Author pre-print in advance of publication in the ACM PACM-HCI

Data Integration as Coordination: The Articulation of Data Work in an Ocean Science Collaboration

ANDREW B. NEANG, University of Washington, USA
WILL SUTHERLAND, University of Washington, USA
MICHAEL W. BEACH, University of Washington, USA
CHARLOTTE P. LEE, University of Washington, USA

Recent CSCW research on the collaborative design and development of research infrastructures for the natural sciences has increasingly focused on the challenges of open data sharing. This qualitative study describes and analyzes how multidisciplinary, geographically distributed ocean scientists are integrating highly diverse data as part of an effort to develop a new research infrastructure to advance science. This paper identifies different kinds of coordination that are necessary to align processes of data collection, production, and analysis. Some of the hard work to integrate data is undertaken before data integration can even become a technical problem. After data integration becomes a technical problem, social and organizational means continue to be critical for resolving differences in assumptions, methods, practices, and priorities. This work calls attention to the diversity of coordinative, social, and organizational practices and concerns that are needed to integrate data and also how, in highly innovative work, the process of integrating data also helps to define scientific problem spaces themselves.

CCS Concepts: • **Human-centered computing → Computer supported cooperative work**; **Ethnographic studies**; **Empirical studies in collaborative and social computing**.

Additional Key Words and Phrases: Data Sharing; Data Integration; Infrastructure; Articulation Work; Data-Centric Science

ACM Reference Format:
Andrew B. Neang, Will Sutherland, Michael W. Beach, and Charlotte P. Lee. 2020. Data Integration as Coordination: The Articulation of Data Work in an Ocean Science Collaboration. *Proc. ACM Hum.-Comput. Interact.* 4, 0, Article 000 (2020), 26 pages. https://doi.org/000000000000

1 INTRODUCTION

In most characterizations of the emerging "data-centric" or "data-driven" model of science, one of the core promises is the ability to mobilize data from one situation on to new and unexpected uses in other situations. As Leonelli [35] has argued, the core characteristic of the data-centric project is to broaden the evidential value of data in this way, making it relevant and permissible as evidence in more and more diverse research situations. Given this focus on the movement and reuse of data, the issues of data sharing, data integration, and data interoperability have become central problematics in the rollout of a new data-centric model of science. Substantial research on data integration

Authors' addresses: Andrew B. Neang, University of Washington, Seattle, USA; Will Sutherland, University of Washington, Seattle, USA; Michael W. Beach, University of Washington, Seattle, USA; Charlotte P. Lee, University of Washington, Seattle, USA.

2573-0142/2020/0-ART000 $15.00
https://doi.org/000000000000

and sharing, however, has found it to be something of a logistical nightmare, and has focused on outlining the factors or barriers to data sharing, such as documentation, access, collaboration politics, standardization efforts, disciplinary differences, and concerns around scooping or misuse that complicate the process of data exchange [4, 10]. And yet in various locales, and in however limited scope, researchers do share data, and effective processes for documenting, exchanging, and reusing data do emerge between groups or collaborations. This study can be seen as a return to the drawing board, or rather the field, to examine how these practices for integrating data get hashed out, established, and routinized. In doing this we want to shift away from questions of when or whether data will be shared and towards questions of how researchers go about exchanging and reusing data.

There is a plethora of terms referring to the social and technical approaches and tools that seek to make heterogeneous data work with each other including data interoperability, harmonization, standardization, federation, mapping, and integration [51]. In eScience, Computer Supported Cooperative Work (CSCW), and Human-Computer Interaction (HCI) more generally, the problematic used to frame studies of these topics has focused on delineating and typologizing factors, barriers, and incentives which contribute to a person's decision to share data. Researchers withhold data due to lack of trust, fear of being scooped, lack of knowledge that their data might be useful, or because of mismatches in formats and standards [4, 56]. At the same time, researchers might be encouraged to share data through data citation, easing the work of contributing data sets, institutional requirements, or secondary benefits [16, 48]. The literature has begun to investigate data sharing as a practice, engaging the ongoing activities of researchers [46]. These studies have brought valuable contributions in the form of ethnographic description and survey analysis, and has turned the attention of scholars towards a diverse array of actors and technologies behind and beyond the datum itself. However, where the literature does engage these practices explicitly, it uses practice as a unit of analysis or explanation, and largely restricts itself to identifying such practices and factors which contribute to when or whether they will occur, or how they can be incentivized [e.g., 65].

Our goal in this paper is to move from the delineation of factors or barriers towards the particular kinds of work that researchers do in developing practices for overcoming barriers and bringing their concerns into alignment. In doing this we frame data integration as a kind of coordinative work rather than an exchange. We draw on the concept of articulation work [61], and many of its refinements in the study of large-scale collaborative cyberinfrastructures, in order to examine how these practices emerge and are negotiated between researchers. In particular we draw on lessons learned in the organizational practices of researchers in large scale cyberinfrastructure projects [2, 26, 50], particularly regarding the work of leveraging and aligning relationships in order to bring new collaborative possibilities into existence [1]. While our goal is to bring larger processes of infrastructuring into consideration, we do so to contextualize data integration, a specific aspect of larger collaborative systems that has become a focal point of research in the area. It is not our goal to refine the concepts of articulation work or synergizing themselves, but rather to use them empirically to examine, in detail, the specific task of data integration, and to surface some productive lines of inquiry for research on that topic going forward. We use the term *data integration* primarily because it focuses our attention on those situations where data is actually reused rather than just shared or published, and secondly because it captures the project of making data from different sources comparable [42].

We examine this topic through an empirical investigation of a multidisciplinary ocean science collaboration called CBIOMES. The purpose of CBIOMES is to create the Collaborative Maine Atlas Project (CMAP), a system that includes not only a database but also additional services to integrate, visualize, and analyze ocean data sets such as satellite data, in-situ observations, and model outputs. The field of Ocean Science, as with many other scientific disciplines, is in flux.

Ocean science researchers study the global marine environment and its interactions with the earth, the atmosphere and the biosphere. Scientists increasingly wish to use multidisciplinary research methods to answer complex environmental problems. This sharing of data necessitates learning about very different methods, processes, and types of data. The CMAP is still in its early stages, and the collaboration processes of data exchange and mutual engagement between researchers from different fields and methodologies are still being worked out. We follow these efforts to make data relevant to new research questions as a way of foregrounding the work of making data integration go well. With this in mind we frame our investigation around a specific question: *How do multidisciplinary ocean scientists coordinate with each other in order to integrate data?*

Our findings highlight the variety of kinds of work that researchers have to engage in to make data integration between disparate research sites go well. This means not only taking on extra work in the form of preparing or documenting data, but in establishing relations and coordinating with others to figure out how the movement of data might be useful, the details of how it will be accomplished and who will do it, and what the proper documentation is in the first place. In this examination the lens of articulation work focuses our attention on the kinds of tasks that go on before and behind the work of data integration, and we follow a number of ramifications of these observations for the way we might approach the problem of data integration in future investigations and interventions. Specifically, we point out the highly contingent situation of data integration efforts on the part of individuals, the essentially novel character of the work arrangements these researchers are attempting to establish, and the pragmatic benefits of focusing on practice over barriers and incentives given this high level of contingency. We further draw a connection between these forms of articulation work, and the ways that new tasks and types of work emerge in a data-oriented scientific initiative.

This paper proceeds by outlining prior work on various conceptions of data reuse in science and on the concept of articulation work, which informs our analysis. We then describe the field site, the CBIOMES collaboration, and our methods of investigation. Our findings section takes the format of a set of scenarios which highlight different situations of collaboration around the problem of data integration. Finally, the discussion outlines the points made above.

2 LITERATURE REVIEW

Efforts to encourage large-scale, distributed scientific collaborations that enable many new scientific discoveries, such as cyberinfrastructure and eScience initiatives, have drawn a great deal of attention from researchers [25]. Early efforts to support large-scale, distributed scientific collaboration tended to be technology-centric, often focusing on the provision of advanced computational capabilities such as high-speed computation, computer architectures that leverage cloud resources, data-repositories, and specialized analytical tools [17, 23, 41, 72]. System developers and researchers alike, however, quickly began to recognize the challenges of scientific collaboration as being fundamentally sociotechnical, which is to say that technologies, social practices, and social structures are inextricably enmeshed and interwoven.

2.1 Data Sharing, Reuse, and Integration

Understanding how to better support the sharing of scientific data within and across disciplines has been an important area of interest to scholars in CSCW [66]. The need for data to be made more openly accessible has been widely recognized, and there is a growing body of research on data sharing practices in various scientific fields (e.g., ecology, biology, and astronomy) [68, 70]. Oleksik et al. [39] noted that improvements to collaborative data generation and reuse hinges on “a deeper understanding of the social and technical circumstances”. Empirical works on data sharing and reuse reveal how lacking contextual information on where and how data are produced (or acquired) [13, 67] can lead to difficulties with interpreting how key variables were constructed [52] and concerns about the data's quality and usefulness [74]. Oftentimes, reaching out to the original data producer is the only way to acquire the contextual information necessary to assess the data's relevance for reuse [52]. Standardization has played a pivotal role in helping address the challenges brought on by heterogeneity (e.g., different data formats, data types, and scientific disciplines) [6]. Standards have been designed and employed to better facilitate data sharing and data reuse among and across different disciplines [27, 28]. For example, Millerand et al. [38] describes how ecological scientists engage in collaborative efforts to identify and reach mutual agreement on terminologies and naming conventions, provide tutorials and training on integrating metadata standards, and translate the standards so that they can be circulated across multiple sites.

Data integration depends on data being shared in a format that is both accessible and understandable. The need for more systematic methods for data integration, whether to make comparisons, develop new computational models, or explore larger, more complex questions has led to increased exploratory research in this area [73]. Pasquetto et al. [46] assert that data reuse is not to be limited to reproducing research, which is an example of independent reuse (reuse of individual datasets) but should also include data integration. Integrating diverse data is necessary but challenging for many scientific disciplines. Studies reveal how researchers across the environmental sciences struggle to bring together data collected at different scales and using different nomenclatures [9, 36, 38], how archaeologists must attempt to combine data collected in different contexts with different methodologies [14], and how neuroscientists wrestle with the need to merge neuroimages with clinical data (e.g. demographic, genetic, and behavioral) [71].

Data interoperability, in particular, is a critical area for the success of new research collaborations and has drawn a great deal of attention in recent years [11]. Yet researchers often take data interoperation for granted, viewing it as an established reality, rather than as an object of inquiry. Ribes stresses that "...interoperability is a fundamentally historical phenomenon. More precisely, data interoperation is a form of front-loaded practical work, negotiation and technical innovation that is thereafter black-boxed, largely forgotten, eventually taken for granted and naturalized as the inevitable technological trajectory for data. Only following interoperation do data flow with the ease promised by its advocates while evoking concerns about unanticipated or unintended uses" [51]. Rather than looking at completed data interoperability, we investigate the negotiation that happens before desired data interoperability is achieved.

2.2 Articulation Work

Articulation work is often characterized as consisting of a set of activities (e.g., planning, coordinating, negotiating) required to bring together discontinuous elements into working configurations to accomplish an overarching project [55, 62, 63]. Scholars in the fields of CSCW and HCI have taken up and extended the concept in different research contexts, often making use of it to highlight the less visible work that allows other, more visible work to happen [45, 60, 64]. Articulation work can be better understood when contrasted to primary work. While primary work typically centers around the "specific agendas and goals of the work situation" [22], articulation work focuses on the "specifics of putting together tasks, task sequences, and task clusters" [61] to keep or get things on track in the face of unforeseen contingencies, disruptions, and breakdowns. Pallesen and Jacobsen [43] noted how work projects in particular "entails a division of labor and thus requires that actors continuously engage in linking or meshing otherwise divided tasks". Articulation work then engages how tasks come about, how people delineate specific units of work and their attendant roles: "Since the plurality of tasks making up their totality, as well as the relations of actors to tasks, are not automatically articulated, actors must do that too, and often in complex ways" [61].

For the last two decades, scholars in CSCW have turned to the concept of articulation work as a lens for studying the nature of cooperative work in order to establish a foundation for designing information systems [54]. They have used the concept to broaden the notion of work to include routine and overlooked forms of labor over the years [37, 49] and in the process have argued how articulation work is continuously unfolding as people, practices, and tools are constantly changing. It is a cumulative type of development; that is, as work becomes more complex (e.g., adding new work tasks, using new tools, etc.), there is more articulation effort required [53].

In order to better disambiguate specific data integration efforts we turn to theoretical concepts specifically pertaining to different types of coordination: local articulation work, metawork and synergizing. Local articulation, or simply articulation work, consists of specific tasks assigned to individuals that "insure the flow of resources," "make arrangements about the division of labor," "match workers' motivations and tasks," and "supervise delegated or assigned responsibilities" [62]. Articulation work is a useful concept because it enables researchers to characterize how actors, actions, and the combination of both are brought into alignment such that project work can be completed successfully. Articulation work has been defined as "the work of making work go well" [62]. It is a collection of "individual and yet interdependent activities [that] must be coordinated, scheduled, meshed, integrated" [55].

Advancing the theory of articulation work, Gerson further refines the notion into the two concepts of metawork and local articulation work. Metawork is about ensuring that different types of activities function together as expected, using pre-defined specifications and representations, to align different units of work [21]. Local articulation refers to what needs to be done to ensure that all the resources are available and operational for activities to be undertaken "in the local situation" [21]. There is a certain overlap between these two facets, and in Gerson's view, they are particularly useful constructs to investigate reach, reach being "the distribution of tasks across organizational, spatial, and temporal boundaries" [21]. In this discussion and others the notions of local articulation work and metawork are carried out within a specific field of work: the objects, conceptual models, and knowledge representations through which a group coordinates and changes state in the process of ongoing interactions. [1] turn attention towards how this state of coordination and mutual objects of work comes about, and use the word synergizing to capture the kind of work involved in bringing a field of work into existence in the first place. There is a recursive relationship between these concepts. The work of metawork, local articulation, and synergizing continue throughout the course of the project. This type of work keeps the project from stalling or halting and ensures the project can continue to progress [1]. The notion of "distance" we discuss below is not the more literal sense of distance discussed in Olson and Olson's investigation of collocated and remote collaboration [40], but we share a concern with how people establish a common ground. We explore a similar issue here as a problem of establishing a common field of work.

Extending Strauss work on articulation in medical worlds to laboratory work, Fujimura [18, 19] introduces the concept of "do-able" problems. In her study of cancer research laboratories, Fujimura demonstrates how constructing do-able problems or successful research projects depends on being able to align tasks among three different levels of work organization: the experiment, the laboratory, and the social world. A research problem is more or less do-able contingent on how difficult it is to carry out the articulation work (e.g., planning, evaluating, coordinating, integrating) needed between those levels of work organization to create alignment [18]. The term 'alignment' is often used to refer to one aspect of articulation work and processes necessary to fit tasks, actors, and projects together [24, 59, 62].

3 RESEARCH SITE AND METHODS

The findings we present here are from an ethnographically informed field study on scientific collaboration and socio-scientific concerns that frame and constitute the integration and sharing of data in oceanography. This study informs a larger funded research project by the Simons Foundation which seeks to create a research infrastructure that will support data integration and facilitate data sharing within the foundation's collaborative project on computational biogeochemical modeling of marine ecosystems. Here we introduce oceanography as a research site along with the scientists collaborating to make a collaborative marine atlas, before discussing our research and analysis methods.

3.1 Oceanography as an Evolving Discipline

Advances in instrumentation, communications, and computational capabilities have ushered in a new era of experimental approaches to science. This change is particularly prominent in the field of oceanography where increasingly sophisticated computers, autonomous sensor systems, and innovative methodologies have rapidly brought about new possibilities for exploration and study — from the smallest marine microbes to the vastest ocean basins [29, 58]. What distinguishes the current state of this discipline from the past is the rapidity and volume of data generated by new data collection methods and the increasing focus on research questions that demand multidisciplinary problem-solving and collaboration. This requires bringing together data from satellites, laboratory cultures, or generated by models as well as data collected by a variety of instruments deployed from seagoing vessels on research "cruises." The oceanographic laboratory has become an evolving assemblage of novel instruments and interdisciplinary experts [31]. As research groups grow larger and require more formalized connections, so too does collaborative science. Because of the range of scientific knowledge required to understand all of the processes involved, ocean science is typically divided into three main disciplines: biological oceanography (marine biology), chemical oceanography (marine chemistry), and physical oceanography. There is a great deal of overlap between disciplines, because many aspects of the marine environment are influenced by interacting biological, chemical, and physical processes.

3.2 Scientists Collaborating to Create a Collaborative Marine Atlas

Our study focuses on members of laboratories that are stakeholders and part of this larger computational biogeochemical modeling of marine ecosystem project (referred to as CBIOMES) funded by the Simons Foundation. The CBIOMES project is comprised of ocean science researchers from 22 academic laboratories in the United States, Canada, and the United Kingdom who are also involved with the foundation's other two interrelated collaborative ocean initiatives in the area of microbial oceanography (known as Ocean Processes and Ecology (SCOPE) and SCOPE-Gradients).

CBIOMES is a relatively new project, having only been funded for approximately two months before we began our fieldwork in Fall of 2017. The project seeks to integrate key new datasets in real time as they are collected at sea to facilitate direct tests of theoretical predictions; to synthesize an atlas of marine microbial biogeography suitable for testing a range of specific ecological theories and quantifying the skill of numerical simulations; and develop new models. As part of the project, researchers aim to build a system – a CMAP – tailored for mapping marine microbial biogeochemistry in a statistically robust manner. This complicated undertaking requires the compiling, annotating, and managing of data sets of widely different qualities, formats, and sizes. Researchers taking the lead on CMAP are currently working in collaboration with the authors of this paper to design and develop software tools and computer architectures for a system capable of storing, integrating, exploring, and visualizing oceanic data sets.

3.3 Methods

For this qualitative, interview-based study, we conducted semi-structured interviews with 43 Simons Foundation-supported ocean science researchers based in 22 academic research laboratories across the United States, Canada, and the United Kingdom. These individuals were stakeholders and part of the foundation's larger CBIOMES project in the spring of 2018. Participants included principal investigators (PI), research scientists, postdoctoral researchers, graduate students, and operational staff members. Table 1 shows a breakdown of interviewees organized by role at the time of our study.

Table 1. Participants Interviewed – Researchers

	Junior Researchers	Senior Researchers
Observationalists	22	13
Modelers	3	5

We recruited a varied set of participants from across the foundation's collaborative ocean initiatives to take part in our study. Participant selection began with using a purposive sampling technique [44] based upon their discipline and role within the ocean initiatives and was followed by a snowball sampling approach [47] wherein we asked interviewees who else we should talk to. These interviews were designed to elicit an understanding about each researcher's background and membership in the Simons Foundation, including the funded collaborative initiatives they worked on, who they worked with, where they obtained data, and how data sharing took place in their respective project groups. The interviews were audio recorded and transcribed and lasted an average of one hour. In order to effectively triangulate our data [20] we supplemented our interviews with over 20 participant observations of group meetings, presentations, and demos by researchers working on the CMAP.

3.4 Data Analysis

Throughout data collection we employed a qualitative data analysis process [5] for which we iteratively open and axially coded the data that were collected in ATLAS.ti to develop and refine themes [12, 69]. The initial rounds of coding helped characterize each individual's scientific research work, how they produced (or collected) data and what challenges they faced with integrating data from other sources, and they revealed different perspectives as to when data should be accessible to others to further explore. Over time we narrowed down our coding to tease out the relationship between data sharing and integration and the risks, rewards, and additional considerations that both frame and constitute this work. Descriptive and analytic memos were written during ongoing analysis of the data. Coding and memoing were iterative: after writing memos, we would return to the coded transcripts to further enhance and refine the coding scheme. Codes in our final scheme covered themes ranging from collaboration (how participants' worked with different individuals), to how participants made their data available to others, and the different social and scientific concerns around sharing data. The research team then identified coding categories relevant to key topics such as distance and the different forms (i.e., familiarity, discipline, and make) participants perceived and felt from other individuals. Through this process, we developed a set of themes that we discuss in our findings.

4 FINDINGS

The CBIOMES collaboration reflected some of the classic difficulties of large collaborative science projects, and demonstrated many of the typical barriers to data sharing and integration. However, researchers also actively negotiated these barriers in order to establish data sharing and data integration efforts. We first set out a problem space in which these efforts took place by defining the concerns and contributing factors that were most salient in researchers' efforts. Next we shift our focus to the specific types of work that researchers engaged in to overcome or capitalize on situated social relations in order to accomplish data integration.

4.1 Unpacking Scientists' Umbrella Term of "Distance"

The ocean scientists in our investigation repeatedly used the notion of "distance" when describing the difficulties of developing new collaborations within CBIOMES. Our analysis disambiguates different notions of distance when collaborating to integrate data to further science. We define three types of distance: discipline, make, and familiarity. Brief descriptions of these "distances" provide background for understanding the concerns and challenges that make sense making conversations and the development of new work practices necessary. The notion of distance is an intentionally vague and highly aggregated term, allowing researchers to reference a variety of difficulties encountered in the course of trying to make use of others' data:

> "...I often rely on and work with researchers with other expertise in other fields such as those from marine or molecular biology. This enables us to take a 'big picture' approach to doing research. However, it's really not as simple as it sounds. The conceptual, theoretical, and methodological distances between our scientific domains is quite wide and often presents challenges when working together" (Bethany, Ph.D. Student)

Bethany elaborated on the challenge of interdisciplinary collaboration: "The challenges around managing these conceptual and methodological distances are really different on a case by case basis." Each collaboration between a certain individual and others scientists is its own case where specific conceptual and methodological assumptions and practices must be reconciled.

Discipline. Discipline has been a common distinguisher (and barrier) in accounts of data sharing [3, 15], particularly when expressed in terms such as "interdisciplinary" or "transdisciplinary". When we use the term *discipline* in this study, we refer to distinctions that emerged in the CBIOMES community delineating boundaries between biological, chemical, and physical oceanography. These three categories implied different background knowledge and training, as well as different primary research objects (e.g., cycling of dissolved nutrients and gases vs. growth or abundance of bacteria or phytoplankton). However, labs from these different disciplines often overlapped in the study of common ecological systems. Cell growth and metabolism, for instance, requires both chemical and biological data and analysis. The premise of the CBIOMES collaboration was to engage research questions that crossed these disciplinary boundaries in order to address research goals that were understood to be previously unattainable through the approaches of a single discipline. Researchers were therefore often operating under the assumption that it would be beneficial to integrate data from one discipline with another, and a number of interstitial categories emerged between these disciplinary approaches such as "biochemical" or "biogeochemical". In this sense, the distinct disciplinary divisions were in actuality somewhat murky within the CBIOMES collaboration, as the overarching goal of the project was to leverage data and methods from different disciplines together. Connected with the notion of discipline, there were also a large number of distinctions between different kinds of methods, such as metatranscriptomics and flow cytometry, which were associated with biological oceanographic approaches, and metabolomics and liquid chromatography-mass spectrometry (LCMS), which were typically associated with chemical oceanographic approaches.

Make. Researchers also made a broader categorical distinction between what are called "observationalists" and "modelers". These categories represent a different kind of methodological divide: Observationalists approach research questions through the collection and analysis of specific datasets. The work of observationalists in the collaboration included everything from data collection on research cruises, data cleaning, searching out other datasets on repositories or data portals, and developing instrumentation or analysis procedures. Modelers, on the other hand, approach research questions primarily through the aggregation of existing datasets, which included finding data sources, doing further cleaning, and developing novel ways of integrating and analyzing different data sources. We separate this category from the notion of methods for two reasons. Firstly, the observationalist/modeler distinction is of special significance to the researchers themselves when it comes to differentiating the work of different research groups. Secondly, the notion of make did not associate strongly with specific disciplines or other methods of data collection and cleaning, with biological or physical modelers making use of a variety of data types. Thirdly, within the developing work arrangements of the CBIOMES collaboration, observationalists and modelers begin to occupy different lines of work which are distinguished and related to each other as work processes, beyond their methodological associations. CBIOMES members never gave an explicit hypernym or umbrella term for the difference between modelers and observationalists (such as "discipline" for chemical and biological oceanography), but the CMAP system used the term "make" to indicate the categories of model data and observational data. We adopt that term here to capture a set of concerns centered around seeking out, acquiring, and integrating existing data sets with mathematical models in contrast with the work of developing measures, collecting data, and cleaning data from the point of collection with a sea float or lab culture. These different ways of collecting data and mobilizing it as evidence create an aspect of distance that is separate from distinctions between disciplines.

Familiarity. Researchers who had encountered each other on collaboration teleconferences, interacted directly through email or individual teleconference, collected data together on research cruises or previously exchanged data in the past found it easier to integrate data. Part of this had to do with familiarity with research methods. On CBIOMES teleconference calls, which involved a number of laboratories, researchers would regularly present some of the research going on in their lab. Through this and other means, one researcher could develop a sense of another's research, in terms of what kind of data they were collecting, what research questions they were hoping to address, and what methods they tended to use. Knowing these aspects of another researcher's work would enable them to better prepare datasets for integration both in terms of knowing what kind of data would be useful to the other researcher, and in terms of providing the right kind of metadata and explanation of the processing that the data had undergone. Familiarity was also built directly through PIs or collaborators putting one researcher in contact with another. In Scenario 4 below, for instance, a senior PI puts a modeler in contact with one of her former students who she knows can generate the kind of experimental data that the modeler needs. Researchers would also go on cruises together, sometimes for a week or more on the same vessel, and worked alongside each other to collect data. These cruises, as well as the planning meetings that lead up to them, were an important site for cultivating familiarity between research groups and particularly between their different methods and research questions. Having gone on a cruise, researchers not only had a better idea of what kind of data existed (which is an accomplishment and an important prerequisite to integration in and of itself) but also could reach out to those who had collected the data, having interacted with them on the cruise. In this way researchers drew on a variety of informal encounters and interactions when integrating data.

Corresponding with much of the literature on data sharing, we found data integration in the CBIOMES collaboration to be a situated and therefore complex process. The factors we point to here are neither exhaustive, nor discrete. Rather, the different dimensions interact with each other in data integration operations. Moreover, as in Crowston et al.'s [8] analysis of continuities and discontinuities, we found that having distance between researchers, along one of the dimensions identified here, was not necessarily an obstacle to collaboration, but rather could be beneficial. Given this complexity, we found it beneficial to shift our analysis away from cataloging the relevant factors and towards the work researchers engaged in in order to bring these multifarious factors into alignment.

4.2 Data Practices for Integration

In the previous section we three outlined dimensions of distance that, together, frame the numerous risks, rewards, and additional considerations of prospective data integration collaborations. The diverse individuals within the ocean science network navigate with and around all of these considerations, all the while being mindful of their particular situation within the network, and leveraging those connections in order to make the work of data integration go well. In this section we turn to the practices that constitute these navigations. Understanding these practices requires coming to grips with the fact that data integration occurs in a complex organizational context. Previous work on scientific cyberinfrastructures has discussed scientific research infrastructures as "ecologies" [57] and the multimorphous nature of human infrastructures [32]. Our investigation here does not trace out the outlines of the network of the "ecology", rather it describes six collaborative scenarios. These scenarios can be considered to be sketches of various kinds of collaborations (i.e. a subset of links between a subset of nodes) that are taking place within the larger multidisciplinary, geographically distributed CBIOMES collaboration. Without narrowing our view to a single collaborative connection over time, we capture the heterogeneity of collaborative relationships within the larger CBIOMES collaboration for which the CMAP research infrastructure is being built. The distinction between observation and modeling is a prominent one for the individuals throughout the CBIOMES collaboration as a whole.

Below we examine collaborative strategies among observationalists and observationalists and also between observationalists and modelers. The scenarios selected are meant to show a variety of kinds of coordination work practices. The first scenario in each section concern first engagements, instances in which researchers realize they might make use of another data type and attempt, for the first time, to figure out how to integrate it with their data. These scenarios highlight synergizing work in particular, as researchers make connections and establish a field of work. They also include descriptions of metawork and local articulation work as researchers negotiate how their research methods might connect, and what metadata and formats are required. The second two scenarios in each section highlight instances in which coordinations are more established in that articulation work between the researchers has been going on for a long time, but are also constantly being "reworked", to use Corbin and Strauss' term for the process of reopening established practices for negotiation and debate [7].

4.3 Collaboration Strategies: Observationalists with Observationalists

Scenario 1: Developing Shared Understandings Across Research Projects and Methods. In order to ensure appropriate use, scientists must not only come up with a common data format but they must also come to a shared understanding of underlying assumptions and data processing techniques. Harold, a postdoctoral researcher in a laboratory focused on ocean geochemistry, illustrated this problem to us in a situation from his work. While Harold's work focuses on geochemical processes, his research contributes to understanding how certain nutrients in the ocean support life. As a result it is useful for him to draw on more biological data types to complement his own data, and he had therefore reached out to Danica, a postdoctoral researcher at another university, whose PI

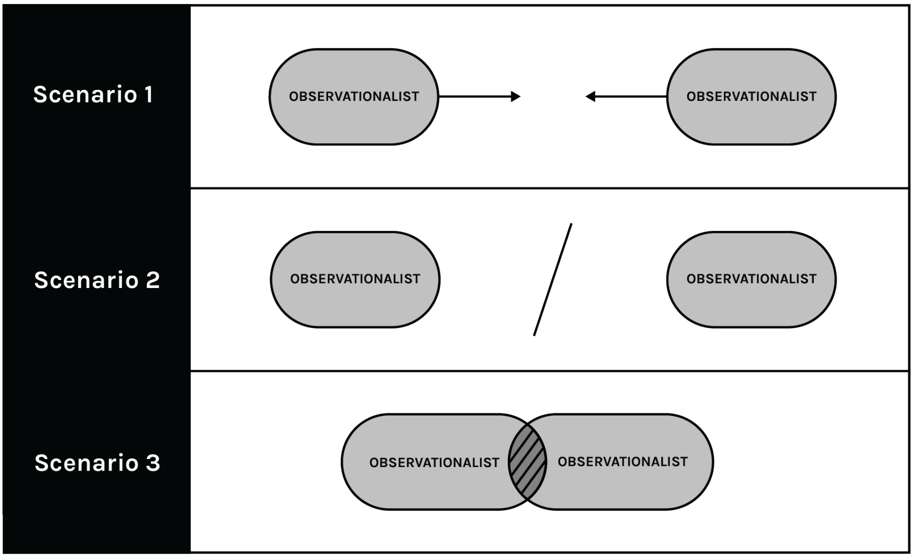

Fig. 1. Three scenarios demonstrating data integration practices between observationalists. Scenario 1 centers on first engagements with using another data type. Scenario 2 and 3 focus on more established coordinations, in which researchers actively avoid using each other's data (scenario 2) and collect data cooperatively (scenario 3).

is also part of the CBIOMES project. Danica's work centers around using metatranscriptomic data to explore relationships between different microbes in ocean energy cycles. The two researchers initially ran into difficulties trying to figure out how to apply metatranscriptomics data to Harold's work however:

> "those data types are obviously very rich, but it takes a lot of work to understand what they're saying. A lot of questions about how you normalize the data, and the ways to determine significance are relatively complex and holistic statistical process. I think that it's one of the challenges, to understand how that data actually... it's not an apples to apples comparison to an iron concentration, which you could say is known within an error that we can report. Numerous BlueJeans calls and email exchanges had to happen with Danica before we could figure out the best way to combine those datasets." (Harold, Postdoctoral Researcher)

Harold draws a distinction between iron concentrations and metatranscriptomics data in that there is a great deal of organizing and making sense that has to go into early processing steps for sequence data. As Harold described to us, "a lot of decisions are made" during these steps, and he would not trust himself to make or evaluate those decisions, since he does not have experience working with that method. Moreover, means for comparing this to iron concentration measurements are not well established or routinized in the field. For Harold and Danica, then, the difficulty of the problem comes in large part from its novelty. Understanding metatranscriptomics alone is not enough. Rather, successful integration emerges from a novel interaction between the two processes of measuring iron concentrations and metatranscriptomics. This interaction requires significant work on the part of both researchers together, through repeated teleconference discussions around how to compare the data.

Harold's discomfort with using data that he does not have experience with resonates with what has been described in the literature as the risk of misuse. Where researchers do not understand each others' methods, the issue of misuse becomes more salient. Researchers showed concern that making data accessible to others without the proper knowledge articulation and preparation (including documentation, annotation, and understanding of how the data was created, and established methods of making sense of data) could result in misuse or misinterpretation of the data. As one junior level researcher put it:

> "The challenges are when we don't speak the same language, or we don't speak the same scientific language. For example, most of my colleagues wouldn't know what to do with my genomics data. You could give them the data in a format that they're used to, but they still wouldn't know what to do with it. This is why I spend a lot of time making sure my documentation is tailored to the needs of whomever I'm sharing my data with … especially those that I know are outside of my field" (Demi, Postdoctoral Researcher)

It is a great concern if an untrained researcher who is inexperienced with a dataset uses unsuitable protocols or scripts for evaluation and analysis. This could result in a publication with potential errors, differing methods or findings, and inappropriate or misleading interpretations. Researchers are also aware of the potential for misuse when data are shared without proper knowledge articulation. One researcher spoke about the limitations to sharing when data are "used without a real good understanding about what the assumptions coming into it are" or when the data are "not processed in a way that best suits the question, because the question has changed" (Hamilton, Postdoctoral Researcher). For these reasons, many of the cases of data integration we encountered in CBIOMES resembled Demi's and Harold's cases. They were not so much a matter of sharing more broadly or documenting more thoroughly. Rather the ability to provide documentation hinged on first figuring out how datasets might be compared and then on methods of evidencing, such as that of processing metatranscriptomics data, could be brought into alignment with others. This work occurred largely through person-to-person discussion over BlueJeans calls.

Scenario 2: Scoping Work and Data Access to Protect Novelty of Research Contributions. In cases where different research groups used the same or similar methods and had similar research questions, researchers had none of the problems of understanding expressed in Scenario 1. While this potentially made it easy to integrate data sets, it also led to situations where researchers carefully scoped how broadly the data would be disseminated. Hallie, a Ph.D. student in a lab focused on ocean chemistry and biochemistry, described to one of the authors the concerns that went into one such decision:

> "There's a group out there that does the same stuff, same motivating questions as us, and we are on good terms with the group and everything, but it just would be uncomfortable for them to have access. ... It would be uncomfortable for me to have access to their data. I'd be curious to go at it, but you shouldn't..." (Hallie, Ph.D. Student)

The kind of metabolomics data that Hallie worked with required a high degree of specialized knowledge to make sense of. Indeed, Hallie explained to us that most of her colleagues in CBIOMES would have difficulty making sense of it (a situation similar to that of Scenario 1), and so sharing it widely before she had published it would not be problematic: "no one would know what to do with it. I wouldn't care". However, metabolomics data also takes significant thought and effort on the part of the researcher, and there was a sense that the person who went through this effort should have "a crack at it first".

These concerns reflect what has been discussed in the literature as scooping: the intentional or inadvertent publication of research or a result that someone else has been working on. However, in the CBIOMES collaboration, the issue of scooping was not an issue of scooping actually happening. To our knowledge there were no cases of someone being scooped by somebody else in the collaboration, and, as in the quotation above, using other researchers' data inappropriately was to be avoided as much as having one's data used inappropriately. The "issue" was more a continual process, observed by people on all sides, of carefully scoping the movement of data to avoid impinging on others' research. Stefan, a Ph.D. student in a biologically-oriented lab described a particular instance of this:

> "What I've done is email the lead PI, in this case, Abigal, and cc my advisor and my collaborators here at the [U.S. University], saying, 'Okay, we have this data set that's complete. We're posting it on the shared folder on the Google drive, so it's there. Let us know if you have questions and let us know if you want to use it, if you intend to use it, or, excuse me, for what type of analysis.'" (Stefan, Ph.D. Student)

Placing the data in a Google drive which is open to the wider collaboration is more constrained than sharing it on a public repository, and he establishes expectations around keeping him and his PI aware of any subsequent usage. The act of cc'ing his PI reflects a broader tendency for responsibility around data sharing to shift upwards along the laboratory and collaboration hierarchy. When asked how they made their data available to others, many Ph.D. students and more junior lab members stated that it was not really their responsibility to make those decisions. Such decisions were often made by PIs, but with responsibility towards the work of their students and postdocs. Stefan himself suggested to us that PIs are often responsible for the decision to make data available (and how to), but that "the priorities are for their graduate students, their postdocs, so sometimes it's at their [the students' or postdocs'] discretion". Some PIs described a process by which the larger Simons-funded projects (CBIOMES, SCOPE, and SCOPE-Gradients) have ongoing conversations to decide "which papers they think will evolve and which data is going to be in which paper" (Ivette, PI). Such decisions were not often made by individuals, and how the data would be integrated into others' research mattered in the decision.

Scenario 3: Upstream Coordination of Work to Reduce Need for Downstream Data Integration Across Research Projects. Two of labs in particular, located at the same university, had established a close working relationship and surfaced a number of practicable ways of integrating different data types. One of the two labs uses metatranscriptomic, and primarily biological methods, whereas the other works mainly with biochemistry-oriented metabolomics methods. Danica, a member of one of the labs, described to us how they were working together on a particular project to analyze nutrient cycling using both metatranscriptomic and metabolomic data collected in the field, in combination with lab-grown cultures. The proximity of the two labs is an important component of this work. Danica, somewhat jokingly said that:

> "Lucky for me, there's a chemical oceanography lab next to ours that uses advanced mass spectrometry tools as part of research work to elucidate microbial processes in natural systems. The PI and Ph.D. students from that lab are currently close collaborators on this sulfur cycling project. We've had a lot of in-person meetings and email exchanges early on to try and figure out how to best combine laboratory studies with metatranscriptomic and metabolomic field analyses." (Danica, Postdoctoral Researcher)

Members from the two labs have worked together to hash out how data produced from disparate methods might be mobilized against the same research questions in a sound way. PIs and students from the two laboratories had also gone on data collection cruises together in the past, an activity which both made them familiar with each other's data collection processes, and established connections between them that facilitated later interactions.

One particular point of coordination in the two labs' data integration efforts centered around using both metatranscriptomic and metabolomic data to develop a fuller picture of nutrient cycling in an ocean ecology. On this point, researchers from the two labs were able to establish a kind of complementarity between the two data types by trying to establish a connection between particular metabolites and the genes that (through the activities of various microbial organisms) ultimately produce them. Gareth, a Ph.D. student who worked primarily with metatranscriptomics data, described this process:

> "We usually work really closely with them, in part because they are down the street. But you know it's also complementary, you know. They look at the chemical in the water. We look at the genes that make them. And it's, you know, I think they really strengthen each other. So we coordinate it, we'd be like 'ok, let's all go out at the same,' we'd sample from the underway tap together, so we basically have, all of our samples are right on top of each other." (Gareth, Ph.D. Student)

Sampling at the same time allows researchers from the two labs to preempt the problem of integration somewhat, precluding some of the difficult contingencies of reusing data by making sure that subsequent comparisons are based on collections from the same location at the same exact time. While this particular practice is not common, nor likely amongst physically dispersed research groups, it exemplifies highly articulated data integration efforts in the collaboration. Data is collected by researchers from each lab with the notion of its potential value in integration efforts in mind, and with an understanding of the data collection process that will make the data more valuable to those efforts.

Across all three scenarios above we see evidence of the three kinds of articulation work: local articulation, metawork, and synergizing. These three scenarios, all involving observationalists integrating data with that of other observationalists, indicate how providing data and documentation for particular cases of data integration require researchers to relate their respective methods and work processes to each other in some way. We see this as a form of metawork, a hashing out of related lines of work, and its relation to a kind of local articulation work, the provision of data for a researcher's ongoing project. Previous work on coordinative protocols coined the term of "brackets" to indicate that the same coordinative mechanism can be use to simultaneously keep actors together and in sync at the same time that they keep actors apart [21]. In a similar fashion, the successful longevity of a collaboration like CBIOMES requires not just figuring out how to integrate data but also, as in Scenario 2, it requires some coordination and mutual awareness to maintain distinction between data sets, research questions, and work.

4.4 Collaboration Strategies: Observationalists with Modelers

Scenario 4: Creating New Shared Understandings to Ensure Appropriate Use Across Research Projects. Data integration challenges live alongside the challenges of imagining and creating processes and tools for new kinds of scientific inquiry. To be sure, this also happens in the observationalist to observationalist collaborations, but given that part of the mandate of the CBIOMES effort is to find new ways to leverage modeling, the integration of observational data into modeling work is a point of active engagement in the collaboration. Thus by design, part of the work of observationalist to modeler collaboration is to close the distances discussed above.

Scenario 4 involves a modeler attempting to integrate data sets into a model where they have not previously been used. The modeler receives a dataset by email or by download from a public

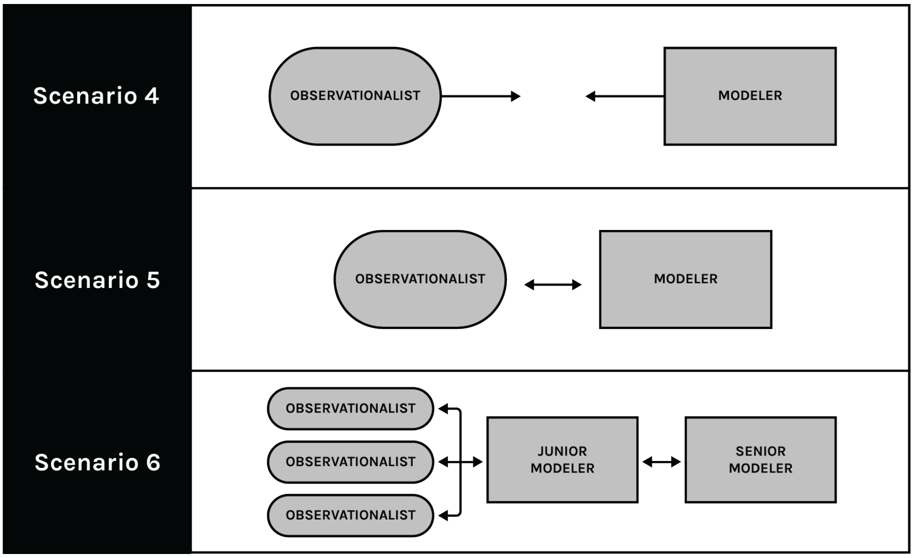

Fig. 2. Three scenarios demonstrating data integration practices between observationalists and modelers. Scenario 4 centers on a modeler's first attempt with using a new data type. Scenario 5 focuses on an instance where an established working relationship between researchers exists while Scenario 6 centers on the emergence of new roles for integration work.

repository, but she is not sure how to make sense of one of the fields. She has not used this particular data type before or been on any of the cruises on which they are collected. There is also no clear documentation of what the fields mean. Through a follow-up with the observationalist who sent the data, she realizes that the field represents "stations" from a number of different cruises rather than from a single cruise. Further problems emerge however, as she also needs to find out what precisely was measured. Another observationalist provides data on chlorophyll, but the modeler discovers that the observationalist in fact derived chlorophyll metrics from a number of other data points, and the uncertainty of that derivation is unclear. The process of derivation affects how and whether the data can be used in her modeling work. As in Scenario 1, critical aspects of the data remain worryingly uncertain for the modeler. While this kind of problem of missing metadata was a problem for observationalists also, it was a persistent and somewhat characteristic aspect of modeling work. Because modelers did not collect empirical data directly, they were frequently in the position of trying to understand data sets generated by others. This involved both aspects of metadata, such as understanding what different fields were meant to indicate, but also more specific metadata around the provenance of the dataset, the transformations and cleaning methods that had been applied to it. A great many more email exchanges are required, through which the observationalist explains the elaborate procedures they go through to process their data. These are in many ways the classic problems of data exchange, but we want to highlight here is the process of becoming familiar or establishing practices for integrating particular kinds of data for particular purposes. Another modeler described this process:

> "The station data that was measured in some of the cruises that we looked at first, but we decided to focus on this underway stuff for now. There's a little bit of, there's some documentation out there, and then we have to deal with it ourselves at this end. Just to both learn what's available and what happened in the actual cruise as opposed to

> what was planned, and all those things. We typically prioritize obtaining and using macromolecule measurements and stoichiometry and those sorts of things because we currently know best how to constrain models using those things." (Harvey, Research Engineer)

Here we can see that the various contingencies and uncertainties around how to use datasets create a differentiated field of data products, where some are familiar and can be used quite easily, whereas there are no established practices for others and a lot of work in the form of discussion with observationalists or poring over documentation would be needed.

From the observationalists' side, it is difficult to know what data is valuable or how it might be used. This includes not only how to prepare the data, but also how it might be used and what aspects of the data collection and processing work might affect how the modeler makes use of the data. Without clear communication from the modeler, it is not clear what data products would be useful. An observationalist described this situation:

> "You know, we can measure all kinds of shit, you know. It's nothing for us to think about measuring something new. It would be, I would love a list from [the senior modeler] saying, hey, I need to know this. This is the grand challenge. Can you get me this in a year? … I think [senior observationalist] and I could go and brainstorm and maybe deliver that. But we need them [the senior modeler] to tell us." (Morris, PI)

Participants ascribed some of these interests on the part of modelers to established practice in the field. Certain kinds of biological data had long been integrated into such models, whereas for other kinds of data, particularly those produced by chemical oceanographers, it was not as clear how it might be added to a model. On multiple levels, then, the problem of making data from an observationalist workable in a modeling context was a mutual problem. Observationalists were unsure of what data would be useful and how the data could contribute to a modeling effort; while simultaneously, modelers were uncertain of what data were available; and did not have the practical knowledge of how to make use of those data. Superficially, this problem may seem very similar to Scenario 2 above, but in this scenario we find that the object of discussion often has to do with how to create a new field of work, and all that entails, at least as often as it was about how to integrate data in existing fields of work.

Scenario 5: Bringing Independent Practices into Alignment to Facilitate Integration. Some of the observationalists had significantly more experience with modeling methods, and could therefore understand the modeler's needs, including what variables and parameters were ideal to include when preparing data for integration into a model. However, alignment between the secondary use of data and the processes of production was not automatic, but rather had to be established and practiced. Zack, a PI who specialized in experimental work with biological laboratory cultures, described a longer running collaboration he had with a modeler who was using his data. Abe, the modeler, had reached out to Zack asking to use data Zack produced for a particular publication on the effects of the day and night cycles on virus and host interactions. However, complications arose in reusing the data that had already been produced:

> "They did really great experiments, but they stopped the experiment really early because it's a lot of work. We tried to ask them to have perhaps more small points after this, just to extend a bit of data, just to see the dynamics. But it seems that we don't have the same goal, really. I know it takes experimentalists a lot of time and money for them to obtain data. It also speaks to the differences in our processes. As modelers, we see a process perhaps more in this way, and the experimentalists think more perhaps in terms of measurements, in terms of methods to measure it." (Abe, Post Doctoral Researcher)

The data set that Zack had initially produced had what Abe considered to be a short timescale, and only varied light exposure, and not variance of nutrients or temperature. Moreover, the data initially came in parts with some parts having different kinds of error or summary statistics reported, requiring Abe to ask for the raw data that had produced it. This prevented Abe from being able to rule out other biological processes than the one he was interested in. All of these things seemed to be implicated in Abe's suggestions that the two labs were engaged in distinctly different work processes. The observationalist was more concerned with specific means of measurement, and, as Zack's lab was paying in money and labor for each datum produced, they worked on a smaller scale when it came to producing data to answer a specific question. The modeler, using data in a secondary capacity, was less concerned with the cost and labor, but more concerned with the standardization of the datasets and the breadth of the data in terms of time and the variation of particular parameters. The issue, then, was not only merging a specific data set, but also figuring out a way of making two distinct production models work together.

Because the initial datasets could not be used, Zack agreed to produce more data that would be tailored to the specific needs of Abe's model. It was fairly difficult at first to understand how this would be done. After some difficulty in communication Zack physically visited Abe's institution in order to get a better sense of what was needed.

> "So I also eventually came out to [U.S. University] to visit Abe with a PhD student from my lab to learn more about his modelling efforts. I think being together in person was really important. Simply put, my lab is composed of researchers that are truly experimenters and we don't always understand what's going on with the modeling which is partially due to many of us not having a mathematical background like those in Abe's lab. We're really speaking two different languages here."(Zack, PI)

In producing a second round of data, Zack attempted to change his practices of data production to account for a particular site of secondary usage. Moreover, as Zack explained, he was subsequently more familiar with the kinds of formats and metadata that would make processing in the model stage easier. Differences in the make of the data were not entirely overcome. From Zack's position, he would have very much liked to contribute data to the modeling effort, but it was not economically feasible within the data production processes of his lab:

> "From my experience, modelers are used to asking for a lot of data from other observationalists and getting it. As much as we would like to do the same, it's not really feasible. We operate a much smaller lab in terms of the funding we have available as well as the number of researchers we have to do the work when compared to other labs." (Zack, PI)

This instance of data integration verges on a more traditional category of paper collaboration, in which two researchers dedicate effort towards a single paper. This is in part because of the bespoke nature of the integration work. Rather than reusing data that had already been produced and were lying around in a data repository somewhere, as is the case with many modeling efforts, Zack actively produced the kind of data needed for the model based on consultation with the other researcher. It is through this particular interaction, however, that processes of data integration become established between two different research situations, and two different "makes" of data.

Scenario 6: The Emergence of New Roles for Integration Work. One major subproject of CBIOMES involves a large number of labs collaborating with a single modeling lab in order to figure out how to bring a variety of data types together. The variety presents a large obstacle to the group; as it includes sequence data such as metabolomics, transcriptomics, and lipidomics; organism counts and measures; and environmental data such as ocean salinity, temperature, and wind speed, collected mainly from ocean buoys. Malcom, a Ph.D. student in the modeling lab, has taken on the work of facilitating this process as part of his doctoral work. Malcom is one of a number of Ph.D. students from the lab who work with specific data producers based on their prior experience with the data types:

> “There are a whole bunch of different groups, maybe nine or ten groups in total that are involved in this specific collaboration, including us. And so our job from around the time I arrived was trying to centralize and standardize the data coming from all of those groups so that they could be put into the same sorts of computational and analytic pipelines and also so that results of analyses of those data sources would be mutually intelligible with each other. And how our group has chosen to divide up the responsibilities of interfacing with those other groups is mostly around our specific familiarities with the data type.” (Malcom, Ph.D. Student)

What is interesting about Malcom’s situation is that data integration is not a process within his work, but rather its central object. In addition to facilitating the integration of data into larger models, the interaction between modeler and observationalist here also implies an effort to establish a larger analytical frame for the data from different researchers. While Malcom has some prior experience with biological data, he must still work with the group collectively and individually in order to figure out ways of making statistical comparisons between, for instance, variables presented by lipidomics and metabolomics data. This process involves Malcom facilitating group telecons with the participating researchers to discuss how this level of data integration might be accomplished, as well as working with observationalists individually to develop a meticulous understanding of their data. Bethany, who is contributing metabolomic data to this project, described this process:

> “He’s been a good person to talk to about it. He knows a lot of statistics. Basically I was worried about the fact that we had time series data and trying to figure out what you can say is significant, and the best way to detect periodicity, and so we’ve just talked about those things. I guess because he’s in the Davis group, which has more statistical expertise than ours. He’s convinced me I’m doing things okay. I think he’s had some ideas about how to do some things slightly differently but it’s not dramatically differently.” (Bethany, Ph.D. Student)

Malcom and Bethany had previously met via telecon to go through each of the columns in her data set one by one so that she could explain what they represented. However, it is important to point out that, as with the modeler / observationalist interaction in Scenario 4, the process of integrating data into a model is not just a matter of the person who produces the data providing a thorough description. Bringing multiple kinds of data together means forming a new understanding of how they might be used as evidence collectively, and where it crosses boundaries between statistical modeling methods and the chemically centered knowledge of the observationalist, a new understanding has to be developed. In other words, Malcom’s work is not just the merging of formats (though that is part of it), but also the relating of distinct tasks. Moreover it is a compromise between different makes of data, the merging of processes for producing datasets tailored to specific research questions with a process of making many such data sets and their parameters comparable.

Integration efforts between observationalists and modelers are characterized by many of the same practices in similar efforts between observationalists. However, the work of modelers is in part defined by the work of data integration. In Scenarios 4, 5, and 6, the ability of a modeling effort to yield results depends on establishing a working relationship between the modeler and the observationalist. We again point to this as a kind of metawork, in which the line of work of the observationalist, in collecting, cleaning, and analyzing a particular kind of data, must be meshed in some way with the work of developing a statistical model. In Scenario 5 this implies not just the provision of correct metadata, but a change in the kind of data generation the observationalist carries out. In Scenario 6, the modeling project serves to establish a field of work, within which the metawork of planning and relating different data collection efforts between the central object of work for a particular Ph.D. student.

5 DISCUSSION: DATA INTEGRATION AS ARTICULATION WORK

The coordinative work of the CBIOMES collaboration can be thought of simultaneously in terms of the personal and intimate craft of piecing together a quilt and in terms of the layered agility and flexibility of the computer networks upon which the collaboration's science depends. The coordinative work of integrating data is certainly about reconciling data with different histories, assumptions, and purposes. However, careful attention must also be paid to the way that these problems interact with the issues of integrating manifold ways of working, including workflow and division of labor, which themselves must be reconciled and reinvented across different time frames and different places.

5.1 Data Integration as Articulation Work

Viewing data integration in terms of articulation work gives us a potential way forward from the discussion of data integration, reuse, or sharing based solely on barriers and motivations. Isolating and typologizing such barriers (such as scooping, misuse, trust, and extra labor) is a valuable contribution, but part of our point here is that the potential factors that might come to bear on decisions around data sharing are multifarious, and have interactional effects. Individual stakeholders are therefore operating from situated positions, and must make their own calculations about the possibility and benefit of data integration. Decisions about formats, processes, and policies of data integration cannot be made once at the PI or institutional level and then be expected to capture the ongoing contingencies of individual researchers' concerns. This is not to say that the notion of data integration is hopeless. Researchers can and do align their concerns in order to successfully integrate data. Similar to the case described by Lee et al. [32], researchers leverage a variety of formal and informal relations across different working groups in order to accomplish this task successfully.

The work of data integration cuts across a number of different levels of articulation work. Researchers engage in local articulation work by figuring out, in a particular situation, how to provide data to others efficiently and address questions collaborators have about metadata fields, acronyms, and so on. This is necessary to make sure that all necessary resources are in place for the data to be reused, and that workers are meshed with tasks through understanding the data and metadata. Researchers also engage in a kind of metawork, which Gerson [21] describes as "making sure that different kinds of activity function together well." Researchers do this by holding teleconferences with collaborators to discuss the documentation they need to make sense of the data or to figure out how to merge datasets with different spatial and temporal resolutions. Local articulation work, such as providing the right metadata elements are developed along with the metawork of meshing different research analyses. Providing the correct metadata is dependent on figuring out who the metadata is for and what metadata is relevant in the first place based on how the data might be reused. It is under particular understandings of the relationship between sites of work that the provision of specific metadata elements and aspects of data provenance can be determined. This is not to say that metawork always precedes local articulation work of that type, but that between researchers in CBIOMES they emerge together. In this sense metawork sets the stage for creating practices and artifacts that can transcend a particular moment or locale.

At a higher level, the CBIOMES collaboration engages in synergizing [1] by bringing individuals, groups, and labs together in new ways to establish common fields of work. Through interactions on research cruises, writing collaborative grants, BlueJeans calls for various collaboration meetings, and attending all-hands meetings, we see researchers develop a mutual sense of each others' work. These interactions allow them to establish appropriate boundaries around data products (Scenario 2), as well as develop mutual understandings of data types, and working operations, including data cleaning practices and formats (Scenarios 1 and 5). Connections between collaborators are also made by third parties, who, positioned between the two groups, are able to connect the needs of specific modelers with the capacities of observationalists (Scenario 6). These activities establish the foundation for collaboration in making the very existence of particular data types known, as well as facilitating discourse around the possibilities of reuse of those data for one's own research questions, and the viability of labor arrangements which would enact those integration efforts. As described in Bietz, Baumer, and Lee [1], these different kinds of articulation work are carried out recursively, such that the central work task of one person may be supra-work to another (whether synergizing or metawork or local articulation work).

The concept of articulation work refocuses our attention on the emergence of new relations between researchers in data integration efforts. Significant work has been done to show how data must navigate the situations of use in which they are mobilized as evidence [34]. Still, in many accounts researchers and labs are posited as discrete or isolated situations for data production or reuse. The contemporary digital repository emphasizes this perspective, presenting a kind of "dead drop" view of data sharing, where one researcher deposits data in a repository, and then another, individually and separately, comes to pick it up and reuse it. By contrast, the notion of articulation work implies that the processes of data analysis and production articulate in relation to other tasks and people, and the work of bringing those sites into alignment is a process of mutual negotiation and engagement. Data integration is a process of working with the various processes that might or must be carried out with the data and distributing them in time and across collaborating laboratories, aligning tasks, and workers with differentiated skill sets. The researchers must jointly figure out whether data can be brought from one place to another feasibly, and whether it can serve a purpose there. Such negotiations are particularly evident in the examples of observationalists and modelers working together and trying to establish mutual understandings of how data might be mobilized from the situation of production to the situation of an integrated model (Scenarios 4, 5, and 6).

This is not to say that the dead drop situation above is impossible, but rather that where such impersonal exchanges are carried off successfully, they are embedded in established practices of one variety or another. As described by Ribes [51], making data interoperable is a historical process, a negotiation that may become embedded in the tools and practices of a research infrastructure, and become invisible in subsequent, day-to-day exchanges of data. In fact it is quite possible, when looking back retrospectively on those data types for which standards are in place and which are now shared and integrated quite freely, to think that such is possible because the metadata was provided, or because the extra work of documenting was committed to the task. However, we want to point to the fact that the work of figuring out which metadata is the right metadata, which documentation is needed, and whether the integration of disparate data sets is even sound or useful is something which itself needs to be determined. Seamless data integration therefore relies on prior work having already been accomplished: the relating of tasks to tasks and tasks to workers in larger arcs of work. In other words, wherever data journeys efficiently, it does so on rails built out through collaborative practice. For the researchers in our study, articulation work is exactly that set of practices which creates new collaborative configurations that allow data to be mobilized in new ways.

As the work of CBIOMES collaborators is nascent, we are observing that the professional roles implied by data-centric science, along with their visibilities, responsibilities, repertoires, and relations to larger arcs of work, are somewhat in flux and still being established—what we call collaborative nascence [33]. The nascence of data integration efforts is also visible in the emergence of niche working arrangements in the collaboration. In Scenario 6, for instance, we see more junior modelers take on the role of go-between (because of their career position and their skillset), trying to figure out potentially useful processions of data from observationalists to modelers. In these roles tasks such as bringing different research groups into dialogue or trying to bring data from one lab to bear on research questions developed by another, become their core tasks, rather than metawork. Crowston et al. [8] similarly identifies this state of nascence in the way that continuities are brought into being around breakdowns or discontinuities in work processes, and examines the role of intermediaries in navigating these breakdowns. As Strauss [61] initially observed, new niche working roles and divisions of labor often emerge from the articulation work around nonstandard or disrupted work arrangements. In looking at articulation work, then, we can examine a “science in the making” [30], rather than a readymade science of established, standardized data sharing. We can furthermore avoid taking the conditions of successful data integration, and the divisions of labor it implies, as given.

5.2 Multi-Sited, Multiple Interwoven Problems

The CBIOMES project can be seen as not only as the home of a highly nascent research infrastructure development project but can also be seen as a complex ecology of many collaborations and many collaborative strategies. CBIOMES is a site where the day-to-day work of data integration is under construction. Relations between data production and data modeling activities that are negotiated in the work of CBIOMES collaborators become embedded in data formats, practices, and skill sets. The articulation work and synergizing work being carried out by researchers in the CBIOMES collaboration are therefore significant as sites where new relations between research labs, disciplines, and methods are becoming embedded.

In this sense the CBIOMES collaboration is representative of a proliferating trend in the sciences: the bringing together of groups who must collaborate across differing evidentiary practices, and bring new tools and methodologies to bear on similar related research questions. We argue that this nascence is critical to understanding the problem of data integration, and that it has implications for the technical interventions we make in large-scale data-integration efforts. Following Leonelli [34], we can see the evident promise of a data-centric model of science as the generative movement of data into novel research situations where they can be reused in new and unexpected ways. However, this implies that, as Tenopir et al. [65] have pointed out, that researchers often do not even know how their data might be used or how it might be valuable to others prior to data integration efforts. As in Scenario 1, modelers are uncertain what data are out there and how they might be used, while observationalists are conversely uncertain what kind of data modelers want, or what they might be able to do with the data the observationalist already has. Our findings suggest that a core process in making data integration attractive to researchers is the process of hashing out, with others, how the data might be reused. In other words, the problem of data integration is not just one of motivating or cajoling researchers to engage in a known process, but of hashing out complex collaborative processes which do not yet exist. The nascence we describe here is not only a matter of standardizing formats, but also figuring out what might be possible, and what is sound, in the transport of data from one situation to another.

This work not only cuts across multiple levels of articulation work, but also involves the tension or resolution of multiple data sharing concerns. In this sense, the researchers in the CBIOMES collaboration are in a long-term process of constructing what Fujimura calls “do-able problems” [18]. Fujimura follows a group of cancer researchers, and observes how they articulate the concerns of their larger social world (what research is publishable and novel in the community) with career concerns around graduating, getting hired, and coping with technical and internal lab concerns. In the CBIOMES collaboration, too, we see the resolution of tensions between research projects, the availability of labor within a lab, the defense of novel contributions for the sake of Ph.D. students' careers, and the efficient publication or sharing of data with other involved labs who have other research questions. A do-able problem, then, is constructed from the alignment of organizational concerns, professional interests, labor requirements, differentiated skill sets, and contribution to a broader community. The notion of a do-able problem does not assume the natural alignment of these different dimensions, but rather highlights precisely the contingency of many factors in resolving what seems like a localized issue. We have tried to show here that working out a do-able process of data integration results not just in a novel data format at a technical level, but can also produce new organizational and professional activities and responsibilities, such as with Malcolm's work as intermediary in Scenario 6.

This work has shown that researchers develop new practices and coordinative relationships to facilitate data integration by building on and reworking the dynamic collaborative network of CBIOMES. The problem of data integration then complicates simplistic notions of exchange and format standardization, and becomes one that is fundamentally about shared sense making, collaborative learning, and of multiply-situated coordination. Collaborators find that the distances that must be bridged and the number and intensity of collaborative strategies vary for each instance of data integration. The work of data integration is, sometimes, the work of creating new do-able problems and creating the means for asking and answering new questions. In all of this, coordination in multiple sites–sometimes interconnecting or overlapping–in a larger collaborative milieu plays an essential role.

6 CONCLUSION

This paper has investigated scientists' everyday activities in navigating the various concerns of data integration. Stakeholders experience a large number of intersecting concerns emerging from their seniority, mutual familiarity, and disciplinary and methodological differences. Researchers work to bring their data collecting, cleaning, and sharing efforts into alignment with downstream situations of reuse. In this process we highlight the various kinds of articulation work which must go on in order to construct lines and arcs of work for the efficient movement of data between situations, to match workers to tasks, and to bring essential resources into the relevant work site.

We outline a number of ways in which a practice-based approach shifts our attention in the study of data integration, and, more broadly, on data sharing and reuse. Firstly, it refocuses our attention towards the practices that researchers enact on the ground in order to navigate this complexity. Secondly, it highlights the work researchers engage in to figure out how to integrate data, and not only their considerations of whether or when to integrate data. This moves us beyond a simple barriers and incentives model, and casts the problem of exchanging data in the sciences not as a system of walls that need to be knocked down, but a system of bridges that might be built. For the designer of data portals and other data sharing systems, this might mean moving away from large, top down standardization efforts, and towards a situated design effort. This would involve looking for sites of collaborative practice where interoperations between data types, methods, and ways of working are already being hashed out for particular ends by researchers and widening these paths, rather than examining why data sharing is not happening in a given situation. Thirdly, it draws our attention to the processes by which new roles and divisions of labor emerge around the project of data integration between newly collaborating fields or methodological communities.

7 ACKNOWLEDGEMENTS

This article is based on work supported by National Science Foundation grants (#1302272 and #1954620) and the Simons Collaboration on Computational Biogeochemical Modeling of Marine Ecosystems/CBIOMES (Grant ID: 549945, E.V.A). We thank our anonymous reviewers as well as Dr.

David Ribes and Dr. Daniela Rosner at the University of Washington and Dr. Drew Paine at Lawrence Berkeley National Laboratory for their feedback. Finally, we are grateful to the members of the Simons Foundation oceanographic research initiatives (CBIOMES, SCOPE, and SCOPE-Gradients) for their participation in our study.

REFERENCES

[1] Matthew Bietz, Eric Baumer, and Charlotte P. Lee. 2010. Synergizing in Cyberinfrastructure Development. *The Journal of Collaborative Computing* 19, 3 (2010), 245–281. https://doi.org/10.1007/s10606-010-9114-y

[2] Matthew J Bietz and Charlotte P Lee. 2009. Collaboration in metagenomics: Sequence databases and the organization of scientific work. In *ECSCW 2009*. Springer, 243–262.

[3] Christine L Borgman. 2008. Data, disciplines, and scholarly publishing. *Learned publishing* 21, 1 (2008), 29–38.

[4] Christine L. Borgman. 2012. The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology* 63, 6 (2012), 1059–1078. https://doi.org/10.1002/asi.22634

[5] Kathy Charmaz. 2006. *Constructing grounded theory : a practical guide through qualitative analysis*. London ; Thousand Oaks, Calif. : Sage Publications, London ; Thousand Oaks, Calif.

[6] Su Yun Chung and Limsoon Wong. 1999. Kleisli: a new tool for data integration in biology. *Trends in Biotechnology* 17, 9 (1999), 351–355. https://doi.org/10.1016/S0167-7799(99)01342-6

[7] Juliet M Corbin and Anselm L Strauss. 1993. The articulation of work through interaction. *The sociological quarterly* 34, 1 (1993), 71–83.

[8] Kevin Crowston, Alison Specht, Carol Hoover, Katherine M. Chudoba, and Mary Beth Watson-Manheim. 2015. Perceived discontinuities and continuities in transdisciplinary scientific working groups. *Science of the Total Environment* 534, C (2015), 159–172. https://doi.org/10.1016/j.scitotenv.2015.04.121

[9] Andrew K. Dow, Eli M. Dow, Thomas D. Fitzsimmons, and Maurice M. Materise. 2015. Harnessing the environmental data flood: a comparative analysis of hydrologic, oceanographic, and meteorological informatics platforms.(ESSAY). *Bulletin of the American Meteorological Society* 96, 5 (2015), 725. https://doi.org/10.1175/BAMS-D-13-00178.1

[10] Paul N. Edwards, Steven J. Jackson, Geoffrey C. Bowker, and Cory P. Knobel. 2007. Understanding infrastructure: Dynamics, tensions, and design. (2007).

[11] Paul N. Edwards, Matthew S. Mayernik, Archer L. Batcheller, Geoffrey C. Bowker, and Christine L. Borgman. 2011. Science friction: Data, metadata, and collaboration. *Social Studies of Science* 41, 5 (2011), 667–690. https://doi.org/10.1177/0306312711413314

[12] Robert M. Emerson, Rachel I. Fretz, and Linda L. Shaw. 1995. *Writing ethnographic fieldnotes*. University of Chicago Press.

[13] Ixchel M. Faniel and Trond Jacobsen. 2010. Reusing Scientific Data: How Earthquake Engineering Researchers Assess the Reusability of Colleagues' Data. *The Journal of Collaborative Computing* 19, 3 (2010), 355–375. https://doi.org/10.1007/s10606-010-9117-8

[14] Ixchel M. Faniel and Elizabeth Yakel. 2017. Practices do not make perfect: Disciplinary data sharing and reuse practices and their implications for repository data curation. *Curating research data, volume one: Practical strategies for your digital repository* (2017), 103–126.

[15] Benedikt Fecher, Sascha Friesike, and Marcel Hebing. 2015. What drives academic data sharing? *PloS one* 10, 2 (2015).

[16] Sebastian S Feger, Sünje Dallmeier-Tiessen, Paweł W Woźniak, and Albrecht Schmidt. 2019. The Role of HCI in Reproducible Science: Understanding, Supporting and Motivating Core Practices. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–6.

[17] Thomas A Finholt. 2002. Collaboratories. *Annual review of information science and technology* 36, 1 (2002), 73–107.

[18] Joan H. Fujimura. 1987. Constructing 'Do-able' Problems in Cancer Research: Articulating Alignment. *Social Studies of Science* 17, 2 (1987), 257–293. https://doi.org/10.1177/030631287017002003

[19] Joan H. Fujimura. 1996. *Crafting science : a sociohistory of the quest for the genetics of cancer*. Cambridge, Mass. : Harvard University Press, Cambridge, Mass.

[20] Mike J. Gallivan. 1997. *Value in triangulation: a comparison of two approaches for combining qualitative and quantitative methods*. Springer, 417–443.

[21] Elihu M. Gerson. 2008. *Reach, bracket, and the limits of rationalized coordination: Some challenges for CSCW*. Springer, 193–220. https://doi.org/10.1007/978-1-84628-901-9_8

[22] Vidar Hepsø et al. 2006. Intelligent energy in E&P: When are we going to address organizational robustness and collaboration as something else than a residual factor?. In *Intelligent Energy Conference and Exhibition*. Society of Petroleum Engineers.

[23] Tony Hey and Anne Trefethen. 2003. e-Science and its implications. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 361, 1809 (2003), 1809–1825. https://doi.org/10.1098/rsta.2003.1224

[24] Steven J Jackson, David Ribes, Ayse Buyuktur, and Geoffrey C Bowker. 2011. Collaborative rhythm: temporal dissonance and alignment in collaborative scientific work. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work*. 245–254.

[25] Marina Jirotka, Charlotte Lee, and Gary Olson. 2013. Supporting Scientific Collaboration: Methods, Tools and Concepts. *The Journal of Collaborative Computing and Work Practices* 22, 4 (2013), 667–715. https://doi.org/10.1007/s10606-012-9184-0

[26] Cutcher-Gershenfeld Joel, S. Baker Karen, Berente Nicholas, R. Carter Dorothy, A. Dechurch Leslie, C. Flint Courtney, Gershenfeld Gabriel, Haberman Michael, King John Leslie, Kirkpatrick Christine, Knight Eric, Lawrence Barbara, Lewis Spenser, W. Christopher Lenhardt, Lopez Pablo, S. Mayernik Matthew, Mcelroy Charles, Mittleman Barbara, Nichol Victor, and Nolan Mark. 2016. Build It, But Will They Come? A Geoscience Cyberinfrastructure Baseline Analysis. *Data Science Journal* 15, 0 (2016). https://doi.org/10.5334/dsj-2016-008

[27] Helena Karasti, Karen Baker, and Florence Millerand. 2010. Infrastructure Time: Long-term Matters in Collaborative Development. *Computer Supported Cooperative Work (CSCW)* 19, 3 (2010), 377–415. https://doi.org/10.1007/s10606-010-9113-z

[28] Youngseek Kim and Ayoung Yoon. 2017. Scientists' data reuse behaviors: A multilevel analysis. *Journal of the Association for Information Science and Technology* 68, 12 (2017), 2709–2719. https://doi.org/10.1002/asi.23892

[29] Kateryna Kuksenok, Cecilia Aragon, James Fogarty, Charlotte Lee, and Gina Neff. 2017. Deliberate Individual Change Framework for Understanding Programming Practices in four Oceanography Groups. *The Journal of Collaborative Computing and Work Practices* 26, 4 (2017), 663–691. https://doi.org/10.1007/s10606-017-9285-x

[30] Bruno Latour. 1987. *Science in action: How to follow scientists and engineers through society.* Harvard university press.

[31] Bruno Latour and Steve Woolgar. 1986. *Laboratory life : the construction of scientific facts.* Princeton, N.J. : Princeton University Press, Princeton, N.J.

[32] Charlotte P Lee, Paul Dourish, and Gloria Mark. 2006. The human infrastructure of cyberinfrastructure. In *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work*. 483–492.

[33] Charlotte P Lee and Drew Paine. 2015. From The Matrix to a Model of Coordinated Action (MoCA) A Conceptual Framework of and for CSCW. In *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*. 179–194.

[34] Sabina Leonelli. 2013. Integrating data to acquire new knowledge: Three modes of integration in plant science. *Studies in History and Philosophy of Biol Biomed Sci* 44, 4 (2013), 503–514. https://doi.org/10.1016/j.shpsc.2013.03.020

[35] Sabina Leonelli. 2016. *Data-Centric Biology: A Philosophical Study*. University of Chicago Press. https://doi.org/10.7208/chicago/9780226416502.001.0001

[36] David Maier, Vernonika M. Megler, and Kristin Tufte. [n.d.]. Challenges for dataset search. In *International Conference on Database Systems for Advanced Applications*. Springer, 1–15. https://doi.org/10.1007/978-3-319-05810-8_1

[37] Helena M Mentis, Madhu Reddy, and Mary Beth Rosson. 2010. Invisible emotion: information and interaction in an emergency room. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*. 311–320.

[38] Florence Millerand, David Ribes, Karen S. Baker, and Geoffrey C. Bowker. 2013. Making an Issue out of a Standard: Storytelling Practices in a Scientific Community. *Science, Technology, Human Values* 38, 1 (2013), 7–43. https://doi.org/10.1177/0162243912437221

[39] Gerard Oleksik, Natasa Milic-Frayling, and Rachel Jones. 2012. Beyond data sharing: artifact ecology of a collaborative nanophotonics research centre. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*. 1165–1174.

[40] Gary M Olson and Judith S Olson. 2000. Distance matters. *Human–computer interaction* 15, 2-3 (2000), 139–178.

[41] Gary M. Olson, Ann Zimmerman, and Nathan Bos. 2008. *Scientific Collaboration on the Internet.* The MIT Press. https://doi.org/10.7551/mitpress/9780262151207.001.0001

[42] Maureen A. O'malley and Orkun S. Soyer. 2012. The roles of integration in molecular systems biology. *Studies in History and Philosophy of Biol Biomed Sci* 43, 1 (2012), 58–68. https://doi.org/10.1016/j.shpsc.2011.10.006

[43] Trine Pallesen and Peter H. Jacobsen. 2018. Articulation work from the middle—a study of how technicians mediate users and technology. *New Technology, Work and Employment* 33, 2 (2018), 171–186. https://doi.org/10.1111/ntwe.12113

[44] Ted Palys. 2008. Basic research. *The sage encyclopedia of qualitative research methods* 2 (2008), 58–60.

[45] Chrysanthi Papoutsi and Ian Brown. 2015. Privacy as articulation work in HIV health services. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. 339–348.

[46] Irene V Pasquetto, Ashley E Sands, Peter T Darch, and Christine L Borgman. 2016. Open data in scientific settings: From policy to practice. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 1585–1596.

[47] Michael Quinn Patton. 2015. *Qualitative research evaluation methods : integrating theory and practice* (fourth edition. ed.). Thousand Oaks, California : SAGE Publications, Inc., Thousand Oaks, California.

[48] Heather A. Piwowar, Roger S. Day, and Douglas B. Fridsma. 2007. Sharing Detailed Research Data Is Associated with Increased Citation Rate (Sharing Data Citation Rate). *PLoS ONE* 2, 3 (2007), e308. https://doi.org/10.1371/journal.pone.

0000308

[49] Neil Pollock. 2005. When Is a Work-Around? Conflict and Negotiation in Computer Systems Development. *Science, Technology, Human Values* 30, 4 (2005), 496–514. https://doi.org/10.1177/0162243905276501

[50] David P Randall, E Ilana Diamant, and Charlotte P Lee. 2015. Creating sustainable cyberinfrastructures. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 1759–1768.

[51] David Ribes. 2017. Notes on the concept of data interoperability: Cases from an ecology of AIDS research infrastructures. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 1514–1526.

[52] Betsy Rolland and Charlotte P Lee. 2013. Beyond trust and reliability: reusing data in collaborative cancer epidemiology research. In *Proceedings of the 2013 conference on Computer supported cooperative work*. 435–444.

[53] Steve Sawyer and Andrea Tapia. 2006. Always Articulating: Theorizing on Mobile and Wireless Technologies. *The Information Society* 22, 5 (2006), 311–323. https://doi.org/10.1080/01972240600904258

[54] Kjeld Schmidt and Liam Bannon. 1992. Taking CSCW seriously. *Computer Supported Cooperative Work (CSCW)* 1, 1-2 (1992), 7–40. https://doi.org/10.1007/BF00752449

[55] Kjeld Schmidt and Carla Simone. 1996. Coordination mechanisms: Towards a conceptual foundation of CSCW systems design. *Computer Supported Cooperative Work (CSCW)* 5, 2 (1996), 155–200. https://doi.org/10.1007/BF00133655

[56] Dan Sholler, Sara Stoudt, Chris Kennedy, Fernando Hoces de la Guardia, Francois Lanusse, Karthik Ram, Kellie Ottoboni, Marla Stuart, Maryam Vareth, and Nelle Varoquaux. 2019. Resistance to Adoption of Best Practices. (2019).

[57] Susan Leigh Star and Karen Ruhleder. 1996. Steps toward an ecology of infrastructure: Design and access for large information spaces. *Information systems research* 7, 1 (1996), 111–134.

[58] Stephanie B Steinhardt. 2016. Breaking down while building up: design and decline in emerging infrastructures. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 2198–2208.

[59] Stephanie B Steinhardt and Steven J Jackson. 2015. Anticipation work: Cultivating vision in collective practice. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. 443–453.

[60] Allan Stisen, Nervo Verdezoto, Henrik Blunck, Mikkel Baun Kjærgaard, and Kaj Grønbæk. 2016. Accounting for the invisible work of hospital orderlies: Designing for local and global coordination. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. 980–992.

[61] Anselm L. Strauss. 1985. Work and the Division of Labor. *The Sociological Quarterly* 26, 1 (1985), 1–19. https://doi.org/10.1111/j.1533-8525.1985.tb00212.x

[62] Anselm L. Strauss. 1988. The Articulation of Project Work: An Organizational Process. *Sociological Quarterly* 29, 2 (1988), 163–178. https://doi.org/10.1111/j.1533-8525.1988.tb01249.x

[63] Lucy Suchman. 1996. Supporting articulation work. *Computerization and controversy: Value conflicts and social choices* 2 (1996), 407–423.

[64] Katie G Tanaka and Amy Voida. 2016. Legitimacy work: Invisible work in philanthropic crowdfunding. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 4550–4561.

[65] Carol Tenopir, Suzie Allard, Kimberly Douglass, Arsev Umur Aydinoglu, Lei Wu, Eleanor Read, Maribeth Manoff, and Mike Frame. 2011. Data sharing by scientists: practices and perceptions. *PloS one* 6, 6 (2011), e21101.

[66] Theresa Velden, Matthew J Bietz, E Ilana Diamant, James D Herbsleb, James Howison, David Ribes, and Stephanie B Steinhardt. 2014. Sharing, re-use and circulation of resources in cooperative scientific work. In *Proceedings of the companion publication of the 17th ACM conference on Computer supported cooperative work & social computing*. 347–350.

[67] Janet Vertesi and Paul Dourish. 2011. The value of data: considering the context of production in data economies. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work*. 533–542.

[68] Jillian C Wallis, Elizabeth Rolando, and Christine L Borgman. 2013. If we share data, will anyone use them? Data sharing and reuse in the long tail of science and technology. *PloS one* 8, 7 (2013).

[69] Robert Stuart Weiss. 1994. *Learning from strangers : the art and method of qualitative interview studies*. New York : Free Press ; Toronto : Maxwell Macmillan Canada ; New York : Maxwell Macmillan International, New York : Toronto : New York.

[70] Michael C Whitlock, Mark A McPeek, Mark D Rausher, Loren Rieseberg, and Allen J Moore. 2010. Data Archiving. *The American Naturalist* 175, 2 (2010), 145–146. https://doi.org/10.1086/650340

[71] R Williams, G Pryor, A Bruce, S Macdonald, W Marsden, J Calvert, and C Neilson. 2009. Patterns of information use and exchange: Case studies of researchers in the life sciences Research Information Network Report. University of Edinburgh Digital Curation Centre.

[72] William A. Wulf. 1993. The collaboratory opportunity. (National Research Council report 'National Collaboratories: Applying Information Technology for Scientific Research') (Computing in Science) (Cover Story). *Science* 261, 5123 (1993), 854. https://doi.org/10.1126/science.8346438

[73] Alyson L Young and Wayne G Lutters. 2015. (Re) defining Land Change Science through Synthetic Research Practices. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. 431–442.

[74] Ann Zimmerman. 2007. Not by metadata alone: the use of diverse forms of knowledge to locate data for reuse. *International Journal on Digital Libraries* 7, 1 (2007), 5–16. https://doi.org/10.1007/s00799-007-0015-8