qualitative methods

I. INTRODUCTION

The ability to share data, and increasingly software, is necessary for conducting many kinds of scientific research and for the replicability of findings. The eScience community has a well-established history of designing and developing software pipelines for scientific data production, processing, and analysis [1,4,6]. These pipelines often enable the sharing of data. In addition to the eScience community, the Computer Supported Cooperative Work (CSCW) community studies the development and application of these software systems, often examining the social and organizational work of creating and maintaining infrastructure [3,8,17]. Furthermore, multiple scholars note the importance of the context of the production of datasets to their sharing and reuse by other researchers, although they primarily focus on issues of trust [5,9].

The existence of tensions between developing customized and generalized systems are well recognized [15]. However, little research has explored in detail the conditions under which infrastructural pipelines are created and how they are shared and used to promote data sharing. We present findings from our examination of a radio astronomy group's work to develop and apply software pipelines in the course of their data-intensive work that begins to address this gap. This type of work investigates both the social and technical, also known as a "sociotechnical" approach. We investigate the research question: *How do radio astronomers design their data and software within a larger experiment's infrastructure?*

A goal of the development and application of software pipelines for scientific research is often to produce as generalized of a system as possible. In an idealized world a given pipeline would readily be able to take in a dataset and some metadata and process it to enable analysis for a given scientific goal, all with minimal manual work required on the part of the scientist. Bietz and Lee's [2] examination of the CAMERA cyberinfrastructure project notes some developers' dream of an "*ideal database*" (p.8) to store all possible genetic sequence data. However, such a system is not feasible due to the myriad research questions and work practices of different researchers and developers.

In this paper we use the term *data product* to refer to the datasets that are the output of instruments or of executions of the processing and analysis infrastructure being created. *Intermediate data products* refer to data outputs that have been processed by a stage of a pipeline but must be further processed to be useful to the scientific goal. Intermediate data products may be shareable with other researchers who would have to do further processing of their own while still accounting for many nuances of the original dataset's design. We use these terms to emphasize the living and in-process nature of these artifacts as products of the research process. In contrast, the term dataset implies a "finished" artifact that is less subject to the vagaries of an emerging and evolving infrastructure but would, however, in theory be more shareable.

Below we show ways in which data products and software are alternately and simultaneously input, output, and processed. We illustrate how all of these things are inextricably intertwined and woven together in the practice of doing research. The increasing scale of data produced from advanced instruments, and the initial design of such experiments and the design of the desired data products that are to be created, necessarily impacts the software and intermediate data products that are developed for and through processing work. Crucial to this co-production are the design decisions made for the scientific experiment and instrument overall. These decisions have enduring impact as scientists grapple with the "big data" that is produced by such experiments.

II. LITERATURE REVIEW

Scientific workflow and data analysis pipeline systems are a prominent topic in the eScience community's research. Belhajjame et al. [1] discuss workflow systems and their specifications. They note that studying the development and

use of these systems is important since workflow systems "*encapsulate knowledge that documents scientific experiments*." Killeen et al. [10] offer a case study of an effort to assemble a workflow system from existing technologies to combine data from many different sources. Finally, Darby et al. [4] note that research data sharing is one of the key challenges of the eScience era.

The Computer Supported Cooperative Work (CSCW) community, a field generally comprised of computer scientists and social scientists working together, also has an extensive history studying the development and adoption of computing systems for scientific research, especially under the banner of Cyberinfrastructure (CI) or Collaboratories (cf. Jirotka, Olson, and Lee [8] for a recent overview). One key perspective is that of Star and Ruhleder [17] who note that infrastructure is inherently relational, with components that are embedded inside of other systems and infrastructures, and are learned as part of membership in a community. Lee et al. [12] furthermore describe "human infrastructure" which consists of different forms of organizing among stakeholders that overlap and change over time and often used in parallel to create and maintain CI.

Bietz et al. [3] examined how multiple stakeholders sustain the development of CI middleware systems at two supercomputing centers where stakeholders engaged in sociotechnical work, as part of the human infrastructure, to develop new software systems for scientific use. Bietz et al. and other CSCW scholars [11,15] studying cyberinfrastructure development commonly note the tension of balancing research versus development priorities. Engaging in research work versus developing systems is an ongoing challenge scientists and systems developers face when producing various software systems for scientific use.

Finally, Jirotka et al. [9] and Faniel and Jacobsen [5] both note the importance of the context of production to the sharing of datasets in eScience work. These scholars both emphasize the issues of trust that arise in eScience projects, and other situations, regarding the sharing of data. Rolland and Lee [16] go beyond trust issues to illustrate the data practices cancer epidemiology post-doctoral researchers actually employ when reusing existing data. These practices illuminate that a much wider lens is necessary to examine researchers work reusing existing data and importantly how the data was produced in the first place.

III. Research Site and Methods

The findings presented in this paper come from our multi-year qualitative study of data-intensive scientific research groups at the University of Washington in Seattle, WA. This study is examining the change in scientific research practice as technology evolves and datasets increase in size over a five-year period. Two of the five research groups, a radio astronomy and a microbiology group, are the research sites for the first author's qualitative dissertation research. This dissertation research focuses on the scientific group's development and use of software in conjunction with hardware systems and datasets. We discuss one of these sites in this paper. This group, its members, and its projects are referred to using pseudonyms.

A. Research Site: A Radio Cosmology Group

The research site we examine is a radio astronomy group engaged in empirical cosmological research that we refer to by the pseudonym Radio. Cosmology as a field aims to understand the origin and evolution of our universe. Key to empirical cosmological research is studying the different phases of the element Hydrogen that are visible to us across the electromagnetic spectrum. One such phase is the Epoch of Reionization (EoR). The EoR is a period in the Universe when primordial stars and galaxies emitted ultraviolet light that reionized the hydrogen in the universe. Studying the EoR requires the development and use of cutting edge radio telescope hardware and software along with the production of significant volumes of data.

One such telescope is the Widefield Radio Telescope (WRT). The WRT is the effort of an international radio astronomy collaboration. The WRT was designed for multiple radio astronomy objectives (at least four areas were originally planned), but it is biased towards EoR research. The Radio group contributes to the WRT as part of the International EoR group, visible in Figure 1 in bold, within the US EoR group. The design decisions made in the late 2000s during planning and construction necessarily constrain the work of the Radio group today, as we will illustrate with our findings.

The principal investigator of the Radio group is Magnus. As of Spring 2014 the Radio group also has three post-doctoral researchers, three PhD students, and two undergraduate students who are active in the research projects being undertaken. The Radio group's work at this time focuses on the development of one of two data processing and analysis software pipelines for EoR science using the WRT. We refer to this pipeline as the US EoR pipeline. Scholars in another country are responsible for developing the second EoR pipeline. By design certain software pipeline components and intermediary data products from each pipeline are meant to be interchangeable. This enables both groups of researchers to evaluate the integrity of each of their intermediary products as they work towards a scientifically analyzable dataset.

B. Research Methods

The Radio group is being studied using three qualitative data collection methods. First, observations are taking place of the group's regularly scheduled meetings. Second, two rounds of semi-structured interviews have taken place with four members of the group to date. Finally, artifacts such as project Wiki pages, publications, software code, presentations, public websites, and email threads are being captured for analysis.

Artifacts from the Radio group have been captured from December 2013 through the present. Meeting observations have taken place periodically from 2012 through the present. The majority of the observations took place between January 2014 and May 2014. Semi-structured interviews of the group

members have taken place in two rounds; the first round in Spring 2013 and the second in Winter 2014. The same four members were interviewed for each round. The Spring 2013 interviews lasted between 33 and 65 minutes (avg. 49 minutes). The Winter 2014 interviews lasted between 56 and 125 minutes (avg. 81 minutes).

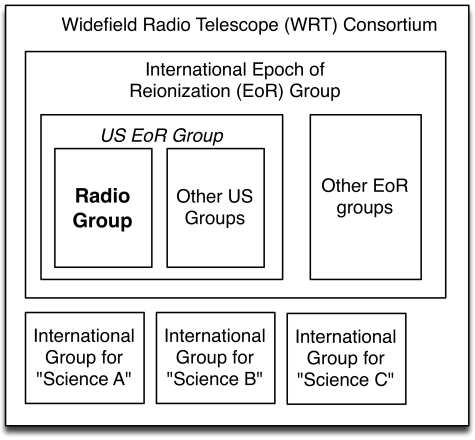

Figure 1. Diagram of the Widefield Radio Telescope (WRT) consortium's four science groups with the Radio group bolded.

The Spring 2013 interviews elicited information about: each individual's 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 used to obtain and analyze data; and with whom the group shares the data and software developed. The Winter 2014 interviews investigated the work that each individual does on his/her project, the software s/he uses and develop for this work, and with whom they work. Each interviewee walked us through their work to collect or produce, process, analyze, and archive data.

Data analysis for each round of data collection has involved coding and thematic analysis for answers to the questions directly asked [13]. As we have worked to trace the network of software and data involved in each person's work we have proceeded by memoing on each individual's work as a member of the project [7]. As each individual's story emerged, we supplemented the data from our interviews with the artifacts collected, such as plots or software code, to round out our understanding of the work. The findings and themes presented in this paper emerge directly from Winter 2014 interviews and are supplemented with our other data.

IV. FINDINGS

Our data reveals how a data processing infrastructure is created as this radio astronomy group iteratively produces software and data. At the same time, members of this group must continually work with the consequences of the WRT experiment's design and construction—a set of design decisions made years earlier. The Radio group's research is producing substantial volumes of data, at least half a petabyte or more per observing season. To process and analyze this massive volume of data with a sufficient degree of accuracy, the Radio group must produce new processing infrastructure as part of the WRT experiment. The Radio group must develop their own processing infrastructure, the various data products that are produced by and used within this infrastructure, and the processes that enable these data products and the software pipelines to unfold together through a series of interconnected design decisions. This work enables the Radio group to pursue their scientific goals. First we briefly introduce the components of this pipeline as background information.

The data processing pipeline the Radio group is developing is composed of three main pieces. We refer to these pieces as ObservationControl, Calibrator, and Power. The first component (ObservationControl) handles the initial conversion of the data output from the telescope's hardware and software systems by executing an externally produced piece of software (FlagAvg). Abner is the PhD student responsible for developing and maintaining this stage. He is also responsible for executing major runs of the pipeline on the group's shared computing cluster. ObservationControl is a series of Bash shell scripts and the C-based FlagAvg pre-processing software that is being developed by another member of the international WRT collaboration.

The second and third components of the Radio group's pipeline, on the other hand, are developed in Interactive Data Language (IDL). It is important to note that the choice of the IDL programming language is a potential hindrance to the sharing of the infrastructure that the Radio group is developing but is an expedient choice for this group due to the two postdoctoral researcher's familiarity with the language. The second component (Calibrator) of the pipeline does the bulk of the processing work on observations to prepare them for analysis. This processing work includes tasks such as: setting up the sky coordinates that the data should be placed in, setting up a model of the telescope's radio beam, calibrating the data against these coordinates and model, and producing images. Igor is the post-doctoral researcher developing Calibrator. Finally, the third component (Power) of the pipeline produces power spectra. Brianna is the post-doctoral researcher producing Power. The creation of power spectrum of radio waves is how cosmologists measure the Epoch of Reionization. As such, Power is the stage where much of the long-term scientific goals are encapsulated. Equivalent stages are being produced for the second and third pipeline elsewheres in the international EoR collaboration.

A. The scientific necessity of building new infrastructure

Radio astronomy research can use single dish or interferometric data. Single dish data comes from traditional dish telescopes that typically physically move and point at a location on the sky from the Earth. Interferometric data comes from telescopes that are composed of a scattered array of antennas that capture electrical signals across different Fourier modes. The WRT is composed of an array of over 100 antennas used for interferometric data collection. Some interferometric telescopes, such as the WRT, can be pointed

digitally by adding delays to the signals captured from subsets of antennas. Each type of data requires different mathematical processing to be usable for answering researchers' questions and contributes to the necessity of building new infrastructure.

There are a variety of pieces of software in the radio astronomy community for working with interferometric data. Some are generalized analysis packages, such as the Common Astronomy Software Applications (CASA) package being developed by an international consortium for next generation radio astronomy telescopes. Other pieces of data processing software are custom built for a particular telescope or project's processing approach. Developing a custom software package is necessary when a research group wishes to have more accuracy in the processing of their data since CASA and other packages are generalized tools and thus not built to work with the distinct data of a particular instrument.

Fundamental to processing interferometric data, such as that output by the WRT, is gridding each measurement to the UV mathematical plane so that the data is pixelized in a standard way. This enables faster computations during later processing since the data has been normalized in its distribution. This gridding and calibration takes place in Igor's Calibrator package. Key to this process of gridding is calculating the beam pattern of the antennas in the scattered array. This calculation and application of the beam's actual model is key for accurately processing data and one of many reasons their approach differs from that of some commonly available software packages.

Calibrator's gridding is designed to accurately account for the antenna shape of each measurement taken with the WRT. In contrast, Igor notes that common radio astronomy software uses, such as CASA, "*one just single, simple shape that's the same for all of them [the antenna beams]*." This shape might be a simple circle, whereas the true antenna beam would have dimples in its edges with stronger data capture in some areas when compared to others. Furthermore, some processing packages do not even attempt to grid using a beam pattern and grid their data, instead using a less rigorous and generalized shape for their beam.

Using a single simple shape as the antenna beam pattern fundamentally reduces the fidelity of the data and restricts the analysis work that will eventually be possible. This is permissible for narrow-field-of-view instruments that are being used to look at bright radio sources. However, by design the WRT is a wide-field-of-view instrument. Furthermore, EoR science is looking for incredibly faint sources. This work in fact subtracts out bright sources from the sky. Incredible accuracy of the gridding and accounting for the actual pattern of each antenna's beam is thus necessary to the scientific goals and a key motivator for the design of this infrastructure, especially its data products and software.

B. Three data products and their composition

The Radio group's EoR processing pipeline works with and produces many data products. Here we describe three such products. The first step is converting the WRT's raw data into a data product that can be processed by the pipeline. The remaining processing steps then produce a series of intermediate data products that are useful to different components of the pipeline in different ways. These outputs are intermediate data products because they have been processed in one or more ways, but typically require further processing to be usable for scientific analyses. Finally, plots that visualize the data are output from multiple components of the pipeline. An idealized flow of the data through each component of the Radio group's pipeline from raw to fully processed and ready for analysis is provided in Figure 2.

Each of these data products encapsulates multiple design decisions (i.e. antenna beam shape, software language choices, algorithm implementation, etc.) that impact the data and software that are produced. Furthermore, any of these products may be shared outside of the Radio group and their immediate collaboration. They will however be of varying use to other researchers due to the design choices they encapsulate, as some choices will not support certain types of analyses. We do however note here that the outputs of Calibrator and Power are going to be shared with the other EoR pipeline in development for processing WRT data, and vice versa. The sharing of these outputs is an intentional part of the international EoR group's design process to ensure that the processing and analysis for EoR science across the entire WRT collaboration are indeed arriving at truly novel scientific findings, rather than just exhibiting an artifact of one pipeline's design, by allowing the collaboration to test and compare each pipeline component's output individually.

1) Data product one: UVFITS files

The first data product that needs to be produced for the Radio group's work is output from an initial cleaning of the raw data from the telescope. EoR observations take place at the telescope and are archived as raw data at a nearby computing center. Once an EoR observation is complete the raw data is automatically transferred to a mirror copy at another institution here in the United States. The Radio group uses the mirrored data store as their "local" data source. Once the "raw" dataset is stored on this computing cluster the US EoR pipeline can be executed and process the raw data, producing intermediate data products along the way.

FlagAvg is pre-processing software produced by a member of the WRT collaboration that flags and removes bad data while averaging the good data to reduce its size. The Radio group uses FlagAvg to output UVFITS files for a specified observation. These UVFITS file are the first intermediate data product in the Radio group's pipeline. These files are necessary for not only the Radio group's continued processing work but may also be shared outside of the US EoR collaboration.

UVFITS is a popular file format within radio astronomy. However, Igor describes UVFITS as "*something that people hacked to make it kind of work,*" and Igor has noted in both of our interviews that the radio astronomy community does not use it as its creators intend it to be used. The US National

Radio Astronomy Observatory's documentation for their CASA software package even notes, "*The UVFITS format is not exactly a standard, but is a popular archive and transport format nonetheless*" [14].

When we asked Igor if there had been any push to formalize the standard further for radio astronomy he noted people have said they would like a standard but

> *in general the data volumes are so large that anytime you go to something standard there ends up being either constraints in what sort of data you're allowed to collect or a lot of wasted space.*

Rather than standardize the format further radio astronomy projects place their data in the format in a non-standardized but understood within the community manner. This is necessary to store the massive data output in as efficient manner as possible. For example, for many frequency calculations on the data it is possible to store one value and a set of relations to this value to recreate the original values that were obtained. Storing this simplification and then restoring the values during a processing step can thus reduce the volume of data stored. This design decision requires that Igor and the other members of the Radio group working with UVFITS files ensure that they understand precisely how data is being stored in this format if they are to take advantage of such efficiencies. If they do not understand how the data has been placed into this transport format then they may end up processing garbage.

2) Data product two: IDL Save files

The next set of intermediate data products are produced by Igor's Calibrator and Brianna's Power packages. These products are stored in IDL save files that contain a custom data structure. These files are saved after different processing steps are completed. Depending on the steps executed there may be anywhere from one to three different intermediate data products that are output and saved from Calibrator alone. IDL save files are shared between these two components of the pipeline and enable Brianna and Igor to collaborate on their work. They may be shared with members of the larger US EoR group or the International EoR collaboration as well. These files and the way they structure data internally are furthermore key to the Radio group's infrastructure maintaining stability as each component slowly evolves since they are the site of interaction and coordination between the various components of the infrastructure.

Multiple versions of the IDL save files are output and saved after multiple processing steps due to the substantial computational processing that must take place for some operations. Storing these intermediate data products then provides a starting point for later operations without requiring that previously completed operations be re-run at that time. The data products output by Calibrator that are of use for the Power component of the pipeline contain image, weights, and variance cubes that are appropriate for power spectrum creation. Those output by Power contain data that is further processed in the creation of power spectra.

The intermediate data products produced by Igor and Brianna contain data that is processed to different degrees since different copies are output after computationally intensive tasks. Igor's Calibrator package reads in the UVFITS data product output from FlagAvg. He refers to this as the data preparation work. The data that is read in will, after processing, be stored in the IDL save files that are the intermediate data product with which Igor and Brianna both work. Once the data from the UVFITS files is read in, along with other information, it is then placed in an internal data structure, "obs_structure." It is important, however, to note that Igor can output UVFITS files and Brianna FITS files if needed to share with persons outside of the EoR collaboration.

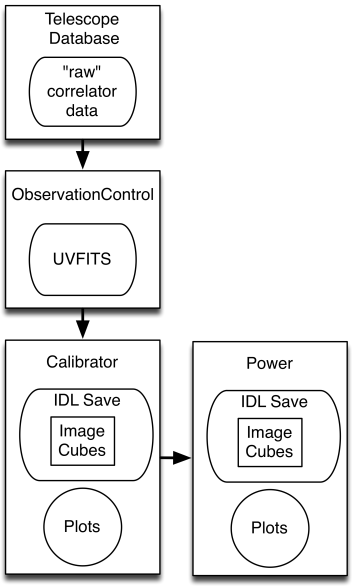

Figure 2. An overview of each component of the Radio group's data processing pipeline and the data products each produces. The arrows display the *idealized* flow of data from one component to another, however the intermediate data products do not always move through in such a tidy manner.

Obs_structure is a data structure Igor designed and wrote in IDL at an early stage of his development work. This IDL data structure stores information about the coordinates being used for processing this data, details about the telescope, and an astrometry structure that the data is placed in. Obs_structure's storage of these details about the data and the data's processing is central throughout the Calibrator and Power components of the pipeline.

The data products that are the "final" outputs of Calibrator and ready for the Power component contain image cubes that are suitable for further processing work. Power, the third pipeline component, takes in the image cubes that Calibrator stores in its IDL save files. Brianna describes taking in data outputs from Calibrator and requiring three "*basic inputs*" (Dirty, Model, and Residual image cubes) to be able to do her

work, although Calibrator has evolved in the last few months to provide five since it is creating them internally anyway. Each of these cubes is a different aspect of the processed intermediate data products in Calibrator. For example, the Dirty image cube is the telescope data before many processing steps have been applied while the Model image cube is their sky and beam model that is subtracted from the dirty image as one of the steps in Calibrator. A Residual image cube is the result of subtracting the Model from the Dirty.

Internally Brianna's Power stage relies upon Igor's obs_structure design since the image cube data products she takes in from the IDL save files are already stored within the data structure that has become a means of coordination within this infrastructure. Power processes the image cubes that are brought in from Calibrator in various ways. After certain processes are executed, intermediate data products are once again written out to IDL save files as a processing operation is completed. Writing out these intermediate files and saving them once again helps save on future computational effort since many of these files are reusable by the pipeline during future executions without requiring that a given computationally intensive task be re-run.

3) Data product three: Plots

Finally, Calibrator and Power also produce plots of the data they are processing. These plots visualize the data in various ways, from the field of view of the telescope produced by Calibrator to the various power spectra images created by Power. Plots are produced so that the Radio group can interpret the data they are collecting and processing. In addition, they serve as an important tool for debugging their pipeline since many processing issues become apparent from these visualizations.

Each of the data products being produced by the Radio group is a mechanism for the members of the group to collaborate and accomplish their ongoing software development tasks as they continue to process their data. UVFITS files serve as a transport mechanism between different radio astronomy projects, in spite of their not entirely standardized design. IDL Save files act as an intermediate data product that moves within the overall Radio group pipeline infrastructure, and across their international collaboration. Key to understanding the importance and role of each data product, and especially their ability to be shared with different stakeholders, is their design and creation in conjunction with the software that produces them, which we examine next.

C. Consequences of experiment and hardware design on data processing infrastructure production

Experiment and hardware design decisions have consequences for the design of the data processing infrastructure. Each of the data products introduced above incorporate and impact the design of the various software components the Radio group is developing, along with that of the WRT instrument itself and to some extent the larger radio astronomy community. Design decisions made for the telescope itself or in any given software component impact how data products are produced or used. Likewise, the data products themselves impact the software components that the group is producing. These decisions impact both the software components and the data products we have just introduced. We discuss some of these interconnected design decisions, with two examples below.

Our first example, on the design of a custom data structure, illustrates how when a researcher creates a central component of a data processing infrastructure they must account for the many design decisions of the larger experiment while simultaneously imparting their own decisions on the infrastructure and their collaborators. The design of the hardware platform influenced the design and use of Igor's obs_structure data structure within the Radio group. The processed data that is stored in an obs_structure in either Igor's Calibrator data products or Brianna's Power data products is one of the reasons this entire processing pipeline is differentiated from other software packages such as CASA. The data stored within this common data structure is calibrated and gridded by accurately accounting for the portion of the sky they are examining with the telescope and the many intricacies of the hardware platform such as the characteristics of the specific digital radio receivers or physical cables in use. The intricacies of the hardware platform and original experiment design have a profound impact on what data products can and need to be generated and saved, and what software must be developed to produce, clean, and process them.

Igor's creation of this structure and its adoption by Brianna enables coordination between their respective software components and the data products. If Brianna were to use a different data structure in her component then she would have had to do more work to use the intermediate data products Igor is producing. Calibrator's processing steps must not only continue to account for specific details of the telescope instrument's instruments but must also store the processed data in the obs_structure. For example, when processing data Igor must setup a beam model of the actual beam from the telescope as the collaboration as a whole understands it to be shaped. This model is referred to as the model image. The model image is subtracted off of the initially gridded, dirty image so that they are left with a residual image. The residual image is what the Radio group requires in the long-term to be able to examine the EoR within their telescope's field of view.

The dirty, model, and residual images capture different design quirks of the telescope as an instrument and are used by Brianna's Power component in her power spectra production and debugging work. Over time the Radio group has determined that having all three image cubes output from Calibrator is beneficial since it helps debugging across the entire research infrastructure. Previous versions of Calibrator only output one of the three images since that was all Igor decided to put out of his component at the time. Having all three image cubes enables the group to more rapidly assess whether an unexpected variation in the residual image is real or an artifact of some piece of the pipeline's processing. Thus

the demands of the larger research infrastructure and the vagaries of the hardware platform, inspired changes to the Calibrator pipeline.

Our second and last example, about a "fourth line" bug, shows how without knowing the particulars of hardware such as the speed of light through particular cables, astronomers cannot properly process data. Brianna's Power component produces many plots of the processed data. At this time the plots are primarily useful as a debugging tool, but long-term they are crucial for analysis work. During Winter 2014 one error that arose in the plots of the dirty, model, and residual images was that of the "fourth line" bug.

In order to detect faint EoR phenomenon the Radio group must work with petabytes of data so that they can reduce the noise inherent in any data collection as low as possible. They are developing their data processing pipeline while the necessary quantities of data are being captured with the telescope. While developing their pipeline the group has tested their system with increasingly large quantities of data, going from initial testing with 2-minute observations to most recently spending months working with 3-hour observations. This process is crucial to their ability to find and fix bugs while scaling the processing capability of their products. However, as the quantity of data increases, separating noise from the interesting data becomes increasingly challenging. Within these image plots the Radio group expected to see a structure of three lines but then a fourth line suddenly began to appear in Brianna's plots. This led to significant debugging work to implement a small yet crucial fix.

The appearance of this unexpected fourth line in the plots led to debugging conversations and tests between first Brianna and Igor, then the entire Radio group, then up to the entire US EoR group, and finally to the point that it was discussed with the entire International EoR project group. These individuals and groups over a period of weeks had to trace through the design of Power and Calibrator all the way back to the physical cabling of the WRT instrument itself to track down what was determined to be the source of the fourth line in their 3-hour observation plots, at least for the time being until larger quantities of data are processed when it may reappear. This debugging process relied upon Brianna, Igor, and the rest of the Radio group examining how beam modeling is done, data stored in obs_structure, how the raw data is converted with FlagAvg, and crucially the design of the telescope itself.

To fix the fourth line bug (at least for the time being) the group eventually determined that Igor's Calibrator component was not accounting for the reflection of the analog signal in these cables in some stages of its processing work. Accounting for the length and speed of light in the cables was known to be necessary up front in the Radio group's work due to the design of the WRT experiment and its hardware infrastructure. However, through their multi-week debugging process the group realized that Calibrator needed to also factor in the reflection of the signal in these cables in its processing steps. Igor then had to modify Calibrator to properly account for this design artifact of the telescope and the metadata regarding the cable used that is associated with the raw data. From here the group could output new versions of Calibrator's intermediate data products so that Brianna could once again test Power and output plots to examine whether the solution worked, at least for a 3-hour observation.

The design decisions and quirks that Power, Calibrator, and every other component of the pipeline account for in their code and data products illustrate the close connections between each product within and across the stages. The hardware and "raw" data products of the WRT embodies the research design and the initial system configuration decisions of this international consortium of radio astronomers. This research design and implementation in turn always impacts any of the data gathered with this instrument. Without a sufficiently detailed understanding of the design of the telescope, the "raw" data it outputs, the computing systems, or the software and data products of every other stage no one individual component could be produced. The Radio group's project infrastructure and datasets would therefore not emerge as a cohering and stabilizing system with which their scientific goals can be addressed.

Long-term the WRT collaboration is required to share its data products with the astronomy community at large, as is common for most any large, international scientific collaboration in the 21st century. However, whether the collaboration must share the "raw" or some form of processed data products has yet to be determined. For researchers outside of the WRT collaboration to be able to scientifically use any of the project's data some form of the processed data products would have to be shared, and potentially some of the associated infrastructure, since the infrastructure and processed data products have taken into account the myriad design decisions and quirks of the experiment. Any external work using such products is therefore dependent upon the design decisions captured in the stabilized system while also imparting further such decisions on its products.

V. Discussion and Conclusion

The design decisions producing data and software products in scientific research are fundamentally intertwined since they are developed simultaneously with decisions made for one product directly impacting the other. CSCW and eScience scholars do not typically explicitly claim that these products are fully separable, however, we also rarely discuss them as intertwined. From our examples of the Radio group we see just how deeply connected the design decisions made during the development of the telescope as an instrument, its raw data products, and the group's computing system are to their efforts to produce a processing pipeline. In turn, the group's design decisions, such as Igor's data structure, are imparted on the data and software that they produce and share with collaborators, and, potentially long-term, their larger community.

The context of the production of all of these components (i.e. the particular hardware in use, software language choices,

implementation of a mathematical algorithm, etc.) along with the scientific theory is often crucial knowledge for the enactment of research infrastructure and the conduct of the science itself. This is especially true as the volume of data being relied upon increases in size and depth and the design decisions expand in scope. Our informants all noted the necessity of producing this infrastructure from scratch because existing solutions no longer hold up under this volume of data or account for its depth. Many existing software packages simply crashed under the massive datasets. Others did not come close to processing data accurately enough for the Radio group and larger WRT collaboration's needs.

The necessity of developing a new data processing infrastructure is a fact of life for researchers engaged in cutting edge experimental work. These researchers must spend a substantial amount of their time doing work that is systems development. They acknowledge that this work is necessary and a fact of their day-to-day lives, yet still do not entirely consider such work to truly be the "scientific" work in and of itself. However, our work clearly illustrates that iterative system development is key for answering their scientific questions and the development of the necessary research infrastructures. Such "non-scientific" work has a deep and lasting impact on the group's scientific outcomes since the design decisions throughout the whole process accumulate over time. The co-production of software and data products in the course of building a data processing infrastructure is intricately connected to early design choices for the overall research. This leads to an interesting interaction between the telescope as an instrument (and infrastructure in its own right) and the new data processing infrastructure being created by the Radio group.

The human infrastructure of the Radio group, and the larger WRT collaboration, dynamically knits together a new research infrastructure through the creation of data products and software pipelines. The initial research and experiment design is constraining the boundaries of the scientist's work and in turn the scientists are configuring this new data processing infrastructure. The researchers' design decisions unfold in the context of an evolving and increasingly detailed system whose complexity must be understood and accounted for to create viable data products that are of scientific value, and potentially shareable.

To better support scientific research, we need to continue to develop our understanding of the practices that enable and produce data products, software, and ultimately infrastructures. To further advance data-intensive projects we must understand the day-to-day work practices that are enabling new sociotechnical research infrastructures to emerge. Some elements of such systems will need to be customized to account for the situation at hand (i.e. the other software, the data products, and the science knowledge that are of relevance). Understanding and supporting the dynamics of such sociotechnical change will help us to create sustainable infrastructures that can adapt and change over the long-term. While the specific products of a given research project may not be sustained, we may yet learn from and iterate upon the design practices that produce these research infrastructures.

VI. Acknowledgments

This material is based upon work supported by the National Science Foundation grants IIS-0954088 and ACI-1302272. We wish to thank our informants along with Betsy Rolland and our anonymous reviewers for their feedback.

VII. References

[1] Belhajjame, K., Goble, C., Soiland-Reyes, S., & De Roure, D. Fostering Scientific Workflow Preservation through Discovery of Substitute Services. In Proc. IEEE E-Science Conference 2011. (2011), 97-104.

[2] Bietz, M. J., & Lee, C. P. Collaboration in metagenomics: Sequence databases and the organization of scientific work. ECSCW 2009 (2009), 243-262.

[3] Bietz, M. J., Paine, D., & Lee, C. P. (2013). The work of developing cyberinfrastructure middleware projects. In Proc. ACM CSCW Conference 2013. (pp. 1527-1538). San Antonio, Texas, USA.

[4] Darby, R., Lambert, S., Matthews, B., Wilson, M., Gitmans, K., et al. Enabling scientific data sharing and re-use. In Proc. IEEE E-Science 2012. (2012), 1-8.

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

[6] Grochow, K., Stoermer, M., Fogarty, J., Lee, C., & Howe, B. COVE: A Visual Environment for Multidisciplinary Ocean Science Collaboration. In Proc. eScience 2010, IEEE (2010), 269-276.

[7] Harper, R. P. Inside the IMF: An ethnography of documents, technology, and organizational action. Academic Press, Inc., 1997.

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

[9] Jirotka, M., Procter, R. O. B., Hartswood, M., Slack, R., Simpson, A., et al. Collaboration and Trust in Healthcare Innovation: The eDiaMoND Case Study. Computer Supported Cooperative Work (CSCW), 14, 4 (2005), 369-398. DOI= http://dx.doi.org/10.1007/s10606-005-9001-0.

[10] Killeen, N. E. B., Lohrey, J. M., Farrell, M., Liu, W., Garic, S., et al. Integration of modern data management practice with scientific workflows. In Proc. E-Science (e-Science), 2012 IEEE 8th International Conference on(2012), 1-8.

[11] Lawrence, K. A. Walking the Tightrope: The Balancing Acts of a Large e-Research Project. Computer Supported Cooperative Work (CSCW), 15 (2006), 385-411. DOI= http://dx.doi.org/10.1007/s10606-006-9025-0.

[12] Lee, C. P., Dourish, P., & Mark, G. The Human Infrastructure of Cyberinfrastructure. In Proc. ACM CSCW Conference. (2006), 483-492.

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

[14] National Radio Astronomy Observatory. (2010). 2.2.4 UVFITS Import and Export. Associated Universities, Inc. http://casa.nrao.edu/docs/userman/UserMansu96.html - x110-1090002.2.4 [2014, May 10].

[15] Ribes, D., & Finholt, T. Tensions across the scales: Planning infrastructure for the long term. In Proc. ACM Conference on Supporting Group Work, GROUP (2007), 229-238.

[16] Rolland, B., & Lee, C. P. (2013). Beyond trust and reliability: reusing data in collaborative cancer epidemiology research. In Proc. ACM CSCW Conference (pp. 435-444). San Antonio, Texas, USA: ACM.

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