From Sea to Servers: Temporalities of Data Management and the Limits of Availability in Oceanography

Yubing Tian

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington
2025

Reading Committee:
Megan Finn, Co-Chair
Jaime Snyder, Co-Chair
Ricardo Gomez
Charlotte Lee

Program Authorized to Offer Degree:
Information School

University of Washington

Abstract

From Sea to Servers: Temporalities of Data Management and the Limits of Availability in Oceanography

Yubing Tian

Co-Chairs of the Supervisory Committee:
Megan Finn
Information School
Jaime Snyder
Information School

A persistent challenge across scientific fields is determining what research data to keep, why, how, and for how long. This dissertation examines how research data is managed and shared in oceanography, focusing on the impact of data policies introduced by the United States (U.S.) National Science Foundation (NSF). Since 2011, the NSF has required data management plans (DMPs) as part of grant proposals with the aim of making research data available over time. While DMPs may be a relatively recent requirement, NSF data sharing policies for oceanographers can be traced back to the World Ocean Circulation Experiment (WOCE), 1990 - 2002. Oceanographers participating in WOCE were required to make their data available two years after collection. A requirement that at the time, was noted as a departure from traditional research practices. Nevertheless, this time norm for data sharing has endured, as it can be found in early NSF Division of Ocean Sciences (OCE) data policies from the 1990s and remains in OCE data policies today. And yet, oceanographers still report difficulties managing and sharing research data. This study explores what happens to data, researchers, and infrastructures under the requirements mandated in federal data policies. Through interviews and document analysis, this dissertation foregrounds the temporal dimensions of managing and making data available

and makes the following three contributions. First, this study shows that data do not all follow the same lifecycle as understood in prescriptive research data lifecycle models. Moreover, while data policy imagines universal availability, in practice, data persists unevenly over time. From DMPs, I identify three forms of planned data afterlives: secluded, splintered, and speculative, to describe how material conditions, disciplinary norms, and institutional arrangements shape what data endure, and how. Second, I explore the ways that researchers struggle to meet policy expectations for data availability. Researchers describe data management as tedious, time-consuming, and hard to prioritize, even for those who understand its importance. Following Elizabeth Shove's argument that practices have distinct temporal characteristics, I argue that data management and sharing lack an established "practice-time profile". This absence leads to end-loaded, last-minute efforts during the project's sunset phases, which can create minor to substantial delays in data sharing. At the same time, some researchers are developing practice-time profiles to better manage their data in preparation for sharing and preservation. Third, I introduce the concept of temporal paradox to describe how data infrastructures built for long-term access are often marked by short-term fragility. Building on Marisa Cohn's "convivial decay," I describe how researchers and data managers preemptively anticipate infrastructural demise, not only in aging systems, but also in relatively new ones. I articulate how this practice of planning for the end, paradoxically, supports the long-term persistence of data. Together, these findings contribute to information science, STS, and infrastructure studies through an empirical account of the temporal dynamics between data practices, infrastructures, and policy in the short-term availability and long-term stewardship of scientific data.

TABLE OF CONTENTS

Chapter 1. Introduction 11
1.1 The emergence of data management and sharing policies in the 21st century 11
1.2 The 2011 NSF DMP Mandate 14
1.3 Why Oceanography and Ocean Data? 17
1.4 Research Questions 19
1.5 Contributions 20
Chapter 2: Time and Temporality in Data, Practice, and Infrastructure 21
2.1 Introduction 21
2.2 Conceptualizations of Time 21
2.2.1 Objective understanding of Time 22
2.2.2 Subjective understandings of Time 22
2.2.3 Chaotic Time and the topology of times 25
2.3 The Temporalities in Data Lifecycles 27
2.3.1 Prescriptive research data lifecycle models 27
2.3.2 Researcher-centered data lifecycle models: Career timelines and affective entanglements 31
2.3.3 Epistemic times of data: Data time, phenomena time, and scientist time 32
2.4 Time and data practices 35
2.4.1 Time and temporality in scientific, information, and organizational practice 35
2.4.2 Retrospective and Anticipatory temporalities in scientific work 37
2.5 Infrastructures and Time 40
2.5.1 Timescales in Infrastructure development 40
2.5.2 Continuities and Discontinuities of Scientific Cyberinfrastructures 42
2.6 Conclusion 44
Chapter 3. Methodology and Study Design 45
3.1 Introduction 45
3.2 Grounded Theory 45
3.3 Research Methods Used to Study DMPs 46
3.4 Study Design 49
3.4.1 Data Generation 50
3.4.2 Data management plan corpus construction 50
3.4.3 Semi-structured interviews 52
3.4.4 Data Analysis 57

3.4.5 Document Analysis 57
3.4.6 Inductive Qualitative Analysis 59
3.5 Study limitations 60
3.6 Researcher's stance and reflexivity 61
3.7 Conclusion 62
Chapter 4. Tracing the origins of OCE's data-sharing time norm 63
4.1 Introduction 63
4.2 Global-scale ocean studies come of age: The World Ocean Circulation Experiment 64
4.3 The origins of a time norm: data management and data sharing in WOCE 72
4.3.1 Envisioning and Implementing the WOCE Data Management System 74
4.3.2 The evolution of WOCE's data sharing policy: Establishing the Two-Year Data Sharing Norm 80
4.3.3 Data sharing policies in the wild 81
4.4 The persistence of WOCE's data sharing time norm in OCE data policies 85
4.4.1 Pre-2010 DMP mandate OCE data policies 88
4.4.2 Post-2010 DMP mandate OCE data policies 89
4.5 Conclusion 90
Chapter 5. Planning for Heterochronous Data Afterlives 91
5.1 Introduction 91
5.2 Exhaustive Planned Hindsight and the limits of research data availability 92
5.3 Secluded Futures of Physical Samples 93
5.3.1 Uneven institutional support for physical samples 94
5.3.2 PI-driven data stewardship 96
5.3.3 Material Decay 98
5.4 Splintered Futures of Models 99
5.4.1 From large to colossal data scales 100
5.4.2 Fragmented Access and Accessibility of Model Outputs 102
5.4.3 Transient Epistemic Utility: Limited lifetimes of model outputs 103
5.4.4 Frozen code, Fragmented Models 104
5.4.5 The promise and perils of cloud computing infrastructures 107
5.5 Speculative Futures of "Uncommon" Data 108
5.5.1 Improvised Access 109
5.5.2 Community over Compliance 110
5.5.3 Stabilizing Speculative Futures 111
5.6 Conclusion 112

Chapter 6. Practice-time profiles of data management and sharing 113
6.1 Introduction 113
6.2 The Missing Practice-Time Profiles of Research Data Management 115
6.3 End-Loaded Practice-Time Profiles of Data Management 115
6.3.1 Key personnel and the dynamics of data availability and delay 116
6.3.2 “We like to first publish and then share” 119
6.4 Making time at “Right Time” for Data Management and Sharing 120
6.5 Towards Integrated Practice-Time Profiles for data management 122
6.5.1 Integrating Data Management into the Intermediary Steps of Research 122
6.5.2 Redistributing Data Management Work Through Automation 124
6.5.3 Anticipating Future Reuse in the Present 125
6.6 Conclusion 126
Chapter 7. Temporal paradoxes in and of data infrastructures 126
7.1 Introduction 126
7.2 Paradoxical Rhythms of Infrastructural Support 128
7.2.1 Alignments and Misalignments in PI and Repository Collaborative Rhythms 128
7.2.2 Facilitation and Drag 130
7.2.3 Aligned infrastructural support: The BCO-DMO - GEOTRACES partnership 132
7.3 Anticipatory convivial decay and the paradox of the long now 134
7.3.1 The uncertain long tail of public data access 134
7.3.2 Anticipating Infrastructural demise 135
7.4 The Curious Case of Recurring Reverse Salients 139
7.4.1 “That’s a niche thing”: how “newness” reintroduces reverse salients 141
7.5 Conclusion 144
Chapter 8. Conclusion 145
8.1 Primary Findings: Temporalities of data management 145
8.1.1 Heterochronous data and their afterlives 145
8.1.2 Practice-time profiles of research data management 146
8.1.3 Temporal Paradoxes of Data Infrastructures 147
8.2 Future Research Directions 149
8.3 From Exhaustive towards Selective Planned Hindsight 150
8.4 Policy implications 151
8.5 Conclusion 152
Bibliography 154

Appendix A. Interview Protocol For 2022 Interviews 162
Appendix B. Interview Workflow Checklist 164
Appendix C. Interview Protocol for 2025 Interviews with PIs, graduate students, and research scientists 165
Appendix D. Interview Protocol for 2025 Interviews with data managers 167
Appendix E. Codebook Used For Second Cycle Coding 169

TABLE OF FIGURES

Figure 1. Examples of research data lifecycle models 28
Figure 2. From Treloar & Harboe-Ree (2008) The data curation continua model 30
Figure 3. From Stahlman (2022) researcher-centered data lifecycle model 32
Figure 4. From Wylie (2024) Scientist Time model 34
Figure 5. Number of OCE DMPs by year the project was funded 52
Figure 6. From WCRP (1986) WCRP's major scientific objectives 66
Figure 7. From U.S. WOCE Implementation Plan (1988) WOCE's organizational structure 68
Figure 8. From Woods (1985) showing how existing oceanographic datasets were not suitable for WOCE 69
Figure 9. From WCRP (1991) showing locations of planned WOCE research cruises 72
Figure 10. From WOCE International Project Office (1997) Overview of WOCE Data Set 73
Figure 11. From U.S. WOCE Implementation Plan (1988) Cost Estimates Summary 75
Figure 12. From WMO (1986) Diagram of planned WOCE data system 77
Figure 13. From Lindstrom & Legler (2001) Location and institutions in WOCE's data management and sharing system 79
Figure 14. From WHPO (1991) Showing time for data sharing after collection 83

TABLE OF TABLES

Table 1. Eight irreducible elements of a timescape, their definitions, and examples 23
Table 2. Interviewees by research discipline 53
Table 3. Interviewees by position held 54
Table 4. Key phases of WOCE and their corresponding years 64
Table 5. Time in WOCE and OCE Data Policies 86

ACKNOWLEDGEMENTS

As the saying goes, every dream needs a team. I would like to thank the people who helped make this research and my PhD journey possible. This dissertation is not only the product of my own efforts but also the result of the encouragement, generosity, and insights of many people who shaped my thinking and supported me intellectually, financially, and emotionally through the ups and downs of this six-year-long adventure.

First, I would like to express my deepest gratitude to my co-chairs, Megan Finn and Jaime Snyder. Megan, thank you for your steadfast mentorship, incisive feedback, and for always encouraging me to clarify and sharpen my ideas. Your enthusiasm for scintillating infrastructure studies and STS research is infectious. You are the type of advisor that Reddit would consider a unicorn. Rigorous, generous, so widely read and knowledgeable, yet still willing to do a deep dive into the data. Jaime, thank you for always seeing the big picture and for coming up with creative analogies to help me see it, too. I don't think I'll ever forget "the chain of logic" or the throughline when writing. Thank you for trying to make explicit the hidden rules and norms of the different aspects of scholarly work. I feel incredibly fortunate to have had both of you guiding this work.

To my committee members, Ricardo Gomez, Charlotte Lee, and Sarah Quinn. Ricardo, thank you for your encouragement throughout the stages of this process. I still remember in our first meeting how you congratulated me on getting into the program, and you then said that the hardest part is actually getting out. I didn't get it at the time, but I know that to be true deep in my bones. Thank you for giving me candid advice and feedback, whether it be about my research, career, or how I should just drive more to get over my then-fear of driving. Charlotte and Sarah, thank you for graciously agreeing to join my committee on such short notice. Despite this, I am grateful for your encouragement leading up to the defense and generative comments during my defense. I feel incredibly lucky to have worked with a committee that challenged and pushed me and was genuinely supportive.

This research first began to take shape thanks to the Data Afterlives project. Thank you to Amelia Acker and Sarika Sharma for being thoughtful collaborators and generous thinkers. Thank you to Thomas Struett for kindly helping me with intercoder agreement for my codebook, which jump-started the final stretch of my dissertation work.

To the research participants who generously responded to the email survey campaign, who shared their time, experiences, and insights with me: thank you. This project could not have happened without you. I learned a great deal from the care you bring to your work and the complexity with which you understand your own data practices I am especially grateful for the trust you placed in me to handle your accounts with care. I hope this dissertation does justice to what you generously shared with me.

To the friends and colleagues who kept me going throughout the years. Thank you to Shawon Sarkar and Lan Thi Nguyen for your accountability and friendship. Will Sutherland and Justin Petelka, thank you for consistently being there to talk shop, sharing readings, reflecting on theory, methods, or the weirdness of research and doing a PhD. During the times when I had no gas left in the tank, our co-working sessions reminded me that showing up and just starting was often the hardest part.

Although it didn't end up materializing, I want to thank Andrew Nguyden, as well as Charlotte Lee (again), for introducing me to one of the research groups in his field site. Your kindness and effort to help mean a lot to me.

Thank you to Shiring Madon, my Master's advisor, who planted the seed of this many-years-long adventure. You believed that I could do this even before I fully believed it myself.

Finally, thank you to my family and friends for holding me up and putting up with me. Thank you to my parents, Tian Ya and Xiang Zhailu, and brother Yizhou Tian, and in-laws Joytsna and Dinesh Zope who always believed in me. To my wonderful husband, Manu Zope, thank you for being steady as a rock. You never once doubted that I could finish this, even when I did. Thank you for creating space for me to do this work, and for making sure that our life remained resonant, full of adventure, nature, and good food. I couldn't have done this without you.

Chapter 1. Introduction

1.1 The emergence of data management and sharing policies in the 21st century

One enduring challenge common across all scientific fields is that of managing research data and ensuring that access to that data is possible in the future. Well managed research data ensures that that data is reusable outside of the project that produced it. One hypothesized way to ensure that research data is well managed and accessible in the future is to plan for it at the outset (National Science Board, 2005). Since the 2010s, funding agencies in the United States (U.S.), such as the National Science Foundation (NSF), have required that grantees author data management plans (DMPs) as a way of assuring that research data is available and accessible beyond the lifetime of their research projects. However, we have yet to understand what the impacts of the NSF's broader science data policy are and how, if at all, the data management and sharing mandate has affected researchers' data practices and workflows since it has been in effect.

To better understand how this science data policy came to be in the first place, I begin with an overview of the discussions around open data in the 2000s that have shaped the introduction of this mandate. Since the turn of the 21st century, open access to data has grabbed the attention of policymakers, scientists, activists, and members of the public. In the U.S. open access to data was at first primarily concerned with government data (Chignard, 2013; Data.gov admins, 2013). A popular definition of what the open part of open data denotes is provided by the Open Knowledge Foundation, "open data and content [that] can be freely used, modified, and shared by anyone for any purpose" (Open Knowledge Foundation, n.d.). The key elements of this definition are that open data should be free of charge and that once shared can be reused. During the Obama presidency, open access to government data would become the heart of open government initiatives. 2009 presidential memoranda titled "*Transparency and Open Government,"* laid out the benefits for open government initiatives. The memoranda argues that open government data enables many benefits, first timely information about government activities would improve citizen's trust in the government through oversight. Second, better decisions would be made through the public's participation in governmental affairs. Third, collaboration would ensue between government agencies but also across sectors with industry

professionals, businesses, and nonprofits (*Memorandum on Transparency and Open Government*, 2009).

Around the same time when open access to government data was a prominent issue, so too was that of open access to research data. Proliferation of data resulted in many disciplines being labelled "data intensive." The prefix "data intensive" goes beyond merely describing the volume, quantity, or complexity of data that is worked with, but marks the beginning of a new form of empiricism in scientific inquiry. Also labelled the fourth paradigm of science, data empiricism dethrones deductive hypothesis-driven inquiry in favor of an approach that blends inductive, deductive, and abductive approaches to investigate a phenomenon (Hey et al., 2009; Kitchin, 2014a). Rather than formulating hypotheses first and then collecting data to confirm or refute hypotheses, data empiricism envisions a method where scientists and researchers could now begin with data first and generate new hypotheses from that data. Although, the roots of open science and data empiricism began separately, they appear combined in science data policies. For the promise of data-intensive sciences to be realized, it is hypothesized that a necessary precondition is the public availability of vast quantities of data.

Making one's research data available is a hallmark practice of open science (Willinsky, 2005). Positive scientific, economic, and political outcomes have been ascribed to open access to research data. For scientists, open access to their research data is thought to promote good scientific practices, bolster their professional career through data citations and support data-intensive and interdisciplinary research agendas, reduce duplicate data collection efforts and promote equity (Arzberger et al., 2004; Mosconi et al., 2019). It is thought that economic benefits will be achieved through innovation and increased competitiveness. Politically, open access to research data will render visible the impact of federal investment in research (OECD, 2007).

Open science and open data initiatives are not without its critics. Open data movements did not emerge out of a neutral terrain. Writing about the political and economic ideology underlying open movements, Mirowski (2018) argues that open science is the continuation of neoliberalism in academic research under a new guise, that of platform capitalism. The goal of open science, Mirowski writes, "seeks to maximize data revelation as a means to eventual monetization" (2018, p. 193). Arguing along the same lines, but written less harshly, Levin and Leonelli (2017) argue that the emphasis on access, dissemination, and reuse of data is made

dependent on the viability of forms of commercial capture. In practice, "openness" in biology, is highly contingent on the research practices at the laboratory level, the research infrastructures and resources available, and professional beliefs. Kitchin (2014b), writing earlier and more focused more broadly on open government data, questions the sustainability of open data initiatives. Attention has been focused on the supply of open data with little thought to their long-term funding and sustainability.

One hypothesized way to ensure that research outputs become publicly available to wider audiences and for long timeframes is to plan for that outcome at the outset. Following this rationale, funding agencies in the U.S., and internationally, have made changes to their grant application process following this rationale. Initial investigations suggest that researchers perceive authoring DMP as a purely administrative task instead of a research task (Carlson, 2017; Miksa et al., 2019). Meanwhile, some scholars question the premise of DMPs as a science data policy, arguing that DMPs have no evidence for their effectiveness (Smale et al., 2020).

Other scholars do not question the premise of the DMP data policy but note the limitations of its current implementation, as a static document, and lack of support for researchers. Several studies have looked at the readiness or preparedness of researchers to meet the requirements mandated by the policy, only to find that researchers do not understand the important of what the policy aims to achieve, do not know how best to comply with the mandate, or because the policy is not strictly enforced (Tedersoo et al., 2021).

More recently, major updates were made to federal policy that doubled down on its quest to ensuring that federally funded research results in open research outputs. A memorandum titled *"Ensuring free, immediate, and equitable access to federally funded research"* authored by the White House Office of Science and Technology Policy (OSTP, 2022) seeks to address what it views as one of the major limitations of the 2013 *"Increasing Access to the Results of Federally Funded Research"* memorandum – optional 12-month embargo on publications and supporting data from federally funded research. The memorandum urges federal agencies to update their public access policies accordingly. Framing the embargo as inequitable: "Financial means and privileged access must never be the pre-requisites to realizing the benefits of federally funded research that the American public deserves." (White House Office of Science and Technology Policy (OSTP), 2022b)

Moreover, Sec. 10344 of the 2022 Chips and Science Act required that:

> *"The NSF shall facilitate public access to research products, including data, software, and code, developed as part of NSF projects. The NSF shall require that every proposal for funding for research include a data management plan that includes a description of how the awardee will archive and preserve public access to data, software, and code developed as part of the proposed project.*
>
> *The NSF shall develop and disseminate a set of criteria for trusted open repositories to be used by NSF-funded researchers and make awards for the development and maintenance of such repositories.*
>
> *The NSF shall support research and development of tools and infrastructure that support reproducibility and support the education and training of researchers on computational methods, tools, and techniques to improve the quality and sharing of data, code, and supporting metadata to produce reproducible research."* (CHIPS and Science Act, 2022).

The remainder of the introduction provides an overview of the NSF's DMP mandate, provides a site justification for oceanography and ocean data, this study's research questions, aims, and contributions.

1.2 The 2011 NSF DMP Mandate

Beginning January 18, 2011, all researchers who sought federal funding for projects were mandated to include a DMP. The DMP is included as a supplement to research proposals, if the DMP is missing from the it is not accepted by NSF's grant submission platform FastLane (National Science Foundation, 2015).

The origins of the NSF's DMP mandate can be traced back to a 2005 National Science Board (NSB) authored report titled "*Long-lived Digital Data Collections*". The NSB is an advisory board to the NSF and is responsible for approving NSF policies and ensuring that they align with U.S. national policies. The NSB urges in recommendation number four of its report that "the NSF should require that research proposals for activities that will generate digital data, especially long-lived data, should state such intentions in the proposal so that peer reviewers can evaluate a proposed data management plan" (National Science Board, 2005, p. 12).

A couple of years later, a 2009 report titled *"Harnessing the Power of Digital Data for Science and Society"* authored by the Interagency Working Group on Digital Data to the Committee on Science of the National Science and Technology Council recommended the need for a process to encourage the management of research data writing that agencies should

"promote a data management planning process for projects that generate preservation data" (Interagency Working Group on Digital Data, 2009, pp. 10–11). Together, the recommendation of these two reports establishes data management, whether it be as a planning process or a plan submitted for peer-review, as a means of ensuring that valuable data created in the present can be used in the future.

Although there were no official policy goals articulated for the NSF DMP mandate upon its arrival, a press release (National Science Foundation, 2010) by the NSF provides insights into what exactly this mandate seeks to achieve. And much was to be achieved with it. Requiring that grantees write DMPs would address and promote:

1) the needs of data-driven science as "Digital data are both the products of research and the foundation for new scientific insights and discoveries that drive innovation",
2) more effective communication for researchers,
3) more effective collaboration for researchers,
4) comply with President Obama's Open Government Directive

The press release also promised that the DMP mandate was just the first step in a comprehensive data policy (National Science Foundation, 2010). A DMP is a document that describes what research data will be collected or produced during the project, how those data will be analyzed, stored, made accessible and archived. Broad guidance, found in the Grant Proposal Guide: Chapter II.C.2.j, (National Science Foundation, 2011) for the content of a DMP include:

- The types of data, samples, physical collections, software, curriculum materials, and other materials to be produced in the course of the project.
- The standards to be used for data and metadata format and content (where existing standards are absent or deemed inadequate, this should be documented along with any proposed solutions or remedies);
- Policies for access and sharing, including provisions for appropriate protection of privacy, confidentiality, security, intellectual property, or other rights or requirements.
- Policies and provisions for re-use, re-distribution, and the production of derivatives; and
- Plans for archiving data, samples, and other research products, and for preservation of access to them

Beyond very broad guidelines, the NSF relegates more detailed guidance around DMPs to its organizational subunits, as what is considered research data, and subsequently researcher's data practices, all vary by the Principal Investigator's (PI's) research area. For example, the Directorate of Geosciences includes four research areas (Suchman, 2022).[1] Each research area, one of which is the Division of Ocean Sciences (OCE), has their own data policy that interprets the DMP mandate for its grantees.

My dissertation will show that, more than a decade after the introduction of the DMP mandate, there is a clearer understanding of what a DMP *is*. First, as a genre of scientific document, it is 2 pages or less and includes anticipated or planned actions on, with, and around research data. These plans should ideally describe how research data will be managed whilst the research project is ongoing and how that data will be available and preserved beyond the conclusion of the project. Interestingly, projects that will not produce data are still required to submit one. Second, DMPs are occluded documents, meaning that these documents are typically not publicly available and are usually bundled together with the grant proposal. The private nature of these documents is likely a part of the reason why empirical studies of DMPs is a young field.

However, we have yet to understand what it is that DMPs *do*. By extension, we have yet to understand what the impacts of the NSF's broader science data policy are and how, if at all, the DMP mandate has affected researchers' data practices and workflows since it has been in effect.

As a first step towards addressing these questions, this dissertation will examine how the DMP mandate has affected the data practices and workflows of oceanographers. A temporal perspective might provide a path forward as in our preliminary work analyzing DMP guidance documents has identified different timeframes PIs needed to plan for (Tian et al., 2021). Different guidance documents articulated different goals attached to these timeframes which have implications for the resources needed for PIs to manage their research data. For example, some 2011 policies had no discussion of different timeframes for data management and the main goal

[1] Suchman here refers to Cynthia Suchman NSF Program Director of Biological Oceanography in the Division of Ocean Sciences (OCE) and not scholar Lucy Suchman. The reference in question is a presentation slide deck that reviews program areas under the Directorate of Geosciences which includes an overview of OCE slides 9-10. Based on the content, the presentation appears to be geared towards early career researchers about how to apply for NSF funding.

articulated was that managed research data support published results. In contrast, policies that explicitly articulate data reuse as a goal also imposed timeframes within which this goal was to be achieved. Whereas the timeframe in some policies was nebulous, e.g. "reasonable" or "timely" timeframe. Policies from OCE specify that this should be done "no later than two (2) years after the data are collected." The next sections concern why it is important to examine oceanographic data; this study's research questions and its contributions.

1.3 Why Oceanography and Ocean Data?

The importance surrounding present day oceanographic research is undoubtedly its role in determining human impacts on pressing global socioecological issues such as climate change, pollution, and biodiversity loss. It is hard to overstate the gravity of our present predicament as options for course reversal, or even improvement, diminish the longer inaction persists.

Oceanic processes play a critical component in Earth's major systems including but not limited to weather, climate, carbon, and geophysical processes such as erosion in coastal areas (Ocean Literacy Network, 2013). Equally important, ocean health is human health, meaning that the ocean and human society are inextricably interconnected (National Oceanic and Atmospheric Administration, 2023). The ocean is home to a staggering abundance of flora, fauna, and other critters large and small, is a source of natural resources such as livestock and oil, is an inexpensive mode for hauling cargo across vast distances, and a major location of recreation and relaxation that many island and coastal communities' economies revolve around. Ocean research and data is needed to make important policy decisions surrounding social issues, of which climate change is one of them. Experts at the 10th annual World Ocean Summit held in 2023 in Lisbon, Portugal surfaced "accessible [ocean] data" as a necessary precondition for decision-making (Economist Impact, 2023). Yet ocean data and the infrastructures supporting its circulation and preservation may directly contribute to negative environmental outcomes (Pendleton & Sorensen, 2021).

The quest to better understand the oceans has a long history. For centuries, seafarers, naturalists and researchers have endeavored to study oceanic processes and to probe the depths of the ocean (Edwards, 2010; Woods Hole Oceanographic Institution, n.d.). The earth's ocean covers more than seventy percent of its surface. This physical vastness has historically posed serious challenges to observing and recording natural processes occurring at sea and below the sea's subsurface (Adler, 2019; National Oceanic and Atmospheric Administration, 2021). Present

day popular media coverage of ocean research declares that researchers, and by extension the general public, know more about outer space than the ocean. An estimated eighty percent of the ocean remains unexplored, leading some to tout it as one of humanity's final unexplored frontiers. Oceanography is one of the youngest fields of science. It only became a more established discipline in the late 19th century after several ocean-going expeditions were completed by the American and European nations. Moreover, contemporary oceanography is only about 60 years old and can trace its roots back to World War II and the Cold War in the mid and late 20th century when understanding the ocean became paramount in the U.S.'s ability to conduct submarine warfare (National Research Council, 2000). Global oceanography, where oceanography studies the ocean as a global phenomenon, is an even younger field as I will discuss in detail in Chapter 4.

In the last couple of decades, calls for investments into oceanographic research have manifested in long term cyberinfrastructure projects. One prominent example is that of the NSF-funded Ocean Observatories Initiative whose vision is to collect longitudinal data from its more than 800 instruments for at least 25 years (Ocean Observatories Initiative Facility Board, 2021). These expensive investments reflect a shift in some of the core research practices of oceanographers: from being ship-based to sensor-based (Lehman, 2018). Although not all types of oceanographers have, or can, fully embrace this shift in work practices, but for those who can, such as physical oceanographers, remote sensing and satellite data have been heralded with bringing big data and, along with it, data science into the ocean sciences (Conway, 2006; Qian et al., 2022).

Scholars in information science and science and technology studies (STS) have examined how oceanographic knowledge production is shaped by researcher's data practices, the collaborative work required to integrate diverse ocean datasets, the development of cyberinfrastructure, and funding (Darch & Borgman, 2014; Halfmann, 2018; Neang et al., 2021; Steinhardt & Jackson, 2014b). It remains unclear how federal science data policies have affected oceanographer's data practices.

Existing research on the effect of science data policies on research practice suggest that data policy requirements have little effect on researcher's self-reported data sharing practices in the short-term, defined as over 2 years (Priego & Wareham, 2020). Levin et al., (2016) investigate the relationship between open science policies in the United Kingdom (UK) and

research practice of biomedical researchers found that openness was understood to mean different things and associated with different practices around research data. Based on their interview data, Levin et al., suggest that blanket open science policies, policies that are one-size-fits-all, "is not always warranted or useful" (Levin et al., 2016, p. 137). Data policies may backfire and not achieve the aims it sets of to do if "they force scientists to disclose results and resources in ways that they deem useless or inappropriate, or by requiring openness at a stage of research where it is more likely to hamper than encourage progress. Moreover, open science policies in practice create an additional "administrative burden" on UK biomedical researchers, to which Levin et al., suggest that UK funding agencies invest more toward supporting infrastructures and personnel in data management and curation services.

Oceanography and ocean data are the focus of this study for many reasons. Contemporary ocean data are argued to play a major role in decisions made about key societal problems such as climate change and biodiversity loss. At the NSF, the Division of Ocean Science supports the U.S. Global Change Research Program (Suchman, 2022). "Sound data management practices" are cited as paramount in making ocean data accessible and returning value to federal investment in ocean research (National Research Council, 2011). At the NSF alone, approximately 350 million U.S. dollars a year is spent on ocean science research (Gonzales, 2020). The total amount allocated has remained the same, but in recent years funding for ocean data infrastructures has surpassed that for "core science" causing discontent amongst some scholars in the oceanography community (Kintisch, 2015; Witze, 2015). At the same time, generating large volumes of ocean data can be environmentally damaging.

1.4 Research Questions

The aim of this dissertation is to explore the ways that the NSF's data management policy mandate has affected oceanographer's data practices. Specifically, I examine what happens to data, researchers, and infrastructures under the requirements mandated in federal data policies.

The research questions (RQs) that this dissertation explores are the following:

- **RQ1** In what ways do the timeframes specified in data policies shape the management, sharing and afterlives of research data?
- **RQ2** In what ways do oceanographers navigate the timeframes specified in data policies in their data management and sharing practices?

- **RQ3** In what ways do data policy requirements to deposit in designated repositories, shape the capacity of infrastructures to support the short and long-term availability of OCE-funded data?

Together, my research questions aim to 1) describe the relationship between data policy mandates and data management and sharing practices in oceanography 2) highlight the divergence between policy mandates and empirical cases 3) surface the temporal dimensions that shape the data afterlives, practices and infrastructure.

1.5 Contributions

Through interviews and document analysis, this dissertation foregrounds the temporal dimensions of managing and making data available and makes the following three theoretical contributions.

First, this study shows that data do not all follow the same lifecycle as understood in prescriptive research data lifecycle models. Moreover, while data policy imagines universal availability, in practice, data persists unevenly over time. From DMPs, I identify three forms of planned data afterlives: secluded, splintered, and speculative, to describe how material conditions, disciplinary norms, and institutional arrangements shape what data endure, and how.

Second, I explore the ways that researchers struggle to meet policy expectations for data availability. Researchers describe data management as tedious, time-consuming, and hard to prioritize, even for those who understand its importance. Following Elizabeth Shove's argument that practices have distinct temporal characteristics; I argue that data management and sharing lack an established "practice-time profile". This absence leads to end-loaded, last-minute efforts during the project's sunset phases, which can create minor to substantial delays in data sharing. At the same time, some researchers are developing practice-time profiles to better manage their data in preparation for sharing and preservation.

Third, I introduce the concept of temporal paradox to describe how data infrastructures built for long-term access are often marked by short-term fragility. Building on Marisa Cohn's "convivial decay," I describe how researchers and data managers preemptively anticipate infrastructural demise, not only in aging systems, but also in relatively new ones. I articulate how this practice of planning for the end, paradoxically, supports the long-term persistence of data. Together, these findings contribute to information science, STS, and infrastructure studies

through an empirical account of the temporal dynamics between data practices, infrastructures, and policy in the short-term availability and long-term stewardship of scientific data.

In each of my empirical chapters (Chapters 5 – 7) I note the particular ways that the present NSF data sharing policy is universalizing. First, it imagines all data as universally preservable and shareable. Second, it imagines the labor available to make data accessible even when this is not at present adequately recognized. Third, it imagines that data infrastructures will be able to maintain access to research data for long time periods.

Given an understanding that the universal data availability imagined in the policy, this study suggests that data sharing follow principles of "selectivity." Selectivity acknowledges that not all research data can receive or merit equal stewardship. Similarly, resources, time and labor are all constrained. As such, instead of casting a wide net over every possible research data, this study recommends the following principles for embedding "selectivity" in data management and stewardship: 1) material viability, 2) epistemic prioritization, 3) reuse potential, and 4) considering data reduction.

As for operationalizing selectivity in for granting agencies, in the Conclusion I discuss further some possible ways for how this can be integrated into the current research funding process with changes to DMP guidance, content, review, and funding mechanisms.

Chapter 2: Time and Temporality in Data, Practice, and Infrastructure

2.1 Introduction

How time shapes data, scientific practice, and infrastructure has been of keen interest to scholars in various disciplines, ranging from information science, STS, to critical data studies and more. This is unsurprising, as time plays a central role in open science, data sharing, preservation, and information infrastructures.

2.2 Conceptualizations of Time

This section introduces three foundational perspectives for understanding time, objective, subjective, and chaotic, to establish how time can be understood as either singular and measurable or as situated, complex, and relational. The subjective and chaotic understandings of

time provide essential grounding for how this dissertation analyzed time as it operates in policy, data, practice and infrastructure.

2.2.1 Objective understanding of Time

One way to understand time is through its mathematical formulation. Conceiving of time solely based on mathematical units is what scholars call "objective perspective [of] time" (Reddy et al., 2006, p. 34). Note that other terms have been used such as clock time, chronos, etc. (Adam, 1998; Kitchin, 2023) The specific units in question are about the duration of time in seconds, minutes, hours, days, and years. Time in its objective form is measurable, quantified, mechanistic, standardized, and linear (Adam, 1998; Kern, 2000). Scholars point towards Newtonian physics, and the time-keeping and accounting tools and technologies, such as in clocks and calendars, as the "recognizable infrastructures of objective time" (Shove, 2009, p.19 in Haider et al., 2022, p. 9)

A point that is often implicit in discussions of the differences between objective versus subjective time is that the former is characterized by what Adam's calls a "de-temporalized" nature. What Adam's (1998) means by time as de-temporalized is that it is flattened, abstracted to serve the purposes of measuring and calculating, but only has a tenuous relationship with the phenomena it presents and represents:

> "*Newtonian science recognizes no contextually based differences in rhythm and intensity, no contextual tempo or timing, duration or change, no times inherent in processes and phenomena … no right time for every season and place, no special days and moments … no stress and pressure of 'deadlines', no valorization of speed, no reverence for the past, no hopes for the future. Instead, time in Newtonian science … is an atemporal time, a time unaffected by the transformations it ascribes*" (Adam, 1998, p. 40)

2.2.2 Subjective understandings of Time

An alternative way to understand time is through a social constructionist perspective. From this perspective, time is contextual and constructed through norms, beliefs, and customs. The term "subjective time" is often used from this standpoint. Other common terms embodying this perspective include kairos and instantaneous time. Given that this perspective maintains that time is not one singular entity, scholars have named different kinds of "subjective time." As it relates to scientific work, Traweek for example identifies career time, beamtime, and downtime in her study of high-energy particle physicists (Traweek, 2009).

Perhaps articulated most succinctly by Reddy et al., the difference between objective and subjective time is that the former conceives of time as separate from human actors, while the latter understands time as inseparable from human actors "a product of norms, beliefs, and customs of individuals" (2006, p. 34). Those of us interested in studying time and temporality are then left in a bit of a bind. Information science, STS, Computer-Supported Cooperative Work (CSCW), and humanist disciplines align with the subjective conceptualization of time.

Given that the subjective time perspective has shown that time is neither singular nor universal but instead is contextual and complex. British sociologist Barbara Adam extends the subjective standpoint of time through her "timescapes" theoretical framework. In a timescape, the time in question is "complex and multiplex, involving many different features and dimensions" (Adam, 2008, p. 1). The dimensions in question are eight "irreducible elements, the combination of which … [constitute a] 'timescape'," (Adam, 2008, p. 1) see Table 1. The "scape" part of the term is an acknowledgement that "time is inseparable from space and matter, and second, that context matters (Adam, 2008, p. 2)."

Table 1. Eight irreducible elements of a timescape, their definitions, and examples

Elements of a timescape (Adam, 2008; Kitchin, 2023)	**Definitions (Adam, 2008)**	**Examples (Burgess, 2010; Kitchin, 2023)**
Time frames	Objective frame of time in terms of clock time and calendar time. Objective in the sense that the frames are stable and fixed.	Second, days, years
Temporality	Time that are embedded within processes that occur in space and in a particular context	Process, irreversibility, impermanence, ageing

Tempo	The intensity at which activities are conducted	Pace, velocity, rate of change. A concrete example is how students have to adapt to the institutional tempo of the university
Timing	Social, political, economic, environmental, religious and socio technical context time	Coordination; Knowing good and bad times for action (kairos). A concrete example is the opening and closing times of institutions.
Time sequence	The understanding of order, succession, and priority	Series, phasing, simultaneity
Time duration	The extent, temporal distance and horizon of time	Duration, continuity
Time past, present, and future	The individual and collective past, present and future	Horizons, memory, anticipation, past present, present present, present future, future present

As time is experienced in space and in a particular context, it follows that analysis of time cannot be divorced from space, matter, and context. Although the co-constitution of time and space is a core assumption of the timescapes framework, empirical analyses of timescapes analytically privilege the temporal over the spatial to foreground the temporal relations without completely obscuring the role that space and context play. Notable exceptions are scholars from geography who, around the same time, in the early 2000s, became concerned with atemporal theorizations of space and spatiality (May & Thrift, 2003). In other words, how theories of space tended to shy away from including time.[2]

[2] May and Thrift lament how geographers had "the tendency … to draw a strict distinction between Time and Space. Within such a dualism, where Time is understood as the domain of dynamism and Progress, the spatial is relegated to the realm of stasis and thus excavated of any meaningful politics" (2003, pp. 1–2)

The particular way that the timescapes framework extends subjective time is by identifying the eight features of subjective time that allow for more granular description and characterization of the particular type of social time, as well as its salient temporal features, under focus. It allows scholars to describe with greater precision which aspect of time is under study and also allows for greater richness. Whilst other scholars who align with the subjective standpoint agree on the relational nature of time, this is an implicit point that the timescapes perspective places front and center.

More recently, Kitchin (2023) takes up Adam's timescapes and extends it to examine the ways that digital technologies and infrastructures have introduced new temporal experiences in different sectors of society e.g. government and more. Temporality is defined as "the diverse set of temporal relations, processes and forms that are enacted and experienced through individual and collective action. Temporalities are embodied, emplaced, materialized and experiential" (Grant et al., 2015 in Kitchin, 2023, p. 4). Kitchin specifically attends to the way that time can be compressed and stretched in different ways for different people and the role that technologies and infrastructures (rather than Adam's focus on late-stage capitalism) play in that shaping and experiencing of time.

2.2.3 Chaotic Time and the topology of times

Yet another variation on the subjective perspective on time comes from a conversation between Michel Serres and Bruno Latour (1995), in which Serres articulates a chaotic theory of time. Akin to previously surveyed articulations pushing back on the classical, i.e. Newtonian theory of time, Serres' chaotic theory of time draws heavily on chaos theory and topology. In advancing this theory on time, Serres takes to task many of the same critiques raised by scholars aligned with subjective understandings of time, but through his distinctive analytical approach allows him to follow familiar arguments in ways that surface unexpected insights.

Latour, interested in Serres' analytical approach, asks how Serres comes to understand ancient Greek philosophers, such as Lucretius, as contemporaries. In his response, Serres takes issue with the implied classical theory of time implied in the word "contemporary." A helpful concrete example of the way a chaotic theory of time presents itself is through the example of a modern car. Serres explains:

> *"In order to say "contemporary," one must already be thinking of a certain time and thinking of it in a certain way… So let's put the question differently: What things are contemporary? Consider a late-model car. It is a disparate aggregate of scientific and technical solutions dating from different periods. One can date it component by component: this part was invented at the turn of the century, another, ten years ago, and Carnot's cycle is almost two hundred years old. Not to mention that the wheel dates back to neolithic times. This ensemble is only contemporary by assemblage, by its design, its finish, sometimes only by the slickness of the advertising surrounding it."* (Serres & Latour, 1995, p. 45).

As I understand Serres' argument, much like others, there is a push back on the notion of the "arrow of time" as implied by the word contemporary. The arrow of time is an understanding of time as "that of the line, continuous, or interrupted" (Serres & Latour, 1990, p. 57). Serres moves on to describe our conception of the arrow of time or linear march of progress as akin to the geocentric model and "narcisisstic" of us to presume that we are always centered temporally. Moreover, Serres remarks that "it can never be demonstrated whether this idea of time is true or false." (pp. 49-50). Instead, Serres maintains that "Time flows in an extraordinarily complex, unexpected, complicated way" (Serres & Latour, 1990, p. 57).

Taking the chaotic theory of time seriously, means that multiple temporal orientations, towards the past, towards the future, are often enmeshed into the present and are made salient. As with the example of the latest model car, we can see multiple temporal orientations the past - through the various scientific and mechanical components, the present - through the assemblage, and likely the future - through advertising. But how can the past, present, and future coexist simultaneously? Serres maintains that the chaotic theory of time

> "*More intuitively, this time can be schematized by a kind of crumpling, a multiple, foldable diversity. If you think about it for two minutes, this intuition is clearer than one that imposes a constant distance between moving objects, and it explains more. … Earlier I took the example of a car, which can be dated from several eras; every historical era is likewise multitemporal, simultaneously drawing from the obsolete, the contemporary, and the futuristic. An object, a circumstance, is thus polychronic, multitemporal, and reveals a time that is gathered together, with multiple pleats.*" (Serres & Latour, 1990, p. 60).

From the above, the four main assumptions that I wish to highlight are that 1) time is paradoxical, or put another way time is relative not only for people but in relation to other times, 2) This allows time to be folded and twisted, 3) Time doesn't flow, rather it percolates, sometimes it passes and other times it doesn't pass, 4) order can emerge from disorder/chaos are

that. I spend a lot of time with Serre's chaotic theory of time as later I present the ways that scholars in STS have operationalized folded time in their study of the data archival and preservation practices.

In reviewing the literature presented above the assumptions of the subjective time, timescapes and folded time perspectives about time are the following, 1) time is not a singular entity, nor is it universal, instead it is complex and multiplex, 2) following from the assumption before then multiple types of times coexist, 3) different types of times are associated with each other, 3) all time is social time – actions and practices do not just occur in time but are produced by time and reproduce time, 4) there are eight irreducible dimensions within a timescape, 6) the dimensions are interrelated, 7) not all dimensions may be salient, 8) time can only be known through space and context – these two are interdependent, 9) a timescapes framework foregrounds the temporal, 10) time is not always linear, 11) nor does it flow unidirectionally.

These perspectives and assumptions make it possible for this study to identify and describe the specific types of time that shape oceanographic data, data practices of oceanographers and data infrastructure. The following section builds on this and turns to literature on research data lifecycle models.

2.3 The Temporalities in Data Lifecycles

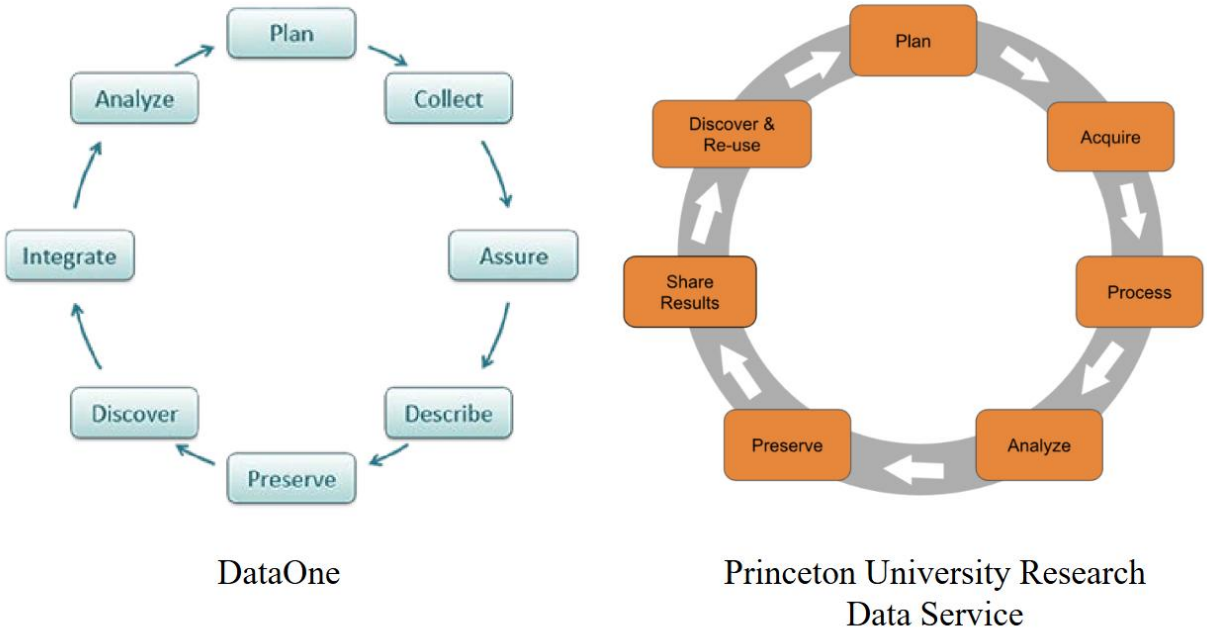
This section focuses on research data lifecycle models. It starts with classical research data lifecycle models where I surface the implicit assumptions about time, how data move through time, and when they are made available. Classical, also known as prescriptive research data lifecycle models, have been critiqued for not being a faithful representation of researcher's data practices. I then present more grounded models of data lifecycles, from researcher-centered models that center career trajectories to epistemic models that highlight temporalities of data preparation and reuse so that data can be used to make evidential claims.

2.3.1 Prescriptive research data lifecycle models

Most models of data and its lifecycle can are prescriptive models (Mosconi et al., 2019). Prescriptive models of research data are abstract and include pre-defined stages for research data (see Figure 1 for examples). Within a given model, the number of stages that research data is thought to pass through, as well as their sequencing, which stage comes before and after, may vary. However, in broad strokes, these research data models share several key similarities.

Movement of data is depicted linearly within these models and the stages are individually contained and fixed.

Figure 1. Examples of research data lifecycle models



DataOne

Princeton University Research Data Service

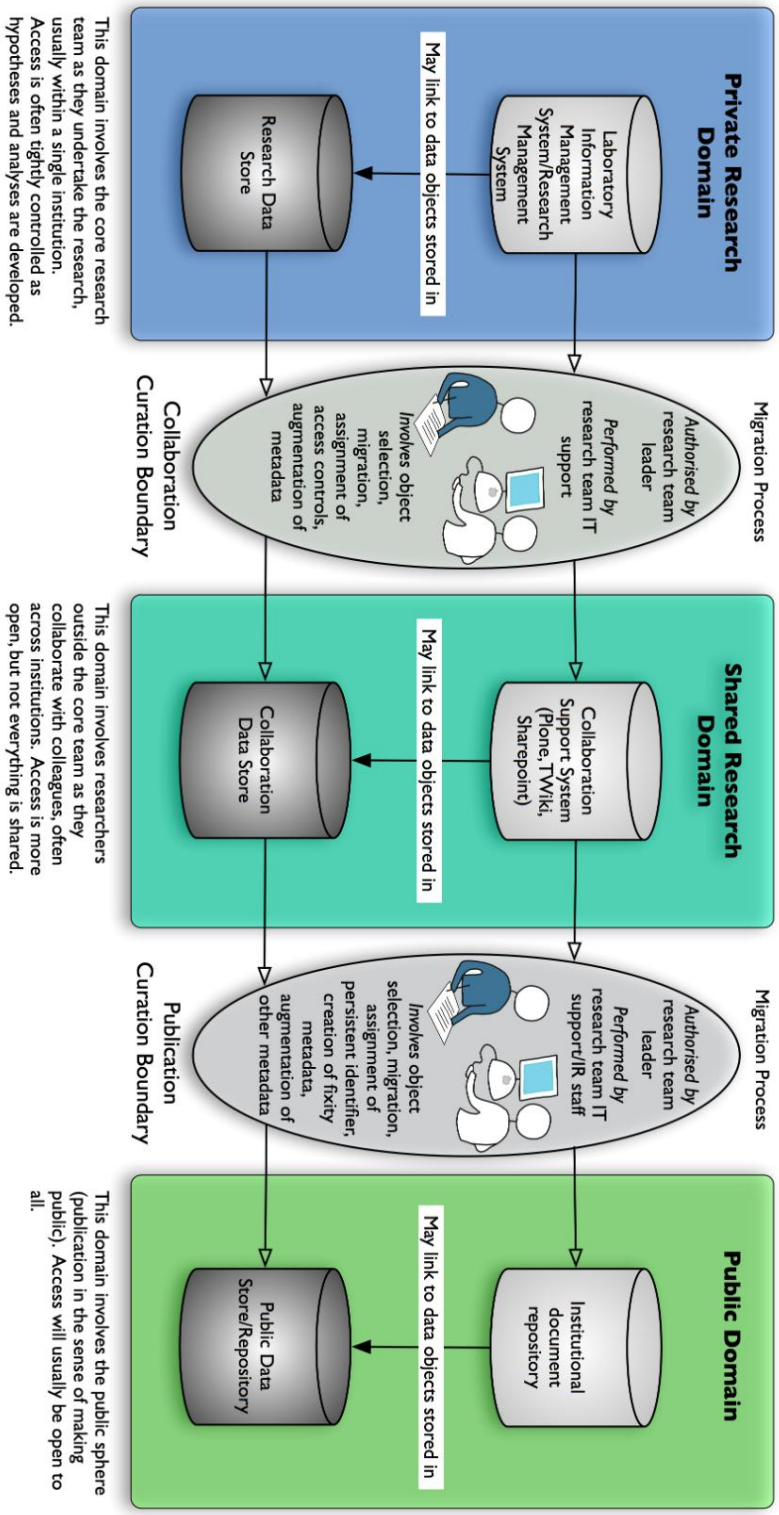
First, positioned at the top of the circular diagram is the planning stage and this is often depicted as the first phase in the data lifecycle. Second, stages that are typically associated with active phases of a research project are presented, such as collection and analysis. Last, sharing and preservation are typically data stages that are associated with research data being made available beyond the research team that produced or collected that data. Data reuse is imagined as directly spawning new avenues of inquiry and further planned research to begin this cycle anew. The exception is the prescriptive research data model from DataOne, which begins with data re-use or what they term the "discover stage," which is reflective of their organization, a network of interoperable data repositories. More importantly, all these models are devoid of the presence of actors that would be involved in enabling research data to move through different stages. Overall, there is a sense that research data inherently possess this lifecycle, with each stage requiring equal attention.

Mosconi et al., argue that these abstract models help illustrate what is being requested by funding bodies, but that they neglect the "collaborative infrastructure in which researchers

actually engage in the business of storing, managing, and archiving data" (2019, p. 753). The authors point to Treloar and Harboe-Ree's (2008) research data model, the Data Curation Continua, as a closer representation of research data as it moves across space and time. The model, as illustrated in Figure 2, represents a more dynamic model of data management.

Notable differences from the Data Curation Continua model and the research data lifecycle models are the recognition that labor is required to move data between private, shared and public domains. The term domain is used to delineate who has access to research data and at what stage. In the private domain the core research team has exclusive data accessibility, in the shared domain those with access expand to include research collaborators, and ultimately in the public domain, the broader research community and the general public is envisioned to have access to research data. The model highlights how PI's and institutional resources play a critical role in moving data between domains. Whilst this model highlights the data management activities involved in producing accessible and open data, the implicit linear assumption of data's lifecycle remains. This model also exclusively concerns the active phases of a research project as there is no treatment of data management planning or data preservation.

Figure 2. From Treloar & Harboe-Ree (2008) The data curation continua model



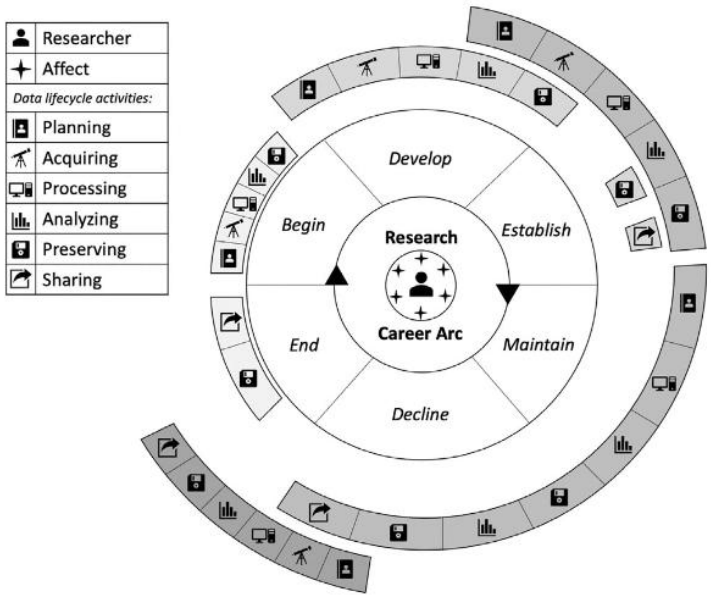
In the prescriptive models of research data reviewed, time is absent. Although there is a type of linear time implied within these models, research data is depicted to move unidirectionally from stage to stage, or domain to domain, as the hypothetical research project progresses. Researchers have argued that this is an overly simplified and unhelpful way of understanding the movement of data, terming these "linear model[s] of data flow". Through the example of ocean data repositories, Baker and Chandler suggest that scholars rethink data flows as non-linear as the movement of data is a "complex system of frequently ill-defined relationships between local repositories and a larger-scale community web of institutional repositories, discipline-specific centers, and national archives." (2008, p. 2134). Although Baker and Chandler are discussing data that have already been made available, a similar non-linear model of research data is likely occurring before the data enters the public domain.

This subsection has focused on prescriptive, or top-down, models of research data and how time is implicit but linear. A second category of models of research data are what Mosconi et al., (2019) term "pragmatic models of data." There is no singular example, or representation of a pragmatic, or bottom-up, model because a pragmatic model understands data curation and management as embedded within researchers' data practices.

2.3.2 Researcher-centered data lifecycle models: Career timelines and affective entanglements

More recently, through interviews with Astronomers, Stahlman (2022) proposed a researcher-centered research data lifecycle model that examines the ways that researchers' career lifecycles are enmeshed with and affect data lifecycles (see Figure 3). Stahlman identifies six affective dimensions: painstakingness, loose ends, altruism, intellectual passion, legacy, and nostalgia. For instance, early-career researchers may face delays in sharing their data, as preparing data for public availability is a painstaking and time-consuming process. These delays are temporary loose ends that early-career researchers may be comfortable with as they prioritize other activities such as publishing and grant-writing. Loose ends, for late-career researchers, however, may instill a sense of urgency and motivation to share data as they may be looking to tie up unfinished work before they retire.

Figure 3. From Stahlman (2022) researcher-centered data lifecycle model



Note. This model unlike classical research data lifecycle models highlights career timelines and affective dimensions experienced by researchers

As it relates to this dissertation, Stahlman's model illustrates that data sharing may not occur at the stage envisioned in prescriptive data lifecycle models. Instead, data sharing and preservation align more closely with the career arcs of the researchers themselves. This is particularly salient for those researchers who, like Stahlman's interviewees, are working with "long tail" data, defined as small in size, heterogeneous, held by researchers and smaller groups of laboratories. Yet, while Stahlman foregrounds the entanglement of data lifecycle stages with researchers' own timelines, other researchers shed light on the epistemic timelines embedded in data itself through researchers' data practices.

2.3.3 Epistemic times of data: Data time, phenomena time, and scientist time

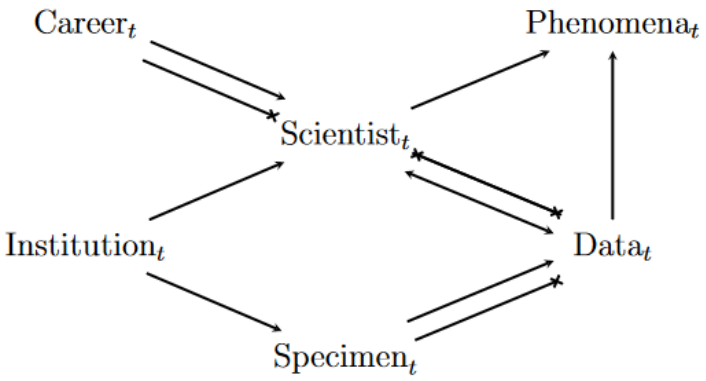
In contrast to the prescriptive models of data discussed earlier, Leonelli and other philosophers of science present a more dynamic and layered account of data. Through a cross-case comparison of two case studies of data reuse in biology, Leonelli identifies two types of temporalities of knowledge production and interpretation: 1) Data time, are "the temporal dimension of data practices used to prepare and manage data so that they can be subjected to inferential reasoning" (Leonelli, 2018, pp. 742–743) and 2) Phenomena time, are the times of the processes and

artifacts that scientists are investigating in their research (Currie, 2021, p. 105). Data times, then are "the ways in which researchers manage time in their work … [and] the time spent in the production, dissemination, and analysis of data" (Leonelli, 2018, p. 743). In other words, data time refers to the time researchers spend making data usable. That is, the temporalities related to transforming collected data into forms that can be analyzed and subsequently used to back up research claims. In other words, the time it takes scientists to transform data into analyzable, evidential, and shareable forms.

The significance of identifying and distinguishing between data times and phenomena times, Leonelli contends, is that first knowledge or ignorance of data time affects how data resusers assess the evidential value of data, leading some data to be deemed unreliable or unsuitable for their reimagined purpose, and others to be deemed suitable. Second, as it relates to the phenomena under investigation, data time directly impacts phenomena time as data times determine the scope and granularity at which phenomena can be explored. Critically, then, data times affect researchers' understanding of the phenomena they are investigating.

Wylie extends Leonelli's model by adding scientist time, see Figure 4., defined as "the temporal component of how scientists and other workers make decisions about research practices based on the values relevant to their particular situations" (Wylie, 2024, p. 2). Foregrounding the role of scientists in the Data Times model is one way to resurface time-based assumptions and values about data that are important for making knowledge claims but become obscured through the knowledge production process. By studying vertebrate fossil scientists, fossil preparators, conservators, and collection managers, Wylie contends that the time experienced by scientists is due to a combination of personal, professional, epistemic and ethical values. Moreover, because time is co-constituted, additional temporal categories that are related to scientist time, include career time, specimen time, and institution time.

Figure 4. From Wylie (2024) Scientist Time model



Note. This figure shows the temporal influences on scientists and how these influence scientists, data and knowledge production. Lower case T (t) refers to notions of time, arrows denote relationships that enable each other whereas constraints are denoted by an x.

Leonelli's articulation of data time and phenomena time foregrounds the ways data's temporal dimensions shape its immediate (re)interpretability and evidential value. However, these issues do not disappear once data moves beyond the research project, instead, they are made more acute. As data circulates, is repurposed, or is left in a database, questions about what happens to data when those practices fade from view are likely to continually resurface. This is because, as time passes or rather percolates, key questions shift from the alignment of data and phenomena to broader concerns with data's epistemic standing over time.

This leads to broader discussions in STS, information studies and philosophy of science that foreground data's afterlives. By this, I mean data's capacity to endure, to decay, to become latent, or to resurface in new evidentiary roles beyond the immediate context of their production and use over time. The following section explores how these concerns are taken up in discussions of legacy data, as epistemically fragile, and as always marked by its past trajectories, but over time, these trajectories may be rendered so faint as to make them effectively invisible.

This section shows that data does not simply move through predefined stages. Instead, data and their availability and reuse are shaped by the multiple and relational times of researchers, institutions and phenomena under study. Taken together, they complicate the notion of a singular, universal data lifecycle model. The next section moves from literature focusing on data to those exploring how time materializes within scientific and data practices.

2.4 Time and data practices

This section draws from information science and sociology scholarship to examine how time is made, managed, and coordinated by researchers. It focuses on how time constraints are not only experienced but also produced. The section also considers how timing can serve as a marker of a competent practitioner, and how data practices, such as planning for preservation, are shaped by temporal orientations, particularly through anticipatory efforts aimed at future goals.

2.4.1 Time and temporality in scientific, information, and organizational practice

Time is important and information science scholars such as Savolainen (2006) noted, is "one of the main contextual factors of information seeking" (p.110 in Haider et al., 2022, pp. 2-3). In earlier information science scholarship, time has been operationalized through the temporal dimensions of frequency and duration. I suspect that this penchant to notice the above-mentioned measures speaks to the embedding of the infrastructures of objective time in our everyday lives.

More recently, Greyson (2016) show how information seeking practices change over time and how time constraints informs how individuals decide how and when to engage with information. Haider et al., (2022) notes how in information science literature "problems of information are frequently formulated as problems of time, and vice-versa" (2022, p. 3). Put another way, time problems, such as time constraints, are indeed information problems. This argument holds face validity, but as I will show in Chapter 6. theorizing time constraints as a lack of information only provides a partial picture. Some time constraints, such as those experienced by oceanographers, are also fundamentally about time.

Whilst the work surveyed above and similar, make some inroads into the role of time and temporality in information practices, I would like to speak also to the limitations of this approach. It would be easy to default to noting the time constraints, limits, and pressures that individuals, groups, or organizations face without articulating the specific ways in which these time constraints are produced and are the result of temporal dimensions, beyond their frequency and duration. For example, studies in data sharing note that participants reported or experienced a lack of time, but further attention and theorizing is lacking.

One way to understand a lack of time, theoretically, is through a commodified view of time, where "time is a scarce resource which practices consume." However, sociologist Elizabeth Shove notes that "temporal arrangements arise from the effective reproduction of everyday life,

or, to put it more strongly, practices make time." (2009, p. 17). What this means concretely is that for any practice to exist, whether that be an information practice or a more mundane one, such as a morning routine, people have to do it. These opposing understandings of time, consumption vs. production, might at first glance seem like another run-of-the-mill dualist conception of time, but upon closer examination, shifting the focus from consumption to production has important analytical implications. For example, under the consumptionist view of time, which understands practices to consume time, we end up with the identification of time constraints. Under the productionist view of time, we can attend to "how temporalities of practice are produced, altered and disrupted" (Shove et al., 2009, p. 5). Furthermore, the productionist view of time is flexible enough to also attend to negative instances of production, for example, to delays and constraints.

Following from this productionist view of time, Shove (2009) introduces the notion of "practice-time profiles" to describe the temporal characteristics of practice and defines it as the "embedded conventions of duration, sequence and timing associated with the competent performance of a practice" (p. 25). Putting Shove's scholarship in conversation with information science research, Shove's intervention is three-fold. First, the portmanteau term "practice-time" suggests the inherent interlinking of practices with time, in a similar way that Adam's timescapes serves as a continual reminder of the co-constitution of time and space. Second, practice-times, rooted in a productionist understanding of time, emphasizes that practices don't just occur in time or merely take time, but actively produce specific times of time. Third, "practice-time profiles" emphasize timing as an important aspect of practices. As the first two points, or interventions, have been discussed, I move to discuss why timing is important for the temporal characteristic of practices.

Time is not just made individually, but are also made and shaped at the collective level. Drawing on Zerubavel's seminal work in the sociology of time, Shove argues convincingly that we hold quite rigid conventions around timing, or what constitutes "the proper time." For instance, as Shove illustrates, "it is almost inconceivable, for example, that an event such as a dance would be scheduled for the morning (even on non-working days (1981:8). Likewise, phoning at 3 a.m. is itself a signal of some kind of emergency" (Zerubavel 1981, p. 8 in Shove et al., 2009, p. 5). In different contexts, both individual and collective, "the proper time" involves different temporal trajectories and characteristics., "taking too long, finishing too quickly or

doing things in the wrong order signals incompetence but [in more forgiving situations] does not result in total failure." (p. 25) Knowing the "proper time" or "right timing" in addition to the sequence of a practice is a marker of competence in a specific practice.

This idea of time being shaped at the collective level is well articulated in CSCW literature. Reddy et al., (2006) define temporality as "the temporal organization of activity as a practical accomplishment of social actors" (p. 31). Scientific projects implicitly contend with aligning different types of temporalities in their work. Jackson et al., (2011) argue that "time matters" in distributed scientific work is a vital yet understudied phenomenon. Issues related to time and temporality are important as their alignment requires work by human and non-human actors and is what makes large-scale and distributed modern scientific work possible. Drawing from ethnographic research in different settings they present four types of temporal rhythms that occur at different levels of analysis and social structures: 1) organizational, 2) infrastructural, 3) biographical, and 4) phenomenal. Organizational and infrastructural rhythms concern the meso structure, biographical concerns the micro and individual, and phenomenal concerns the macro structure. Tensions surface when different temporalities are misaligned.

In summary, time is made and managed both individually and collectively, for example, in large distributed scientific work. Navigating these multiple personal, phenomenal, and collective times is challenging, resulting in conflicting temporalities and misaligned temporal rhythms that require work to re-align them. (Jackson et al., 2011; Reddy et al., 2006).

2.4.2 Retrospective and Anticipatory temporalities in scientific work

Temporal orientations, is a practice and affective condition. Science has long sought to extend its temporal reach by enabling knowledge claims to be made in the present about the past as well as the future.

One anticipatory, i.e. future-orientated practice is that of planning and plans have also been understood as "scalar vehicles." Steinhardt and Jackson use this term to denote the ability of plans to coordinate, contain, align and balance. In the context of their ethnographic research, distributed collaborative scientific work, plans are a formalized way to contain the sprawl of projects that are distributed in space and time (2014a, p. 1512). Plans act as scalar vehicles in two ways. First, plans coordinate sociomaterial worlds, and the interactions of actors across spatiotemporal scales that would not be possible in less formalized strategies. Second, plans balance and align interests of different social structures – macro, meso, and micro. Even when

plans end up unbalanced and misaligned in practice, the plan is the best attempt at achieving those goals. By paying attention to what is being balanced and aligned, Jordan and Jackson contend that plans reveal the new materials forms that are imposed to support the alignment of local culture to that of the broader arc of scientific work.

Meanwhile, some collective practices, i.e. anticipation practices are those that oriented spatially distributed human and non-human actors towards a specific vision of the future (Steinhardt & Jackson, 2015). Whilst scholars in various disciplines have examined anticipation in various contexts, in this dissertation, specifically in Chapter 5, I draw on Joanna Radin's (2015) concept of "planned hindsight" which speaks to how an imagined future, that of the long-term availability of agricultural samples, is being mobilized in the present to justify data practices. I employ planned hindsight to analyze DMPs and identify the ways that the data afterlives imagined by data policy diverges from the data afterlives that researchers plan for.

The term "planned hindsight" was introduced by systematic taxonomists, the science of biological classification, during a 1984 NSF-sponsored workshop report, as an approach for collecting and preserving frozen tissue samples to enable their future reuse.[3] This proposition emerged in response to concerns that biological specimens were being collected in an ad hoc and uncoordinated fashion, limiting their long-term scientific value. Planned hindsight was a call to structure present-day collection and curation practices around the imagined needs of future scientists, asserting that careful planning would increase the ensure that scientific worth of these samples would be retained if not amplified over time. Put another way, planned hindsight is the future being used to justify decisions, actions, and practices in the present.

As Radin (2015) argues, planned hindsight enacts a complex and layered temporalities. Rather than bridging the present actions to future outcomes, it generates contradictory logics of

[3] Brief overview of the origin of the term planned hindsight: "This position was made explicit in an influential 1984 report to the US National Science Foundation, based on a special workshop sponsored by the Association of Systematics Collections (Dessauer & Hafner 1984). The goal of the workshop was to take stock of the hundreds of known existing, ad hoc collections of frozen non-human tissues that had been made for specific projects and to provide guidelines that would aid their future reuse. The authors of this report, systematic taxonomists Herbert Dessauer and Robert Hafner, organized their recommendations around a temporal strategy that they called 'planned hindsight.' They wrote that while 'samples collected to explore one problem may prove to be of historical value for investigating many future problems in distantly related areas ... their value in retrospective studies should increase dramatically with carefully planned sampling programs.' Dessauer and Hafner were concerned that scientists had mitigated the future value of their collections by having created them 'independently without oversight'" (Dessauer & Hafner, 1984, pp. 10-11 in Radin, 2015, p. 363).

"prophecy and prognosis" (p. 372) or put simply of fantasy and fact.[4] The value of preserved tissue samples can never be fully determined in advance but emerges through their reuse. In this way, planned hindsight reveals both the ambitions and contradictions of scientific preservation initiatives.

Radin, like M'charek (2014), draws on Serres' theory of chaotic time. Through operationalizing the idea of folded time shows how data practices can reveal temporal relations by making the past or future more salient. In M'charek's case what is being folded is the present into the past. This folding, M'charek argues erases traces of race from the He-La sequence's present. Meanwhile, in Radin's study (2015) the future is being folded into the present, the future is made more salient. Put simply, we are doing things today, in service of a specific future.

Braun (2023) and other scholars, such as Van Allen (2023), demonstrate how different types of temporal orientations from personal histories (retrospective) to imagined futures (anticipatory) affect the curatorial practices of scientists in museums preserving butterfly samples. Van Allen (2023) argues that different temporal orientations are present in scientific practices "what is saved, is not only the biological matter of the butterfly specimen, but also the hopes and fears of the preparator, their moments of projecting into the future for what a particular specimen may be used for, and for what the role of preserving specimens for that imagined future may be" (p. 299). As not all parts of each specimen can be preserved these intricate retrospective and anticipatory practices impact "what parts of a specimen are saved or discarded, deemed to be precious or to be biowaste. Each of these pieces of the specimen then carry their own timeline, and their own set of possibilities or endings. (299).

Temporal orientations are not just an individual experience but are a collective one. Wylie (2019) argues that different disciplines and epistemic communities have their own collective temporal orientations that affect their practices and epistemological assumptions about data, writing "it is common for a multi-field community to tacitly hold a plurality of temporal orientations which all of which shape members' practices and interactions." (p. 21)

[4] Radin's insightful analysis of the contradictory logics inherent to planned hindsight "Ideas about the need to preserve bits of endangered bodies circulate both as prophecy and prognosis, fantasy and fact, which makes the project of freezing tissue appear worthy of investment, even if its ultimate value can neither be clearly delimited nor guaranteed. The maintenance of these collections depends upon local spaces of experience – what a given specimen or set of specimens can actually be used for as well as what they can plausibly be imagined to be used for – in any given moment." (2015, p. 372).

By shifting the understanding of time as a resource to be used up to something produced through action, this section underscores the situated nature of time in data practices. Whether it is coordinating work across distributed scientific teams or aligning activities with disciplinary norms, time is actively shaped through social and institutional arrangements. As this section has focused on time and data practices, the next section instead surveys literature on time and infrastructures.

2.5 Infrastructures and Time

This section surveys scholarship in infrastructure studies to examine how time shapes the development, maintenance, and transformation of cyberinfrastructures. It discusses how infrastructures evolve across multiple timescales while also structuring how time is experienced by users and maintainers. Key concepts such as the long now, infrastructure time, and the continuity and discontinuities in infrastructural development highlight how infrastructures are understood as stable and robust. This section also introduces convivial decay to foreground how aging infrastructures are actively managed in their final phase.

2.5.1 Timescales in Infrastructure development

Time has been an enduring interest to infrastructure studies scholars. One way that time has been examined is by looking at the lifecycle or development stages of infrastructures and how time is a scalar dimension of infrastructure (Edwards, 2003). Drawing on Giddens' structuration theory (1984) Edwards argues that:

> *"outside rare moments of creation or major transitions, infrastructures change too slowly for most of us to notice; the stately pace of infrastructural change is part of their reassuring stability. They exist, as it were, chiefly in historical time. Partly because of this, infrastructures posses the power to shape human time, shaping the preconditions under which we experience time's structure and its passage." (Edwards, 2003, pp. 194–195).*

Some key things to draw from Edward's earlier writing about time and infrastructure is that infrastructures and time are co-constitutive. Infrastructures shape how we experience time and the structure of the time experienced. At the same time, time also shapes infrastructure, as Edwards drawing on Gidden's structuration theory advances that on "short" timescales such as that of human lives infrastructures appear stable and durable. Their instability and empheral

nature can be seen on long-term historical or geophysical timescales. “The large technical systems group convincingly showed that these and similar patterns can be found in the history of many major infrastructures… individual infrastructures follow a life cycle, a developmental pattern visible only on historical time scales.” (p. 200).

In the late 2000s, time became central to discussions of the base-level tensions in infrastructural development, particularly those between short-term concerns such as funding decisions and the long-term timescales required for infrastructures to develop (Edwards et al., 2007). A focus on longer-term timescales was important as scholars such as Karasti and colleagues argued extending our analytical lens to take into account different time horizons “allows exploration of infrastructural development issues, such as emergence vs. intentional development, and openness vs. closedness of solutions” (p. 380)

To explore these longer temporal issues in infrastructures, environmentalist Stewart Brand’s concept of the “Long Now” has been employed as an organizing principle “The long now ties together the concerns and scales, and encourages a consideration of how today’s planning will effect tomorrow’s technologies through the practical work of designing, (re)constructing, and then maintaining these systems (p. 377). Put another way, we should work in the present today, in service of the specific future of sustainable infrastructures tomorrow. Infrastructure developmental work, then, is similarly oriented in time as the museum workers preserving and archiving agricultural samples in Radin’s (2015) field sites - also falls under the purview of “planned hindsight.” Short-term demands and long-term goals have been identified as a critical aspect of and enduring challenge in infrastructure development (Ribes & Finholt, 2009). Having established that the long now is important theoretically and practically, a follow-up question is how long concretely is this long now?

Karasti et al., (2010) extend this concept by moving beyond simply recognizing the tension that different time horizons pose for infrastructural work but that the interplay of the two timescales, short and long term, can be a synergistic approach to infrastructure development. In their study of the development of a metadata standard in ecology, they identify two distinct temporal orientations in information infrastructure development work, namely “project time” and “infrastructure time.” Amongst other qualities, project time, is characterized by a time scale of 3-5 years, whereas infrastructure time spans at least multiple decades to 200 years. Karasti et al., do note that the temporal scales for information infrastructures, in particular those based on

digital technologies, i.e. cyberinfrastructures, may change at a rapid tempo, as compared with those information infrastructures that don't rely so heavily on digital technologies. Nevertheless, their estimated floor, or minimum expected lifetime, of cyberinfrastructure is still in multiple decades (2010, p. 402).

2.5.2 Continuities and Discontinuities of Scientific Cyberinfrastructures

Edwards et al., (2007) suggest that the longer time scale for the development of cyberinfrastructure is 200 years. As I understand it, their argument is not that cyberinfrastructures fundamentally have a lifespan of 200 years, although others have (e.g. Karasti et al., 2010). Instead, the argument highlights how today's cyberinfrastructures can be traced to changes in the nature of scientific work and information gathering activities of scientists (but also of the state) that have occurred since the 1800s, culminating in today's cyberinfrastructure. There are two main changes that have culminated in today's cyberinfrastructures.

The first is the division of labor in the sciences into distinct disciplines, the era of "x-ology." A desire to accumulate data, samples, specimens is a continuation from this era. Edwards et al., (2007) note that natural scientists inspired by encyclopedists began creating vast repositories of data. Many of these data collections are still around today and are housed in institutions such as botanical gardens and museums of natural history. Presently, data are primarily collected and held in digital forms.

The second is changes in scientists' communication patterns from two-way, via public and private letters, to n-way. The point of noting these changes to scientific work is because Edwards et al., argue that modern-day scientific cyberinfrastructures should be understood as coming from these developments. By recognizing the changes that have come before, we should recognize the continuities and influence of these changes that have become embedded in organizational and institutional practices and norms. At the same time, "there is also genuine discontinuity."

Due, in part, to the growth of open research data, there has been a growth in the number of data repositories over the last two decades (Pinfield et al., 2014). As the number of data repositories proliferates, various types have been defined along their institutional forms, disciplinary and topical scope. For example, institutional repositories – defined as those associated with a single university (University of Washington's ResearchWorks or University of Michigan's Deep Blue Data) or generalist repositories that accept a wide range of data types and

span across research disciplines, such as Dryad. Other repositories may house data specific to a research area, associated with a government research center, or cater to select types of data (Marcial & Hemminger, 2010). Still, other types of repositories may house rare and unique data that play a role in the formation of young interdisciplinary research areas (Darch & Borgman, 2014).

Data repositories are temporal institutions and projects as they promise the longevity or legacy of research data that is deposited within them and "increase [the] availability [of data] beyond a project's original plan or an individual investigator's career" (Baker & Chandler, 2008, p. 2132). In addition, repositories have transformed how and when researchers access others' data. Whilst previously, data may have been accessed through journal publications and information exchanges within one's professional networks, data can now also be found through data repositories. Importantly, data infrastructures facilitate or hinder data sharing and reuse. Data repositories provide an avenue for data sharing and mediate how scientific data are shared between data producers and data consumers by "rules of exchange" specified by controlled access and use policies of data repositories (Borgman et al., 2019; Eschenfelder & Johnson, 2014).

With increasing world-wide emphasis on providing access to research data, data management plans (DMPs) have emerged as the expected way for researchers to formalise and communicate their intentions to stakeholders, including to their funders. This review paper focuses on a thematic analysis and presentation of empirical research on DMPs, a literature that is surprisingly limited, likely due to the young age of the field. Research shows that, despite the benefits associated with data sharing, DMPs have potential that is not being realised to the fullest. Researchers in scholarly communication and information science primarily have evaluated DMPs using text analysis methodologies, often supplementing them with surveys or interviews. Future study, especially in the areas of machine-actionable DMPs is promising; such research is needed to further explore how DMPs can best be utilised to support data sharing (Hudson-Vitale & Moulaison-Sandy, 2019).

But more than that, Edwards et al., (2007) contend that "from one view at least, cyberinfrastructure is principally about data: how to get it, how to share it, how to store it, and how to leverage it into the major downstream products we want our sciences to produce." (31) Data repositories, Borgman et al., (2019) contend, become a critical element of knowledge

infrastructures. Knowledge infrastructures refer to "robust networks of people, artifacts, and institutions that generate, share, and maintain specific knowledge about the human and natural worlds" (Edwards, 2010, p. 7). The focus here is on the dynamic, yet durable, relations between humans and non-humans that support research and knowledge work.

However, not all infrastructures are anticipated to be stable or robust, as I present in Chapter 7. Cohn (2016) introduces the concept of "convivial decay" to describe how an aging space mission infrastructure was not abandoned but actively managed through its end-of-life. Through this conceptual lens, infrastructure repair work shifts to embracing decline rather than resisting it. Instead of understanding decay as a byproduct of time's passage, Cohn argues that infrastructural endings are socially negotiated. In other words, what decays and endures are not just technical components, like hardware or software, but also the relations among people, systems, and institutions.

By foregrounding the ending phase of aging infrastructure, convivial decay addresses a gap in infrastructure studies, which has often focused on formation, development, and repair (to sustain rather than manage end-of-life). By reframing decay as a meaningful phase of infrastructural life, one that is shaped by care, coordination, and situated decisions about what to preserve, let go, or transform. In Chapter 7, I use convivial decay as a sensitizing concept to analyze time in data infrastructures.

Time is central to the design, maintenance and development of infrastructures. This section highlights how infrastructures not only develop in time but also how they shape how time is experienced and organized. These insights highlight the importance of a temporal perspective to understanding data infrastructures.

2.6 Conclusion

This chapter surveyed how time and temporality are conceptualized and operationalized across multiple scholarly disciplines, mainly focusing on information science, STS, and infrastructure studies literature. It began by presenting three perspectives on time, objective, subjective, and chaotic, and their assumptions about what time is. It then examined implicit assumptions of time in research data lifecycle models. Prescriptive models depict data as moving through fixed, sequential stages, often omitting the labor behind data practices. In contrast, researcher-centered models highlight how data sharing is enmeshed with career trajectories. The final sections turned to how time is embedded in research practices and infrastructure development. Time is not only

managed individually by researchers, but also coordinated across teams, institutions, and infrastructures. Temporal orientations, such as anticipated future reuse of data, or following established curatorial practices, and aligning diverse timelines, deeply shapes data work and infrastructural planning. Taken together, the literature reviewed in this chapter understands time not as mere context, flowing in the background, out there independent of human and non-human actors. Instead, specific types of time, and times, are produced through practices, plans, expectations and are managed at individual, collective, and structural scales.

Chapter 3. Methodology and Study Design

3.1 Introduction

In this chapter I present the methodological orientation and study design of my study. In the first part, I discuss constructivist grounded theory, focusing on how it conceptualizes documents and how this perspective shaped my approach to working with DMPs. I then situate my study within the broader body of DMP scholarship, highlighting common methods, how my study builds on prior work, and where it differs. The second part of the chapter describes the primary data, DMPs from funded NSF OCE research projects and in-depth semi-structured interviews were generated and analyzed. I will detail how the DMP corpus was constructed through an email survey, the sampling process for interviewees and the explain the development of interview guides. I also describe how I analyzed DMPs and interview transcripts through inductive qualitative coding. The chapter concludes with discussion of the study's limitations and my reflection on the methodological challenges and considerations I encountered through the iterative process of data generation, analysis, and theory-building.

3.2 Grounded Theory

Grounded theory is an approach to qualitative research first proposed by Glaser and Strauss' *Discovery of Grounded Theory* (1967) that advances that theory be developed inductively from research data rather than applying existing theoretical frameworks deductively to one's research data. At its core, grounded theory provides a methodological framework, that of the constant comparative method, for generating theory from qualitative data. The philosophical

underpinnings of the 1967 grounded theory, now widely known as "classical grounded theory" are postpositivism and symbolic interactionalism (Aldiabat & Le Navenec, 2014). This study follows Charmaz's (Charmaz, 2014) variant of grounded theory as they re-root grounded theory in constructivism. In constructivist grounded theory, "Data do not provide a window on reality. Rather, the 'discovered' reality arises from the interactive process and its temporal, cultural, and structural contexts" (Charmaz, 2014, p. 524). Put another way, neither data nor theories are discovered but are constructed. If generating data through interviews, for example, then data are constructed as a result of the researchers' interactions with research participants.

A key data type that this dissertation works with is documents. Charmaz notes that documents are often overlooked and underappreciated in grounded theory studies, writing that "researchers often review documents but undervalue their potential for theorizing. Analysis of documents may seem far removed from first-hand observations or interviews but think of documents as texts. Most qualitative research entails analyzing texts" (2014, p. 106). Following this train of thinking yet another step further, it is understandable to see why Charmaz contends that it is indeed possible to conduct a grounded theory study with extant documents. To be clear, this is not the nature of my dissertation study, but I emphasize this argument, as Charmaz is emphasizing that at the core of grounded theory is not the data types that a researcher works with, but their approach to working with their data to produce theory through the researcher constructing categories, and theory, from that data (2014, pp. 107–108; 393). As this study's aim is to investigate the effects of the NSF's DMP mandate and analyze DMPs, the section surveys previous studies on DMPs and highlights the research methods used to study DMPs.

3.3 Research Methods Used to Study DMPs

I review DMPs in this chapter to focus specifically on how previous studies have approached DMPs methodologically. Rather than reviewing the content or function of DMPs, the primary aim of this section is to examine how other scholars have leveraged DMPs as objects of study, in order to situate and inform my own methodological approach.

One of the earliest documented usages of DMPs was by researchers involved in technically complex projects in the mid-1960s (Smale et al., 2020). During the 1970s and 1980s, DMPs were a more common presence in some engineering and scientific research areas (Smale

et al., 2020). Besides the same term used to name these documents, and that they concerned research data, there is very little in common between DMPs in the second half of the 20th century and contemporary DMPs. These documents differed in their usage, intended audience, content, and aims.

Empirical research about DMPs is a relatively young research area (Hudson-Vitale & Moulaison-Sandy, 2019). Research in this area falls under two kinds. The first group of studies analyzes DMP guidelines to understand what is expected of researchers. DMP guidelines are authored by funding agencies, or research institutions for Australia-based researchers, and provide instructions to researchers about what content should be addressed in their DMPs. In the U.S., the NSF's DMP guidance documents have received the most attention with one study also analyzing the National Institutes of Health's (NIH) DMP guidance documents (Dietrich et al., 2012; Pasek, 2017; Tian et al., 2021).

NSF guidance for DMPs has remained broad by design, which is interesting for researchers examining the guidance itself and how researchers interpret it, through DMPs. An important ambiguity in NSF DMP guidance documents is the absence of an explicit definition of what constitutes research data; the very artifacts DMPs are meant to address. As a result, it is often left to researchers to determine and articulate what counts as data. The existing NSF definition refers to research data as "the recorded factual material commonly accepted in the scientific community as necessary to validate research findings" (National Science Foundation, 2015, p. 3). While the vagueness of the definition is unhelpful in practice, it is understandable as defining research data is a complex task. What counts as data varies significantly across disciplines and subdisciplines.

The task of providing more concrete guidance has been punted down to NSF's individual units and subunits, such as to the Directorate, Division and Program levels. There were indeed significant differences in the content of NSF's smaller unit guidance documents. For example, some programs, such as OCE, have a recommended list of approved data repositories for researchers to select from depending on the types of research data that they work with. Though at times, more specific in particular ways, the overall vague and equivocal nature of DMP guidance documents was still present at the NSF's smaller units (Dietrich et al., 2012; Tian et al., 2021).

Overall, DMP guidance documents provide minimal direction, beyond providing high-level areas that researchers should address, i.e. policies for re-use, re-distribution, and the

production of derivatives. The result is that a large portion of the content included in DMPs is left up to the researchers' discretion about what content to include as well as the level of detail dedicated to each section within the DMP (Carlson, 2017).

The second group of studies analyzes DMPs to understand researcher's contemporary data management practices and to evaluate how prepared researchers are to adhere to funders' data policies. Unsurprisingly, disciplinary differences can be observed in researcher's DMPs. When focusing on data sharing activities and how researchers approach the dissemination of their research outputs this is very apparent (Bishop et al., 2022; Parham et al., 2016). Some studies observed that researchers may not be prepared to meet the requirements of the mandate, or that they have misunderstood the purpose of the DMP mandate (Carlson, 2017). Other studies have analyzed DMPs to identify areas where institutional research support services are missing and use the results from their studies to guide the introduction of new services (Carlson, 2017; Whitmire, Boock, et al., 2015). Most of the empirical studies that analyze DMPs came from the Data management plan as Research Tool (DART) project. This research initiative was funded by a National Leadership Grant for Libraries development grant. This project was a collaborative project across five U.S. universities that analyzed a total of 500 DMPs, 100 for each institution, to develop and pilot an evaluation rubric to standardize the review of DMPs (Whitmire, Carlson, et al., 2015).

Researchers who study DMPs have primarily used empirical qualitative methods. Document and thematic analysis are common, often supplemented with interviews or surveys. With the exception of the DART project (Parham et al., 2016; Whitmire, Carlson, et al., 2015), most DMP corpuses are constructed by collecting DMPs from a single institution rather than by research domain (Carlson, 2017; Whitmire, Boock, et al., 2015). Focusing on a single institution is helpful for identifying local data and research support needs, however it is less suited to examining the broader impacts of the NSF mandate on researchers' practices within a specific discipline.

This study builds on previous research by understanding DMPs as documents that shed light on researcher's data practices and their plans for their data beyond the life of the project. Like earlier studies, it combines document analysis with interviews and, like the DART project, focuses on DMPs from funded projects. However, this study departs from earlier research in two key ways. First, the primary goal is not to evaluate DMPs for compliance or to develop

evaluation tools, but rather to examine them as artifacts of data work in the context of oceanography. Second, the DMP corpus used here includes documents from multiple institutions within a single discipline. Additional characteristics of the DMP collection and process of constructing the corpus are discussed in greater detail in the following section.

3.4 Study Design

This dissertation employs qualitative research methods to generate and examine the relationship between the NSF's DMP mandate and oceanographers' data management and sharing practices. In particular, I focus on the temporalities of these practices. The primary data I drew on includes a corpus of 379 DMPs from NSF OCE-funded research projects from the mandate's first decade, 2011 – 2021. I also drew on 34 in-depth semi-structured interviews with PIs, graduate students, and research scientists, funded through OCE, and data managers and archivists who work with oceanographic data. The OCE DMP is part of a larger set of 967 DMPs collected by the Data Afterlives project. The broader dataset is unique in that it contains DMPs from multiple institutions and only from funded projects. Including DMPs from funded proposals is important, as these DMPs were reviewed and approved by peer reviewers as part of the grant evaluation process.

A combination of qualitative methods are employed to triangulate study findings. The goal of triangulating is that of "convergence and corroboration" (Bowen, 2009, p. 28). These two terms are related to the idea of triangulating evidence between different sources but mean slightly different things. As I understand it, convergence refers to the same information being found in different types of data, whereas corroboration refers to sufficient evidence to substantiate a claim.[5] Whilst corroboration occurred in earlier stages of this dissertation project, convergence didn't occur for me until the later stages. One example is finding out about the World Ocean Circulation Experiment (WOCE) in OCE data policy documents. This led to reading WOCE primary and secondary literature, which led to me tracing the timeframe specified in data policy to WOCE. Not only that, I came to learn and appreciate that key data repositories and infrastructure in oceanography were developed during this experiment, which

[5] Bowen (2009) doesn't explicitly provide definitions for these two terms. My understanding of them comes from the following text in the article "Fifth, documents can be analyzed as a way to verify findings or corroborate evidence from other sources… When there is convergence of information from different sources, readers of the research report usually have greater confidence in the trustworthiness (credibility) of the findings" (p. 30).

in turn led me to see the impact of WOCE in DMP segments. Not only in segments where PIs will explicitly cite WOCE data standards but also in more subtle ways in terms of noticing which types of data had infrastructural support and others did not.

The following sections describe the process of collecting the DMPs and conducting interviews, which together form the empirical foundation of this study.

3.4.1 Data Generation

Data generation, following Garforth (2012) is used in place of data collection as generation acknowledges the researcher's instrumental role, together with research participants, in producing the data and artifacts analyzed for this study. As mentioned above, the data I relied on for this project involved DMPs and semi-structured interviews. I discuss each in further detail below. The activities described were reviewed and deemed exempt by the University of Washington's Institutional Review Board, protocol number STUDY00019671.

3.4.2 Data management plan corpus construction

The first data generation activity consisted of collecting DMPs of active and non-active research projects funded by NSF's Division of OCE between January 2011 and June 2021. In the tradition of constructivist grounded theory, these documents are considered "extant documents" in that I, the researcher, did not play a role in their construction (Charmaz, 2014; Ralph et al., 2014). Although the mandate went into effect in January 2010, there is a time lag where projects funded in 2010 grants for those projects were generally written in 2009 or earlier but took time to review and fund.

DMPs are occluded documents as they are typically appended to the project's grant proposal and are generally not publicly available. Furthermore, they are considered PI's personally identifiable information. During the early phase of the Data Afterlives project, the research team collected publicly available DMPs via DMPTool (Bennett et al., 2021). However, one significant limitation of creating a DMP corpus via DMPTool was that it was hard to ascertain whether the DMP was associated with a funded project or if it had been created for other purposes, such as getting to know the tool. Even in cases where we were able to locate a matching NSF award, it remained unclear whether the DMP found on DMPTool was the DMP that was submitted during the grant application process.

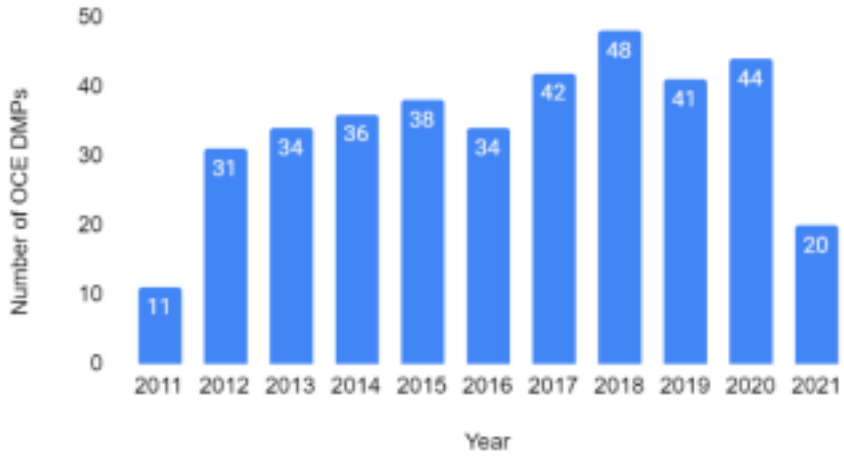
Given all the above, the project team decided that asking PIs directly through an email survey campaign would be the best approach for soliciting DMPs of successfully funded OCE projects. Whilst my dissertation focuses on DMPs funded by the Division of OCE, the research team also solicited DMPs from the Division of Biological Infrastructure, Civil, Mechanical, and Manufacturing Innovation, Secure and Trustworthy Cyberspace, and Science and Technology Studies.

To begin the DMP corpus construction, we first created a super spreadsheet collecting metadata of all NSF OCE-funded projects between 2011 and 2021 in June 2021. The super spreadsheet was created using information compiled from NSF's publicly accessible award database using the advanced search interface (https://www.nsf.gov/awardsearch/advancedSearch.jsp). The resulting super spreadsheet included the following information about funded OCE research projects: Project title, Funded amount, PI name, Co-PI name, Email contact, Type of award, Awarded institution, Project's public abstract.

The exclusion criteria for research projects included those funded through RAPID, EAGER, CAREER, EDU, RCN, Workshop, Symposium, RCU grants. These were excluded as these grants focus on teaching, conferences or from short duration projects. Grants under $100,000 were also excluded from our corpus. Moreover, if a PI had more than one project funded during 2011 - 2021, we selected the oldest grant for that PI. This decision was made to not fatigue participants with more than one DMP solicitation request. The email survey campaign was conducted from July to August 2021.

Google Mail Merge was used to create a template auto populated with information from the master spreadsheet to personalize the email request sent to each PI. The personalized elements included the PI's last name; the NSF award number associated with their research project and the research project's title. An application was used to automatically collect email attachments and store them in an encrypted folder that housed the larger research teams' materials. In total, 1689 emails were sent to OCE-funded PIs. 400 OCE DMPs were received, of which 379 qualified for this study. This resulted in a response rate of 23.39%.

Figure 5. Number of OCE DMPs by year the project was funded



From September to December 2021, manual data cleaning was conducted to verify that the DMPs received matched our inclusion and exclusion criteria. Verification included opening the email reply from the PI, opening the file attachment, and ensuring that it matched the one in the research team's encrypted folder and that it was the DMP for the project that we had requested. In a small number of cases, PIs shared their DMPs by pasting them as text into the body of the email, which was not collected by the application. In these cases, the DMP text was copied from the PI's email response into a new Google Document in the encrypted folder. Yet, in other cases, PIs shared DMPs for projects that we had not requested. These DMPs were included in our corpus if we could verify that they were for NSF-funded projects and we could retrieve the associated award metadata. At times, PIs shared DMPs for projects that were not funded by NSF, e.g. National Oceanic and Atmospheric Administration (NOAA), or were not DMPs, i.e., post-doc mentoring plans or project management plans. In these cases, these non-DMP documents were excluded from our corpus for analysis. Figure 5. Shows the total number of OCE DMPs received and included in the final corpus organized by the year that the project was funded. Except for 2011 and 2021, we received more than 30 DMPs per year.

3.4.3 Semi-structured interviews

The second data generation activity is semi-structured interviews with PIs of funded NSF OCE research and their project collaborators, this included data managers, or individuals holding various titles, that work to manage and share oceanographic data (see Tables 2 and 3).

Interviews help to surface perspectives, descriptions of practices, and reflections in participants' own words (Dumit, 2004). Likewise, interviewees can provide big-picture details and provide descriptions and reflections on important activities and events that occurred in the past that may have little official documentation (Tracy, 2013). Semi-structured interviews helped to provide context about the research project and the DMP creation, which following Charmaz, is key to placing texts into context; self-reported data management practices and their perceptions on how, if at all, the DMP mandate has affected their data practices. The interview guide for semi-structured interviews included some scripted elements and open-ended questions. This form of interviewing allows interviewers to improvise and explore emerging themes and topics that a structured interview would not.

I conducted interviews in two rounds. The first round was conducted between March - June 2022, and the second round was conducted between January - March 2025. In the first round of interviews, twenty-one 1-hour interviews were conducted. Initially, purposive sampling was used to contact PIs who had, in their email replies to our DMP solicitation request, expressed a willingness to discuss their DMP and their research project further with us. Simultaneously, snowball sampling was used and was integrated into the interview guides as the concluding question of the interviews. At first, the purpose of this question was to speak with another researcher who had worked on the same project, however as the interviewing progressed, I was referred to individuals for whom we did not collect DMPs but who our interviewees felt had a great deal to share about data management and sharing or DMPs.

Table 2. Interviewees by research discipline

Research disciplines	Number of Interviewees
Applied ocean physics and engineering	1
Biological oceanography	2
Chemical oceanography	4
Estuarine Coastal Oceanography	1
Geophysics	6
Marine Science	5
Oceanography	2

Other	4
Paleoceanography	2
Physical Oceanography	7
	Total = 34

Table 3. Interviewees by position held

Position held	Number of Interviewees
Assistant Professor	1
Associate Professor	2
Data Manager*	6
Director	3
Graduate Student Researcher	1
Postdoctoral Researcher	1
Professor	7
Project Manager	2
Research Scientist**	11
	Total = 34

*This category combines various titles. They were combined as these individuals primarily performed data management tasks.
**This category combines various titles including senior scientist, assistant research scientist, research geophysicist, senior oceanographer. They were combined as these individuals conduct research and perform data management tasks. Depending on the institution, some of the interviewees holding these titles were also PIs of NSF OCE-funded projects.

Researchers who agreed to be interviewed were provided with the consent form for their review, a copy of the interview guide, and a dedicated Zoom link for their scheduled interview. During interviews, before starting the audio and video recording, I introduced myself to the interviewees and thanked them for taking the time to speak with me. I let them know with advance notice that I would start recording the interviews. For the verbal consent

process, I asked interviewees if they had read the consent form, if they had any questions for me about the study, and its purpose, and if they consented to audio and/or video recording of their interview. If interviewees had not had a chance to review the consent form, the interview would not proceed until they had a chance to read the form.

In the interviews, PIs were first asked questions about their research broadly and what data types and formats of data they typically worked with. This first section aimed to better understand the PIs research background. The second set of questions asked interviewees about the funding process for their project, who performed data-related work during the research project, and how collaborators learned about data management best practices. The aim of this section was to better understand the context surrounding the funded project and to ask about events and actors that may not have been written into the DMP. In the third and fourth sections, interviewees were asked about the DMP mandate and the role of DMPs in their day-to-day work. The aim of this section was to get at the question of how, if at all, the NSF's DMP mandate affected their data practices. This section also asked about which data repositories the PIs deposit their data in and their experience making their data available to others. To my surprise, this line of questioning yielded responses that surfaced temporal dimensions of data. The last set of questions asked the interviewee about discipline-related changes in data collection and dissemination methods, data challenges and open access to data discussions, and who else on their project I could speak with (see Appendix A for how questions were phrased).

Before each interview, I reviewed information and materials that would give me an initial understanding of the interviewee's professional background and information on the research project for which we had requested the DMP for. These materials included, the DMP, the researcher's professional, laboratory, or project website(s), and the interviewee's curriculum vitae (CV) (see Appendix B for interview preparation workflow). After the interview concluded, I wrote narrative research memos for each interviewee and the interviews were transcribed by a professional transcription service.

In the second round of interviews, twelve 1-hour interviews with thirteen interviewees were conducted between January - March 2025. The number of interviews differs from the number of interviewees as for one interview, the PI I interviewed invited their project's database manager to participate. Key differences between the first and second round of

interviewing were how interviewees were sampled and the topics covered in the interview guide.

Again, using the NSF award database, in October 2024, I created a spreadsheet collecting metadata of all NSF OCE-funded projects that had ended between 2022 and 2024. I used the same exclusion criteria from constructing the DMP corpus for consistency and sent out 102 interview requests to PIs starting with those with newly ended projects. As mentioned above, I was surprised by how many interviewees brought up the temporalities of doing data management and sharing, though not in these words, and the challenges they experienced during the ending stages of their NSF-funded research projects, and these were explored further in my subsequent interviews conducted in 2025. Before this second round of interviewing, I had also completed the first round of coding, which gave me greater clarity into which areas I wished to explore further and which areas I lacked data for.

The second interview guide focused more on data management practices focused on sharing and preserving data and who was responsible for that work (see Appendices C and D). As the first round of interviewing highlighted the challenges of managing data during the project's close out phases, questions also focused on the ending stages of their research project. Interviewees were first asked to describe their research focus and their recently concluded NSF OCE-funded project.

The first section established context and helped situate the data-related work and decisions within the broader aims of the specific project. The second section focused on the ending phases of the project, including how data sharing was approached, the types of research artifacts shared, and who had access to these. Interviewees were also asked about their experience using specific repositories and the formats of the data they deposited. The third section of the interview asked participants to reflect on the work it took to prepare their data for sharing. Questions about how much time it required, what challenges they encountered, and what (if anything) made the process easier helped to elicit accounts of data work and difficulties that researchers encountered. In the final section, participants were asked to reflect on what they might do differently in future projects, what advice they would give colleagues and what practices or skills they saw as important for effective data sharing. This wrap-up section, similar to the first interview guide, concluded with asking interviewees for their recommendation on who to interview.

The interviews continued until saturation was reached. All the names and individually identifying information are anonymized in this dissertation. Square brackets are used to make grammatical edits or are used to blur individually identifying information without stripping the quoted DMP or interview segment of excessive context making the segment difficult for a reader to evaluate its importance. This type of deidentifying qualitative data is termed data blurring (Campbell et al., 2023).

In this section I described how the DMP corpus was constructed and how interviews were conducted. In the next section I will discuss how DMPs and interview transcripts were analyzed.

3.4.4 Data Analysis

The artifacts resulting from the data generation activities above include DMPs from the email campaign, transcripts from interviews, and research memos from interviews. During data analysis, research memos will also be generated. Document analysis will be used to analyze DMPs, and inductive qualitative analysis will be used to analyze artifacts from interviews using the qualitative coding software Atlas.ti.

3.4.5 Document Analysis

Document analysis is "a systematic procedure for reviewing or evaluating documents" (Bowen, 2009, p. 27). Document analysis can be a study's sole method, or it can be used in combination with other data sources and analyses, as in this dissertation project. Apart from being an additional source of research data to verify findings, Bowen (2009) contends that document analysis provides a researcher with many advantages. First, documents can aid in uncovering the context within which research participants operate. Documents can surface important events, decisions, rationales, and key people and institutions involved. Second, when combined with interviews, for example, document analysis can become the source of questions (Bowen, 2009).

In this study, the DMPs from funded projects provided a starting point for the initial set of interviews with oceanography PIs to learn more about the context of their projects and the role that the DMP played in their data management and sharing practices. Third, documents can provide insights into longitudinal phenomena and speak to changes over time that are challenging to discern with other methods or are practically challenging to execute.

In this study, the DMP corpus spans the first decade of the NSF's DMP mandate, document analysis of DMPs can surface answers and questions about what changes and developments in managing ocean data have changed over the past decade, what has remained the same, and about how the DMP, as a genre of documents, itself has changed. During the period of study, OCE guidelines were updated, and these changes are discussed further in Chapter 4. As such, although, one key benefit of document analysis is to corroborate and verify findings from other data sources, key differences and incongruities may arise when comparing different sources of data, for example between documents and interviews during analysis. On the surface, key differences are infelicitous, as it makes it difficult to make claims, categorize, and theorize. Something that I too experienced during the analysis and memoing stages of this project. However, what has helped is sitting with Charmaz's perspective that sharp differences "can spark insights about the relative congruence – or lack of it – between words and deeds," as points of contradiction can enrich the theories being developed in the study (2014, p. 113).

Preliminary cycle coding of OCE DMPs occurred from April to August 2022 in the qualitative coding software MaxQDA. The codebook employed for first-cycle coding drew from Hess and Ostrom's (2007) framework of knowledge commons and supplemented with concepts from information infrastructure literature. One of the codes included in this first codebook was "temporality," which captured references to time in DMPs. However, this approach had limitations. Because all mentions of time and temporality were coded under a single, broad code, I anticipated there being difficulty in distinguishing the different temporal dimensions which was the focus on this dissertation. Additionally, since the original codebook had been designed specifically for analyzing DMPs, I anticipated challenges in applying it effectively to interview transcripts, which covered different topics.

To address the limitations of the first codebook, during February to April 2024, I developed a second codebook focusing specifically on time and temporality. This version drew on Adam's (1998) timescapes framework and Steinhardt and Jackson's (2015) anticipation work. The codebook went through four rounds of revision (See Appendix E for final codebook used). My co-advisor, Megan Finn, provided feedback on the codes and their definitions, while Data Afterlives colleague, Thomas Struett, generously tested two versions of the codebook to help me finalize it. During each round of testing, Thomas and I

independently coded the same 75 DMP segments and four interview transcripts. After completing each round, I reviewed the coded output to assess where we had applied the same codes, where we had applied different codes, and where a code used by one of us was not used by the other. Noting each case helped to identify which codes were well-defined and which definitions were ambiguous and required further clarification. I documented these insights in a report summarizing the round's outcomes, which I then presented at a meeting with Megan and Thomas. In these discussions, they provided feedback on changes to be made to the codebook, including suggestions for refining code definitions, and ideas for eliminating ambiguous definitions and redundant codes. This process of intercoder agreement was not aimed at calculating a numerical score for intercoder reliability, rather, the aim was on improving the definitional clarity of the codebook and making sure that the codebook was applicable to both DMPs and interview transcripts.

The second cycle of coding of DMPs and interviews occurred during July and August 2024. During this phase, those DMP segments coded under "temporality" were revisited and coded using a codebook that was designed with Adam's temporal dimensions, and other sensitizing concepts "planned hindsight", "practice-time profiles", "collaborative rhythms", "convivial decay" and "reverse salients" from the STS. Interviews transcripts were likewise coded in the same manner.

3.4.6 Inductive Qualitative Analysis

Interview transcripts were analyzed through inductive qualitative analysis (Saldaña, 2013). After second cycle coding I wrote analytical memos to gain a deeper insight into emerging themes and memoing about codes that co-occured frequently to aid in establishing links between themes. During the second round of interviews, I transcribed and analyzed the transcripts in concert with data collection. It was only as my second round of interviews was concluding that I felt that through comparing DMPs, interview transcripts, my own written memos and discussions with my advisor and colleagues that I had settled on the themes of "PI time-related challenges", "data times" and "repository times" with subthemes of data management and sharing practices that either "save time" or are "time-consuming." Subsequently I selectively coded all DMPs and interviews transcripts following these themes.

This way of coding my data was helpful as I felt a throughline developing for my empirical chapters through these three themes. As a novice to empirical inductive qualitative

research, I made the mistake of thinking that the process was over. I submitted a draft of my empirical chapters to my reading committee a couple of months ago and received feedback that my chapters were too descriptive and lacking theoretical frames. This was a turning point for me, what I thought initially as a setback I now realize was just the next step of the process of inductive qualitative research. At first, I begrudgingly compiled a list of readings to re-read and sat down to read them. I was surprised to find that, with my data in mind, I found myself engaging with papers and their arguments in a way that I had not done in this way before. As I read, I found myself constantly asking questions about whether the arguments and claims of the paper were something I saw in my data? If yes, what were the similarities and differences between our cases? If I didn't see it, why not? I would write these down besides my general notes about the paper. Later I pulled out the writing that compared the literature and findings against my data and I began to group similar concepts together. This, together with articulating my emerging theorizing during meetings with my co-chair Megan Finn, were critical steps for how I arrived at the final sensitizing concepts of "planned hindsight", "practice-time profiles", "collaborative rhythms", "convivial decay" and "reverse salients" in Chapters 5-7.

3.5 Study limitations

There are various limitations to this study. For the email campaign, it may be that NSF-funded oceanographers who responded to our email campaign were those who had a more positive outlook on the DMP mandate and/or who viewed data sharing more favorably. We attempted to assuage possible worries that PIs may have had that this study was an audit of sorts by letting them know in our email campaign that we would not be verifying whether they executed their DMPs.

A similar case of self-selection is likely present in the interview sample. Purposive sampling was employed to identify researchers who responded to our email campaign and had expressed an openness to discuss their DMPs and the research projects with us further. Moreover, as interviews progressed, the snowball sampling strategy selected led to interviews with individuals for whom we did not have a DMP but either worked at a repository that was frequently mentioned in PIs' DMPs or was responsible for either directly managing ocean data or for overseeing the data management activities of ocean data. The tradeoff with

understanding the data management practices of individual projects and broader perspectives of data management practices in oceanography.

3.6 Researcher's stance and reflexivity

In constructivist grounded theory reflexivity is concerned with the researcher bringing one's ideas and preconceptions into the research, how those shape the inquiry, and how their preconceptions could be applied uncritically to one's data. Reflexivity is defined by Charmaz as:

> "*the researcher's scrutiny of the research experience, decision, and interpretations in ways that bring him or her into the process. Reflexivity includes examining how the researcher's interests, positions, and assumptions influenced his or her inquiry. A reflexive stance informs how the researcher conducts his or her research, relates to the research participants, and represents them in written reports." (2014, p. 555).*

To center reflexivity, Charmaz recommends that researchers keep a "methodological journal." A methodological journal is a collection of memos in which one writes about "methodological dilemmas, directions, and decisions" (2014, p. 288). I have been keeping a lab notebook as part of my involvement in the research activities of the DataAfterlives project. I used this notebook to log questions that I had about applying the codebook used during first cycle coding, notable or interesting excerpts from DMPs or briefly summarize the content of DMPs. This is not to say that my lab notebook began as a way to engage reflexivity, but that in looking back at it I see similarities with what one would write about in a methodological journal.

That said, I found myself doing the work of reflexivity, as described by Charmaz, as I wrote analytic memos about my data and codes. As I reviewed endless rows upon rows of coded DMP and transcript segments, again and again, and yet again. It wasn't until my fourth round of selective coding that I felt the method keep true to its promised outcome. For me, the work of doing constant comparison was frankly, at times, deeply uncomfortable. I felt the most discomfort when I didn't know if the process was leading me in the right direction. In any case, when we're talking about creating theory, is there even a right direction?

During this process of constant comparison, I spent a lot of time in doubt. I doubted if I had the right data to support the arguments I was making. I doubted whether the codebook I created could do justice and adequately categorize the different temporalities at play in the

data? I doubted whether I could represent my interviewees well, interviewees who had so generously shared with me their time, stories, and experiences, in good faith. How could I depict the nuances of our conversations when I have to chop it up into tiny bits? I didn't memo about all of my doubts, but they lingered and are present even as I worked on writing the dissertation. Perhaps it is these doubts that are the substance of methodological dilemmas.

My dissertation project captures a partial picture, of the complex and multiplex temporalities of data management and sharing in oceanography that have been affected by the NSF mandate. It offers a small step forward towards presenting and theorizing what challenges oceanographers face when making their data publicly available and a select exploration into the human and non-human actors that shape the timing of the public availability of oceanographic data.

3.7 Conclusion

This chapter presented the methodology and study design of my dissertation, which examines how the NSF's DMP mandate has shaped oceanographers' data management and sharing practices, focusing on time and temporality. Drawing on constructivist grounded theory, I analyzed a corpus of 379 DMPs from funded NSF OCE projects and conducted 34 in-depth semi-structured interviews with PIs, graduate students, research scientists, data managers and archivists. Through triangulation of DMPs and interviews this study design aimed to seek convergence and corroboration to surface patterns, contradictions, and complexities that a single data source might obscure.

This study builds on prior research on DMPs while departing from the goal of evaluating researcher's compliance with the mandate and development of assessment tools. Instead, this study understands DMPs as situated artifacts of data work, embedded in institutional, disciplinary, and infrastructural contexts. Similarly, interview guides were designed to gather researcher's accounts of data practices but also sought to elicit the ways that the policy mandate materialized in their data practices over the course of their research project.

This chapter also described how data were generated, analyzed and synthesized, highlighting the methodological challenges and considerations I encountered whilst doing the

work. Reflexivity, for me manifested primarily through doubt and played a central role in how I navigated the open-ended and uncertain nature of constant comparison and inductive qualitative research. I also describe how reengaging with theory once I had a solid grasp of my data helped to overcome the challenges I experienced. While this study offers a partial view, it provides a grounded account of the ways that time and temporalities of data work shape and constrain the availability of oceanographic data.

Chapter 4. Tracing the origins of OCE's data-sharing time norm

4.1 Introduction

This Chapter provides background and important information for OCE data policies that ask researchers to make their data publicly available "within two (2) years of collection". Data policies for OCE have existed since 1988. As I re-read the policies starting with the most recent ones and working backwards, I came across a document mentioned in the 1994 (NSF 94-126) policy:

> *"All WOCE data shall be made available no later than two (2) years after collection, unless specifically waived by the international WOCE Scientific Steering Group (SSG). However, several WOCE programs require principal investigators to submit data collected to a Data Assembly Center (DAC) for the purposes of quality control and data synthesis within shorter time periods. Detailed program requirements for data submission may be found in WOCE Report No.104/93, WOCE Data Management …" (NSF 94-126).*

I wasn't sure where this search was going to lead, but I was delighted to find that the time norms for data availability, of "two (2) years after collection" came from WOCE. In this chapter, I provide a brief overview of WOCE and focus on data management and sharing with the aim of providing context to why a time norm for oceanographers was deemed necessary and included in OCE's data policy, and reasons provided for why the time norm is the way it is, i.e. why 2 years? Why after collection? Importantly, this chapter traces how the time norm for sharing data 2 years after collection that exists in OCE data policies today is historically produced.

4.2 Global-scale ocean studies come of age: The World Ocean Circulation Experiment

During the mid-20th century, it became clear through World War II and the Cold War that the ocean was the next arena for warfare (Oreskes, 2021; Rainger, 2000). For instance, knowing the topography of the seafloor and being able to locate objects underwater are a few military examples of why it was a high priority to better understand our oceans. During this period, advancing basic and applied research of the ocean became critical national security priorities, only secondary to space sciences and their contributions to the Space Race (Van Keuren, 2000). However, as the Cold War entered its second half in the mid to late 20th century. By the end of the 1970s, spurred on by institutional and public pressure, pivoted to a new mission that of tackling our changing climate (Oreskes, 2021). Pressing issues at the time included the impact of acoustics on marine mammals and the carbon dioxide in the atmosphere (Oreskes, 2021; Van Keuren, 2000).

Both scientists and sponsors of science, i.e., nation states and funders, began to see that to understand the climate we had to not only better understand our oceans but also how ocean processes interacted and affected atmospheric processes. Much like meteorologists before, who in 1979 concluded the Global Weather Experiment, oceanographers desperately wanted their own global-scale research agenda and positioned themselves as foundational for advancing climate science, one of humankind's most pressing and existential crises (Woods, 1985; WCRP Publications Series No. 6, 1986). This section presents a brief overview of WOCE. This account could be presented in many ways but in this chapter, I focus on the community-planning and scientific advancements in oceanography in the 1970s that positioned oceanographers in a good place to conduct a global large-scale experiment and the WOCE planning phase that began in the early 1980s and continued throughout the decade (see Table 4).[6]

Table 4. Key phases of WOCE and their corresponding years

WOCE Phase	Years	Source

[6] I focus on planning and the lead up to WOCE as my aim was to trace the data sharing norm of data availability "two (2) years after collection." Given this, I focused on WOCE's data management plan and discussions around how scientists planned the data system in order to confirm whether this norm came from WOCE or from elsewhere.

Community-led Planning Phase	Throughout the 1970s	Thompson et al., (2001)
Announced	1982	Thompson et al., (2001)
Planning Phase	1983 – 1990	WOCE International Project Office (2002)
Data collection Phase	1990 – 1998	WOCE International Project Office (2002)
Analysis, Interpretation, Modelling and Synthesis Phase	1998 – 2002	WOCE International Project Office (1997, 2002)

Throughout the 1970s, there was an increasing realization of the importance of the ocean to climate change. The 1973 GARP Atlantic Tropical Experiment (GATE) expedition was the first major integration of oceanographic data collection within a global meteorological experiment (Thompson et al., 2001, p. 33)[7]. This was followed by the Scientific Committee on Oceanic Research's (SCOR) 1974 shifting the scientific focus of First GARP Global Experiment (FGGE) from weather prediction to climate change during a meeting in Canberra, Australia (Thompson et al., 2001, p. 33). This culminated in international momentum and support for understanding the ocean's role in climate research. By 1977, SCOR's president wrote a letter to the Secretary-General of the World Meteorological Organization (WMO) that spotlighted the importance of "the need to consider the impact of ocean variability on climate," (Thompson et al., 2001, p. 33), followed a couple of years later by the launch of the World Climate Research Programme (WCRP) in 1979, where climate change became a central concern, justifying a global ocean experiment.

Oceanography was seen as vital for all aims of WCRP, but a global-scale experiment was required to answer goal c. (see Figure 6). As such, at the May 1982 Tokyo Conference on Large-Scale Oceanographic Experiments in the WCRP, preliminary plans for WOCE were presented. The aim of the presentation was to demonstrate how WOCE would be vital to meeting the objectives of the WCRP, and the preliminary plans presented were "welcomed by

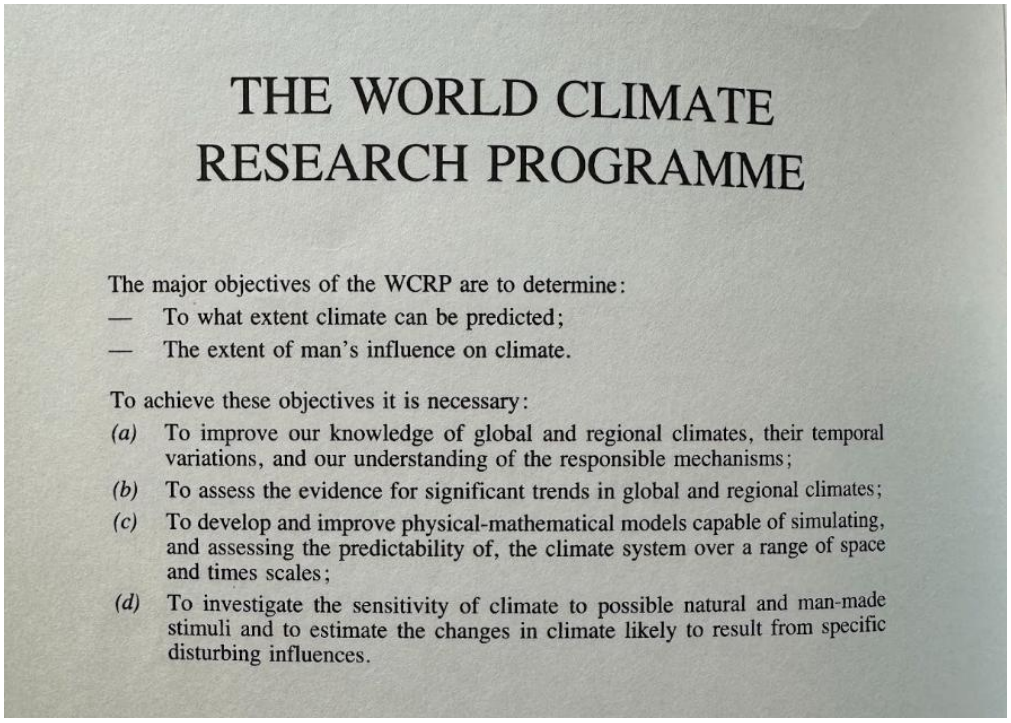
[7] GATE is an acronym for GARP Atlantic Tropical Experiment. GARP stands for the Global Atmospheric Research Programme which the World Meteorological Organization (WMO) established in 1966. SCOR Working Group 43 on 'Oceanography Related to GATE' developed and implemented an international field oceanographic programme within GARP which was at heart a meteorological experiment. GATE was noted as the "first significant step towards cooperative studies of the effect of ocean processes on climate" (Thompson et al., 2001, p.33)

the conference and [subsequently] emphasized in the conference report" (Thompson et al., 2001, p. 36). As planning began in earnest, the two main scientific aims put forth in the 1986 Scientific Plan for WOCE (WMO, 1986, p. ix) were as follows:

1) To develop models useful for predicting climate change and to collect the data necessary to test them.

2) To determine the representativeness of the specific WOCE data sets for the long-term behavior of the ocean, and to find methods for determining long-term changes in the ocean circulation.[8]

Figure 6. From WCRP (1986) WCRP's major scientific objectives



THE WORLD CLIMATE RESEARCH PROGRAMME

The major objectives of the WCRP are to determine:

— To what extent climate can be predicted;

— The extent of man's influence on climate.

To achieve these objectives it is necessary:

(a) To improve our knowledge of global and regional climates, their temporal variations, and our understanding of the responsible mechanisms;

(b) To assess the evidence for significant trends in global and regional climates;

(c) To develop and improve physical-mathematical models capable of simulating, and assessing the predictability of, the climate system over a range of space and times scales;

(d) To investigate the sensitivity of climate to possible natural and man-made stimuli and to estimate the changes in climate likely to result from specific disturbing influences.

[8] The term "representativeness" is being used to denote two separate but related ideas. As the 1986 MWO report explains "The WOCE data set will strongly influence oceanographers' ideas about the ocean circulation in the next century… Since it cannot represent the circulation as it changes over future decades, the strategy is to use the data set (including the surface forcing functions) to develop models capable of accurately predicting the changes that occur in the ocean circulation as one part of the planetary climate system. For this strategy to work it will be necessary that the WOCE data set, or specific parts of it, are representative enough of the large-scale long-term behaviour of the ocean so that its use to develop and verify models for climate prediction will have produced models that are valid for tat purpose," (p.33). The two ideas that the term is referring to is first, whether the collected data is true to the phenomena of ocean circulation during the period that it is collected 1990 – 1997 and second that the data set will be able to stand in for future ocean circulation, not observed but simulated through models.

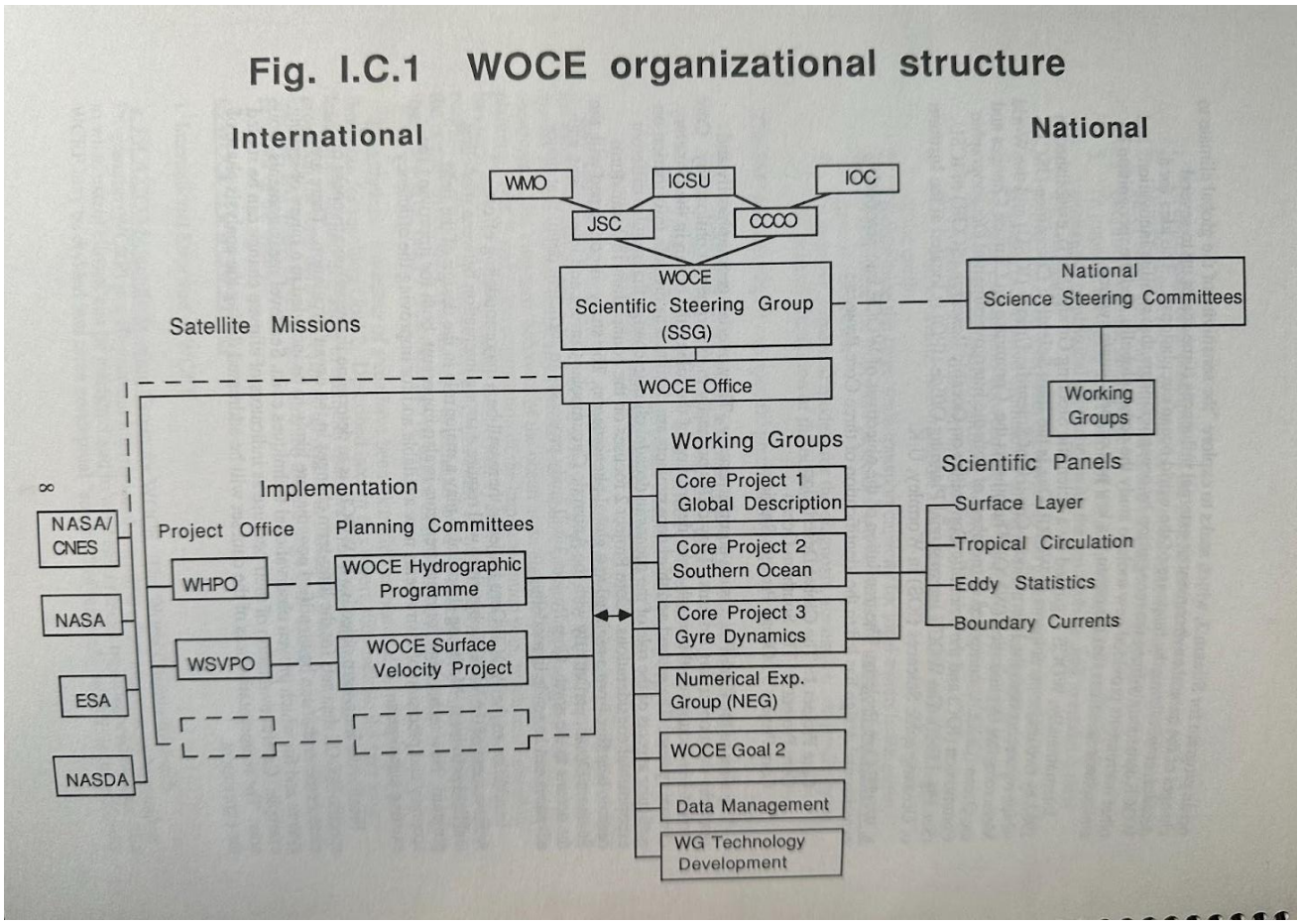
Note. This figure is from the inner jacket of WCRP (1986) Publication Series No. 6 WMO/TD – No. 122. It shows the major scientific objectives of the WCRP. WOCE was understood as critical to achieving objective c.

WOCE was a multinational effort that involved scientists from 30 countries (WOCE International Project Office, 1997, p. 2). The Experiment was coordinated under the WCRP with monetary contributions from at least ten individual countries (WOCE International Project Office, 1992, p. 22)[9]. In the U.S., WOCE was supported by major federal agencies such as the NSF, NOAA, the Department of Energy (DOE) and the National Aeronautics and Space Agency (NASA) (U.S. Planning Office for WOCE, 1990, pp. 4–6).

Organizationally, WOCE consisted of Scientific Steering Group, national committee, and planning offices (see Figure 7). The U.S. planning office was located at The University of Texas A&M. The planning phase took several years, with fieldwork officially starting in 1990 and concluding in 1997 (WOCE International Project Office, 1997, p. 2).. Data analysis, and modelling activities, including model-data integration and synthesis, continued until 2002 (WOCE International Project Office, 1997, p. 2). WOCE closed out with a conference in November 2002 in San Antonio, Texas (National Centers for Environmental Information, 2002). Of course, international support for WOCE was critical but planning for the Experiment was a community-driven "bottom-up" effort. Planning can be traced to the early to mid-1970s through various oceanography and GARP working groups, where community dialogue and planning started in the early to mid-1970s (Thompson et al., 2001, p. 33).

[9] Contributions to the WOCE International Project Office Operating Costs in 1992 from individual countries listed in the order of most, 271,000 U.S. dollars, to least, 1,000 U.S. dollars, contributions include: United Kingdom, United States of America, Canada, Germany, France, Japan, Australia, Netherlands, Spain, and Argentina. (see WOCE International Project Office, 1992, p. 22). The U.S. dollars are their values in 1992 and have not been adjusted for inflation. I note that contributions are from at least ten individual countries as it is unclear given the primary and secondary sources reviewed whether countries made recurring or one-off contributions, it is possible that throughout the course of WOCE that some countries made one-off contributions for one year but not in others.

Figure 7. From U.S. WOCE Implementation Plan (1988) WOCE's organizational structure

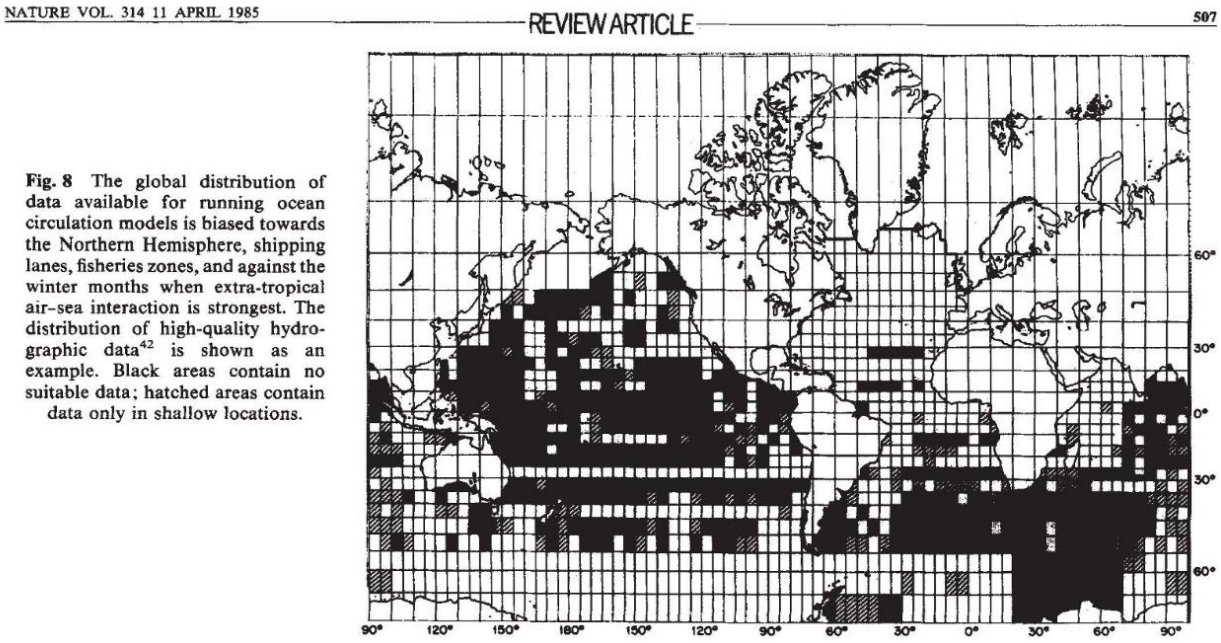


Note. Of interest to this dissertation one of the seven working groups is responsible for Data Management.

Oceanographers recognized that whilst there existed archival data that could contribute to the scientific aims of WOCE, it quickly became apparent that it would not be enough to look at extant data to move the scientific aims forward in any significant way (see Figure 8). This was because the data available were limited in their spatial and temporal coverage, there was uncertainty about the data's quality, and a fundamental misfit between the existing data and the data required to refine ocean models (Woods, 1985). Furthermore, with the exception of the Atlantic Ocean, observational ocean data were incredibly sparse. Given that almost all observational data were collected during the summer months which provided more favorable conditions for field research as compared with other times of the year. The Pacific and what

documents call “The Southern Hemisphere” were severely underdetermined (Woods, 1985). These areas of the oceans were incredibly difficult to study due to their remoteness and difficult conditions.

Figure 8. From Woods (1985) showing how existing oceanographic datasets were not suitable for WOCE



Note. From Woods (1985, p. 507), black areas indicate where no suitable data is available, and hatched areas show where data exists for shallow locations, but do not exist for deep locations.

Importantly, the community of oceanographers saw itself as technologically and methodologically ready to take on this massive endeavor. Several ocean expeditions were conducted in the 1970s, i.e., GATE, POLYGON-70, Mid-Ocean Dynamics Experiment (MODE), POLYMODE, that explored the mesoscale variability of the ocean through in-situ measurements, prototyping, refining, and developing instrumentation[10]. For instance, these

[10] In 1970 soviet scientists carried out POLYGON-70 using novel instruments such as moored current meters and CTDs later used in WOCE. The 1971-74 Mid-Ocean Dynamics Experiment (MODE) further tested observational tools including the newest moored current meter arrays and CTDs, which whilst much improved was reported to fail. MODE was also significant in that researchers conducting fieldwork and those who developed models collaboratively planned fieldwork. POLYMODE is described as a series of experiments conducted in the North

expeditions helped refine techniques for studying the deep ocean. Importantly, for WOCE, these research cruises demonstrated the feasibility of multilateral field campaigns (Woods, 1985).

Another advancement that elicited great excitement from oceanographers was NASA's 1978 Seasat satellite mission. This satellite was the first to be fitted with instruments capable of taking critical observations of the sea surface. Although it orbited Earth for a short duration - just 105 days before it failed due to the spacecraft electrical system short-circuiting - oceanographers rejoiced (Evans et al., 2005). Seasat was a successful proof-of-concept that earth's oceans could be studied from space.[11] This promised data and observations that would yield spatial coverage that could not be achieved through other methodological techniques. Seasat became the precursor to a series of ocean-observing satellites, TOPEX/POSEIDON (T/P), European Remote-Sensing Satellite-1 (ERS-1), European Remote-Sensing Satellite-2 (ERS-2), launched into the atmosphere during the 1990s into the early 2000s that would collect data for WOCE (U.S. Planning Office for WOCE, 1988; WOCE International Project Office, 1993). Moreover, as global ocean models, their development and improvement were the key scientific aim of the program, there was concern that existing computing technology was not able to do so during that time. In the late 1970s, the first Global Ocean Circulation Model was developed by researchers at Princeton University - so all was good to go.

As is the case with all projects, but made more acute by its large scale, when WOCE transitioned from paper to the field, unsurprisingly, many aspects did not go according to plan. Not all scientific aims were treated equally, with some scientific questions getting more attention and resources than others (Lehman, 2021). In the early 1990s, as WOCE was preparing to launch off the starting line, it was met with unexpected headwinds. The satellite launch of both ERS-1 and T/P were unexpectedly delayed by more than a year (Thompson et al., 2001; U.S. Planning Office for WOCE, 1990, 1988); funds and resources promised, or were being counted on by researchers and WOCE planners, had been scaled back

---

Atlantic by U.S. and Soviet scientists which highlighted the importance of eddies in the role of ocean circulation. The 1973 GATE cruise is presented in footnote no. 7 (Thompson et al., 2001).

[11] Woods (1985) writes about Seasat "the most spectacular advance in ocean observation has come from the demonstration by Seasat during a 100-day mission in 1979" (p.508). The enthusiasm for collecting observational ocean data is not just that it is from space, but rather from space, oceanographers will be able to collect global ocean data. Woods continues "the technique offered the prospect of a truly global perspective of the ocean circulation for the first time" (p.508).

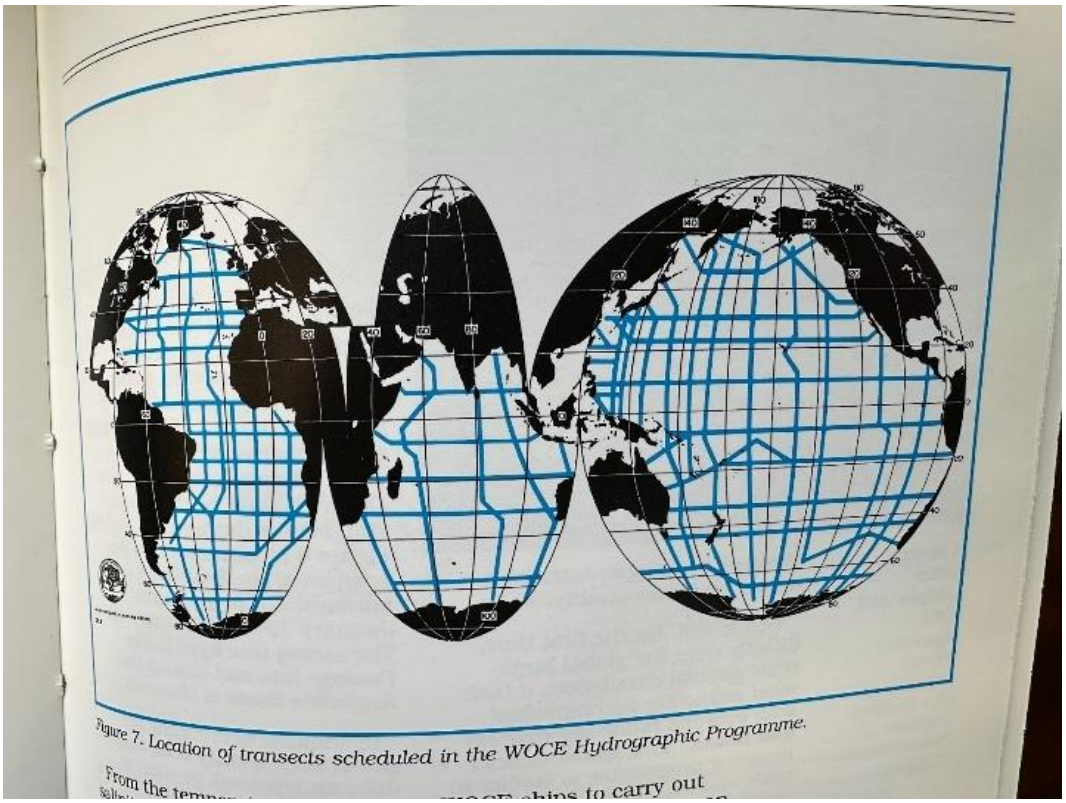
substantially (U.S. Planning Office for WOCE, 1990, pp. A2–A3)[12]. One of the key staff at the planning office overseeing data management had retired and the SSG were having a hard time finding a suitable replacement. For a while there in the early days, the U.S. planning office for this global experiment was staffed by a "one-man operation" (WOCE International Project Office, 1992, p. 19).[13]

When data collection began in earnest, the vast majority of WOCE PIs were not sharing their data or were sharing it behind mandated timelines. The delays were so substantial that the SSG thought that this would be one of those issues that would impact the integrity of the Experiment. Researchers were not the only source of data sharing issues, satellite data from ERS-1, was difficult to access as the European Space Agency hoped to sell it one day, and therefore was not eager to share it (WOCE International Project Office, 1992). There was genuine anxiety that the lack of resources and all these unexpected challenges would irreparably hinder the scientific aims that WOCE set out to achieve.

[12] In the 1990 U.S. WOCE Implementation Plan there is a section in the executive summary titled "Effect of Federal Budget Limitations on U.S. WOCE" in which the report notes "The implementation of the U.S. contribution to WOCE is underway. The federal funding levels upon which U.S. WOCE was planned, however, have not materialized, and U.S. WOCE is proceeding with very restricted funding levels. This may jeopardize the planned cohesive, synoptic "snapshot" of the ocean circulation that is critical to the achievement of WOCE goals. The U.S. WOCE [Science Steering Committee] SSC believes that if the funding situation can be improved substantially in 1991 and beyond, then the integrity of U.S. WOCE can be recovered and maintained. The SSC is hopeful that its efforts, couple with those of the Interagency Panel for U.S. WOCE, will achieve the needed funding enhancements" (p. A2)

[13] In 1992, staffing issues involved a key personnel retiring and difficulties finding someone to fill the position. This issue was further compounded by the lack of funds to hire a replacement. However there was optimism for monetary support from Canada but even if the funds were to be provided, it would not be available until a year later in 1993: "The staffing of the [WOCE Hydrographic Program Office] WHPO remains a problem as M. Stalcup, who retired, has not been replaced, nor are there funds for a replacement. Swift strongly urged the SSG to seek support for this position which is especially critical for the interactions between investigators and the [Data Quality Expert] DQE team. WHPO efforts to obtain a secondment for this position have gone slowly. There is some possibility of Canadian support but not until 1993 or later. The WHP [Special Analysis Center] SAC is essentially a one-man operation." (WOCE International Project Office, 1992, p. 19)

Figure 9. From WCRP (1991) showing locations of planned WOCE research cruises



Note. From WCRP (1991, p. 23) showing the planned research cruises in WOCE's hydrographic Programme and their extensive coverage of the earth's oceans

For all the worries about what WOCE could've achieved, what it did achieve was executing an experiment that generated unprecedented spatial and temporal scope for oceanographic observations (see Figure 9). Just shy of the 21st century, oceanography had, through WOCE, developed institutions, infrastructures, methods, and technologies deemed appropriate for the subject of its discipline: the global ocean. Global oceanography had, at long last, come of age.

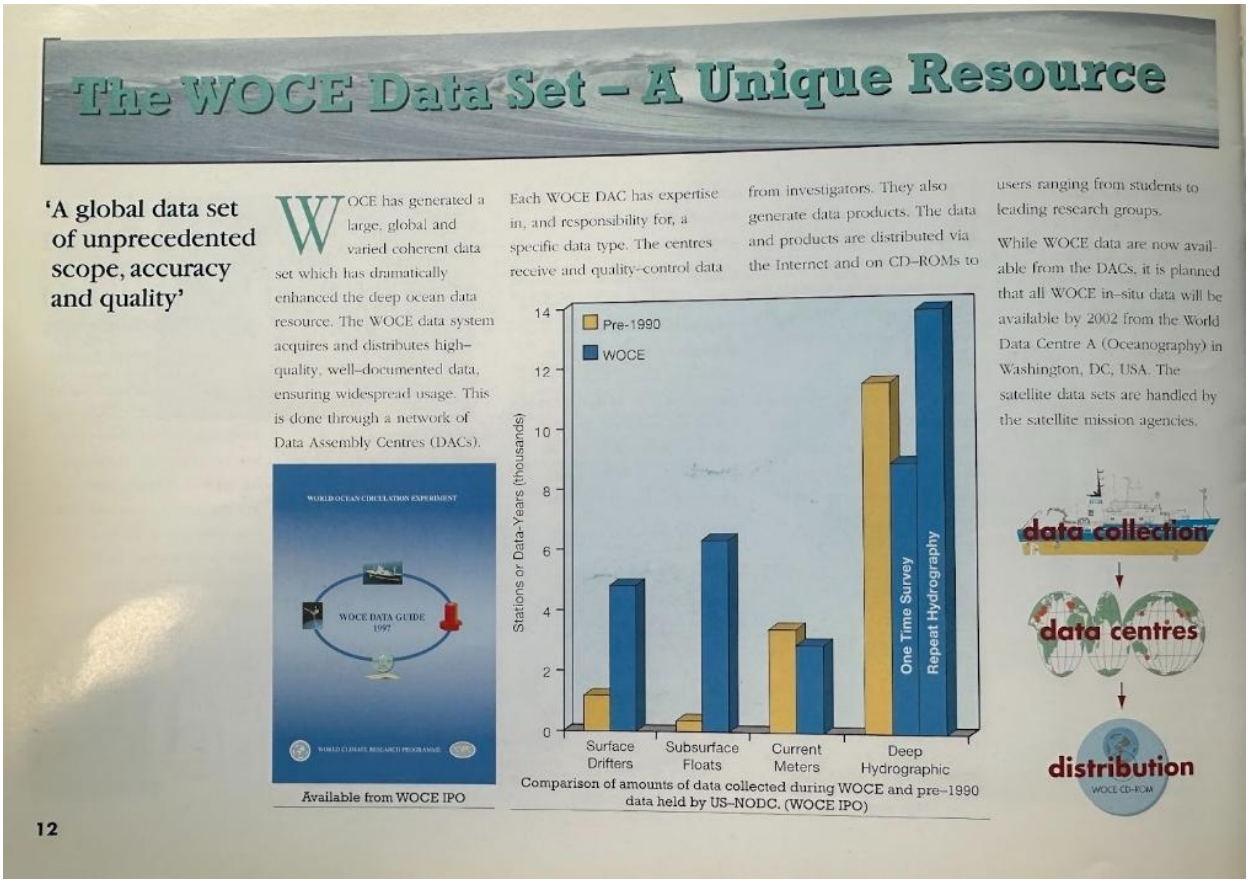
4.3 The origins of a time norm: data management and data sharing in WOCE

The previous section focused primarily on the planning and lead up to WOCE and some of the issues encountered as the data collection phase began. As the conclusion to the previous section foreshadowed, some of the issues experienced concerned the lack of data sharing. This

section focuses on WOCE's data management system, data sharing policy and data sharing in practice.

Interestingly, summary reports and publications on WOCE that came out during the tail end of the experiment would hail the data management system and data sharing as a resounding success. After all, WOCE set out to produce a global ocean data set, and not only did it do so, but it produced one of "unprecedented scope, accuracy and quality" (WOCE International Project Office, 1997, p. 12), also see Figure 10. To achieve this feat, "PIs were asked to submit their data within the unprecedented short period of 2 years following the fieldwork. In some cases, even a 6-month period was suggested - unrealistic perhaps at the time but now at the end of WOCE and the beginning of new field programmes not at all so. Such is the change of attitude that has been fostered" (Thompson et al., 2001, p. 41). One question then is, how did WOCE achieve this feat?

Figure 10. From WOCE International Project Office (1997) Overview of WOCE Data Set



Note. From WOCE International Project Office (1997, p. 12) An Overview of the global and unique WOCE Data Set

4.3.1 Envisioning and Implementing the WOCE Data Management System

The observational data to be collected through WOCE was always understood to be a core component of the Experiment, and not just a byproduct. After all, the Experiment hinged on being able to collect global data. However, data could only become global when compiled into a cohesive dataset and have that dataset does not remain in the hands of the PIs who collected it, but it required that it be shared with WOCE PIs who would take that data to create parameters for ocean models and/or to test their theoretical ocean models against observational data. For this to happen, data management was critical, "There will be intensive scientific involvement in the development of WOCE data management practices. Since the primary purpose is to provide data needed to meet the scientific objectives of WOCE, it is essential that data management and scientific planning be integrated" (US WOCE SSG 1998, p.126-7).

But much like today, whilst oceanographers recognized the importance of data management during their planning efforts and on paper, in practice, the fact of the matter was that data management was boring. Roger Revelle, who was at the time the chair of the Intergovernmental Oceanographic Commission's (IOC) Committee on Climate Change and the Ocean (CCCO) remarked playfully that "*data management was one of the most important problems facing both WOCE and [Tropical Ocean and Global Atmosphere Programme] TOGA and that researchers needed to ensure that it was done properly, even though the topic bored him to tears!*" (Thompson et al., 2001, p. 40). As the work of science was being done all over the world's oceans by teams of scientists hailing from more than 30 countries, there would be seven different streams of data, and add on top of that that the experiment itself would take multiple years to complete - a data management system and infrastructure was required.

Figure 11. From U.S. WOCE Implementation Plan (1988) Cost Estimates Summary

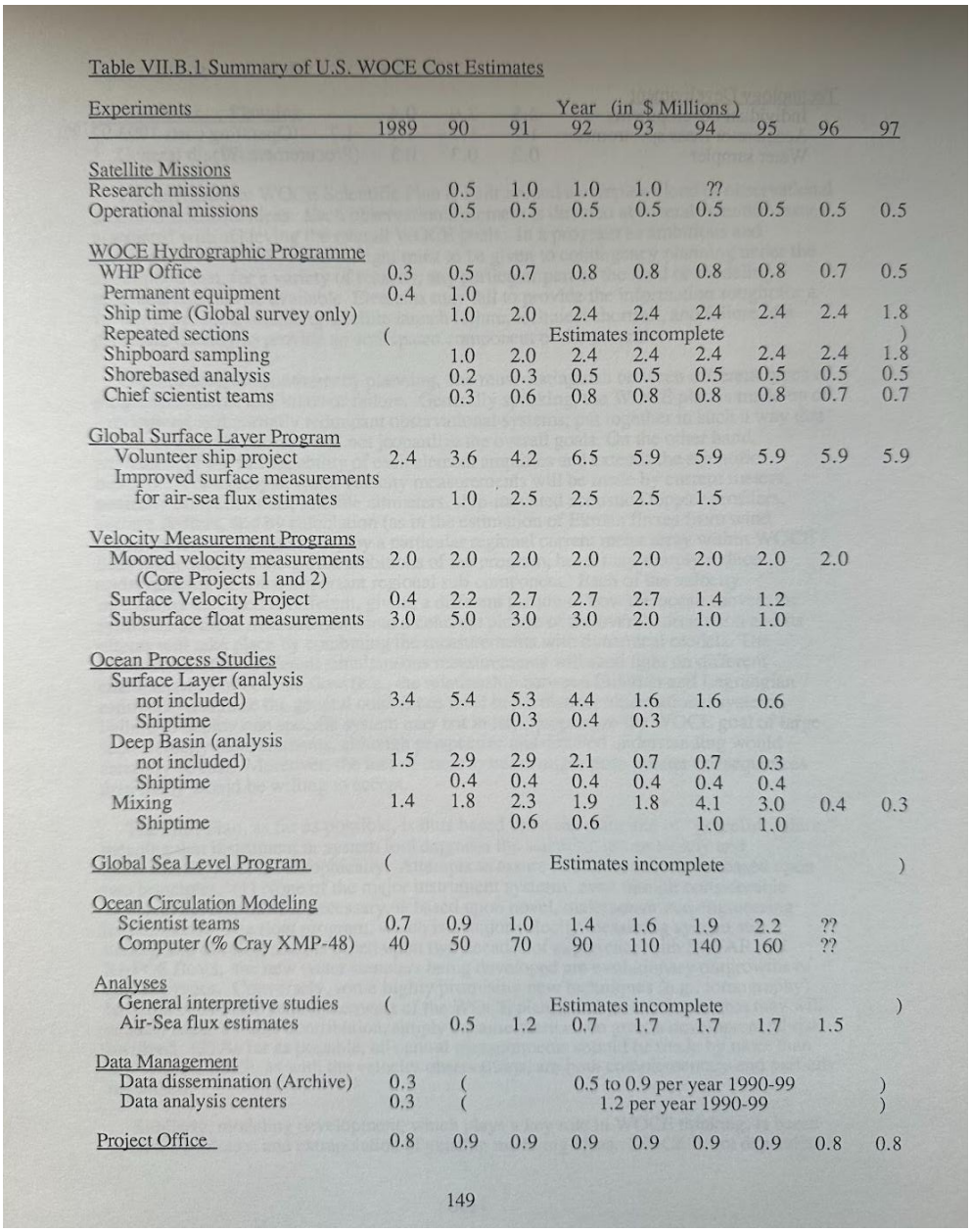


Table VII.B.1 Summary of U.S. WOCE Cost Estimates

Experiments	Year (in $ Millions)								
	1989	90	91	92	93	94	95	96	97
Satellite Missions									
Research missions		0.5	1.0	1.0	1.0	??			
Operational missions		0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
WOCE Hydrographic Programme									
WHP Office	0.3	0.5	0.7	0.8	0.8	0.8	0.8	0.7	0.5
Permanent equipment	0.4	1.0							
Ship time (Global survey only)		1.0	2.0	2.4	2.4	2.4	2.4	2.4	1.8
Repeated sections	(			Estimates incomplete					)
Shipboard sampling		1.0	2.0	2.4	2.4	2.4	2.4	2.4	1.8
Shorebased analysis		0.2	0.3	0.5	0.5	0.5	0.5	0.5	0.5
Chief scientist teams		0.3	0.6	0.8	0.8	0.8	0.8	0.7	0.7
Global Surface Layer Program									
Volunteer ship project	2.4	3.6	4.2	6.5	5.9	5.9	5.9	5.9	5.9
Improved surface measurements for air-sea flux estimates		1.0	2.5	2.5	2.5	1.5			
Velocity Measurement Programs									
Moored velocity measurements (Core Projects 1 and 2)	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	
Surface Velocity Project	0.4	2.2	2.7	2.7	2.7	1.4	1.2		
Subsurface float measurements	3.0	5.0	3.0	3.0	2.0	1.0	1.0		
Ocean Process Studies									
Surface Layer (analysis not included)	3.4	5.4	5.3	4.4	1.6	1.6	0.6		
Shiptime			0.3	0.4	0.3				
Deep Basin (analysis not included)	1.5	2.9	2.9	2.1	0.7	0.7	0.3		
Shiptime		0.4	0.4	0.4	0.4	0.4	0.4		
Mixing	1.4	1.8	2.3	1.9	1.8	4.1	3.0	0.4	0.3
Shiptime			0.6	0.6		1.0	1.0		
Global Sea Level Program	(			Estimates incomplete					)
Ocean Circulation Modeling									
Scientist teams	0.7	0.9	1.0	1.4	1.6	1.9	2.2	??	
Computer (% Cray XMP-48)	40	50	70	90	110	140	160	??	
Analyses									
General interpretive studies	(			Estimates incomplete					)
Air-Sea flux estimates		0.5	1.2	0.7	1.7	1.7	1.7	1.5	
Data Management									
Data dissemination (Archive)	0.3	(		0.5 to 0.9 per year 1990-99					)
Data analysis centers	0.3	(		1.2 per year 1990-99					)
Project Office	0.8	0.9	0.9	0.9	0.9	0.9	0.9	0.8	0.8

149

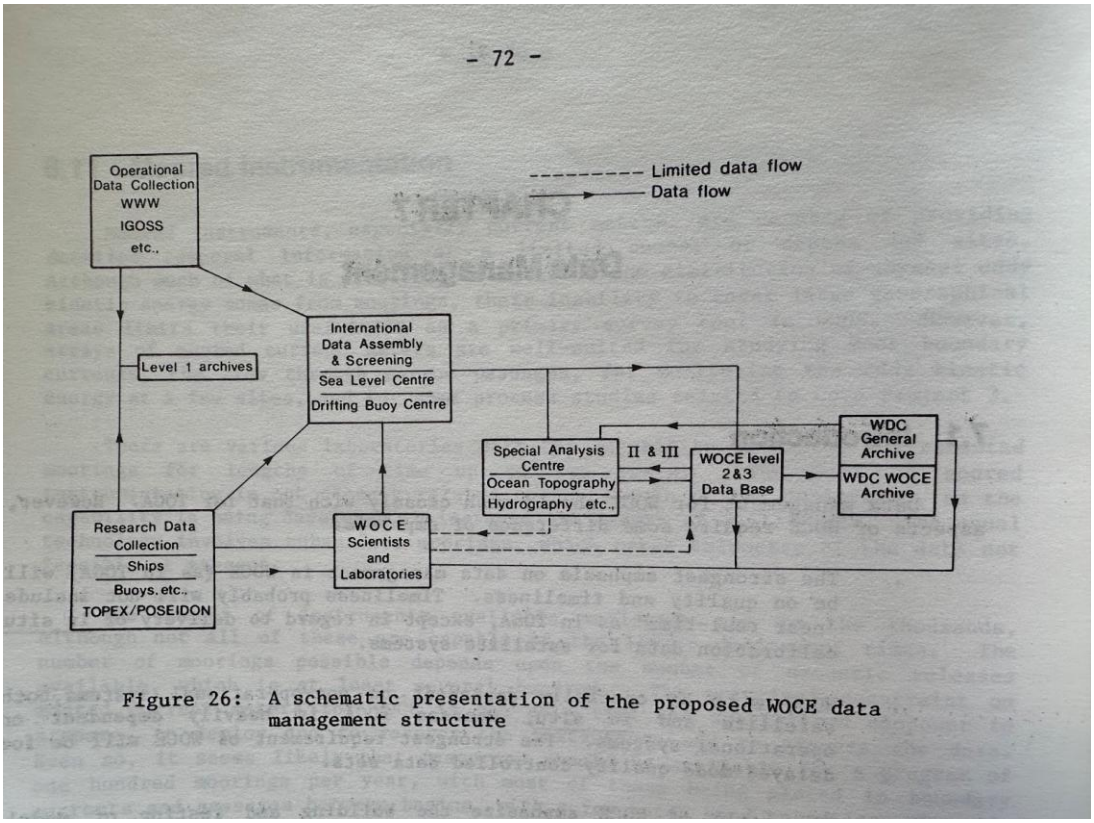
Note. From U.S. WOCE Implementation Plan (1988, p. 149) Of interest, U.S. Cost Estimates Summary for data management over the duration of the experiment was estimated at between 17.6 – 21.6 million U.S. dollars before adjusting for inflation

In the early stages of the planning process, it became clear that data management and data sharing were key to WOCE's ability to meet its scientific goals. Organizationally, WOCE

created a dedicated Working Group for data management (see Fig 6) and in early and final versions of the WOCE implementation plans, a dedicated budget was set aside for data management. The cost estimates in the budget on data management alone throughout WOCE was estimated at between 17.6 – 21.6 million U.S. dollars annually (See Figure. 11)[14]. The funds would include hiring 22 dedicated personnel for this task (compared to previously 0 people working in this field), the largest percentage increase of personnel across all hiring categories (see Table VIII.B.4: Estimates of Additional Manpower Requirements for WOCE in U.S. Planning Office for WOCE, 1988, p. 160). In addition, recognizing that the data flows and data system would have to be in place before any data was ready for sharing, data management activities were planned to start at the beginning of the Experiment. In the mid to late 1980s a pilot of one key piece of the data system was begun at the University of Delaware.

[14] The following is from the U.S. WOCE first implementation plan. "Costs for the "Archive" functions, the U.S. Data Management Unit and data dissemination activities to handle WOCE data are estimated at between $500K to $900K per year for the period 1990-1999; costs for 1989 might be $300K. The U.S. funding needed for WOCE data analysis centers is estimated by assuming a total of six U.S. centers at an average cost of $200K/yr." (U.S. Planning Office for WOCE, 1988, p. 137). Note that the dollar amounts have not been adjusted for inflation. K denotes 1,000 U.S. dollars, e.g. 1K is 1,000 U.S. dollars.

Figure 12. From WMO (1986) Diagram of planned WOCE data system



Note. From WMO (1986, p. 72) showing an early draft of the WOCE data management system showing the envisioned data flows between scientists, archives and other key organizational structures like the Data Assembly Centers

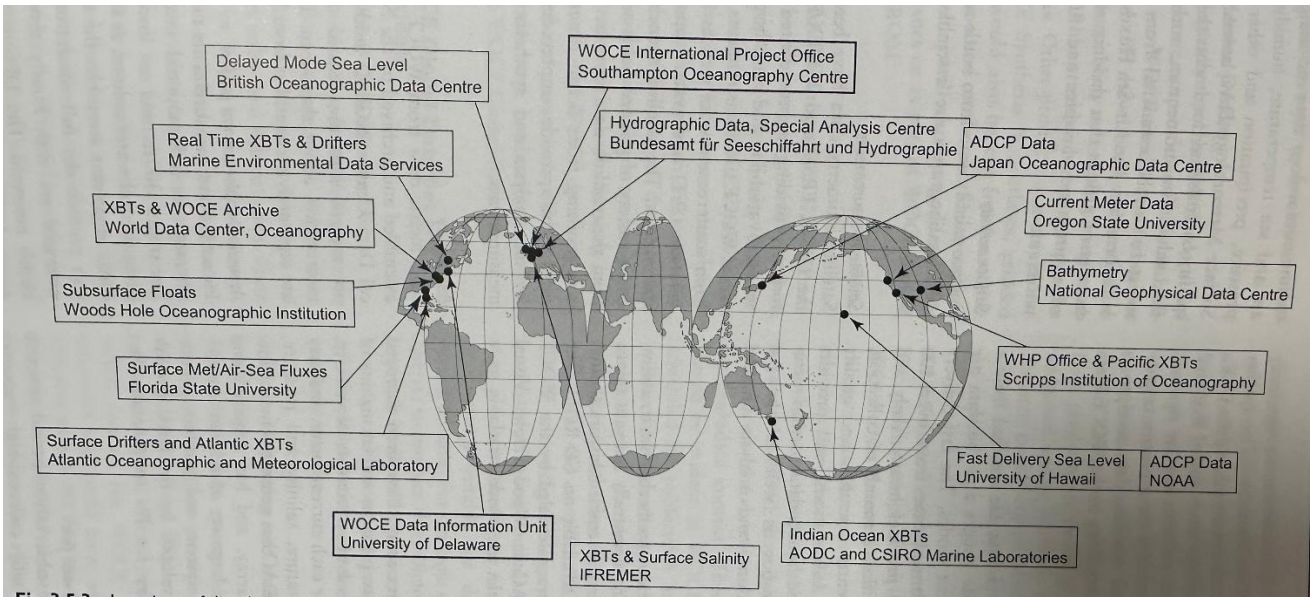
Desirable qualities that the envisioned data system was expected to enable was high-quality and timely data delivery and sharing. High-quality data required that the data be reported in a standardized manner, that information was to be provided about the quality of the data, and that sufficient metadata (appears in primary WOCE documents as "environmental data" and "ancilliary data") be included alongside the data. It was also important that no data be lost and as such archives or long-term data stewards required identification. But how were WOCE to do achieve this?

WOCE, and other large-scale oceanography research programs of the day, such as TOGA, were strongly influenced by a 1982 National Research Council "*Data Management and Computation: Volume 1: Issues and Recommendations*" authored by the Space Science

Board's Committee on Data Management and Computation (CODMAC). In primary WOCE sources such as Nowlin (1985) and Webster (1985) this report is referred to as the CODMAC report. Through evaluation and cross-case study comparison of seven space science missions, CODMAC articulated foundational principles for data management that remain the bedrock of present-day data management and data sharing policies (National Research Council, 1982). In the early planning stages, WOCE data management meetings happened together with its sister program TOGA (Nowlin, 1985; Webster, 1985). In search for best practices for data management the two programs reviewed available literature at the time and drew heavily on the CODMAC Report (Nowlin, 1985; Webster, 1985). Key principles articulated by the report included the importance of scientists' involvement from planning which data to be collected, from managing the data during the active phases of research, all the way to ensuring data's dissemination and archival. Documentation was highlighted as critical for data to be useful to researchers who were not a part of the original research.

These principles can be seen in WOCE, where planning documents discussing data management articulated the importance of metadata, documentation, and standards - key elements that we understand today as critical to infrastructural development "In addition, the system should set standards for data products, processing, and dissemination … Backup datasets should be preserved. Data loss due to the decay of storage media must be prevented." (US WOCE SSG, 1998, p126). In the 1988 draft implementation plan the goal of data management was to ensure that "The reception of US-WOCE data packages from the producers in a timely and useful manner (data package = data + associated quality control flags + processing history + sensor identification, etc.)" (US WOCE SSG, 1998, p.126). This would be a change for oceanographers as they would be required to perform additional data management work in addition to the work of collecting data "The WOCE data management system will have a number of attributes that differentiate it from conventional oceanographic data systems. Data must be 'tagged' to quality information and descriptions of its collection, processing, calibration and validation." (US WOCE SSG, 1998; p126).

Figure 13. From Lindstrom & Legler (2001) Location and institutions in WOCE's data management and sharing system



Note. This diagram from Lindstrom and Legler (2001, p. 187) shows the locations of institutions that formed the WOCE data management and sharing system as well as the types of data that the institution was responsible for conducting quality control

Key components of the WOCE data systems included Data Analysis Centers (DACs), National data dissemination centers (NDCs), Specialty data dissemination centers (SACs), and a "control tower" the WOCE Data Management Unit (DMU) (see Figures 12 and 13). The DMU, as the metaphor control tower alludes to, would oversee the WOCE data management network. It was recognized that it would be intractable to create standards, prescribe data flows, and ensure the timely delivery of the diverse data streams. As such WOCE SSG set the overall vision and scope of US WOCE data management, each of the various "'programs' (such as hydrography, velocity, sea level, etc.)" were seen as better positioned to and instructed to "set standards for where to measure, what to measure, and how well to do so," (U.S. Planning Office for WOCE, 1988, p. 127).

The WOCE Data Management Working Group was to "coordinate and control" the data management activities occurring at the program levels and ensure that data was being delivered by PIs to the appropriate DAC/SAC, NDC in a timely and useful manner, ensuring the archival and sharing of data packages to other oceanogrpahers in a timely and useful

manner, ensuring that quality control had been performed to US-WOCE standards, gather and share information on the state of the US and International programs, providing a timeline essentially of which data are where and what stages of the data lifecycle they are in as well as collecting other non WOCE data products that may be useful to WOCE. (US WOCE SSG, 1988, p 129).

4.3.2 The evolution of WOCE's data sharing policy: Establishing the Two-Year Data Sharing Norm

Delivering data in a "timely" fashion was one of the key objectives of the WOCE data management system, as discussed earlier, because data needed to be quality controlled and standardized by DACs and subsequently passed onto the modelers for their modeling work. As such, a data sharing policy was formally included since the first draft of U.S. WOCE's implementation plan that articulated this priority clearly: "As a requirement of participating in U.S. WOCE, institutions will agree to calibrate, document and pass in a timely way to the appropriate centers," (U.S. Planning Office for WOCE, 1988, p. 136).

Although the task was framed as a requirement, the data sharing policy, read in its entirety, made it clear that it recognized that this would be an unpopular ask. The text in the plan recognized that many PIs would be reluctant to do so as a culture of broad data sharing, and certainly not at the speed desired by WOCE, was not a "traditional practice." Furthermore, the implementation plan recognized that at the time, PIs who collected data viewed the data as inherently belonging to them. Naturally, as career progression is dependent on publishing novel research, supported by data, there were concerns that sharing data would open up PIs who did so to having their work scooped.

To assuage these concerns and "ensure that the rights of data collector are not violated", WOCE data sharing policy stated that "Publication of analyses based solely on these data may not occur for **two years after data collection**. Any publication using these data must attribute the originating Principal Investigator as the source, WOCE [DACs] will not release data to a National Data Distribution Center without prior permission of the originator." (US WOCE SSG, 1988, p.136 - emphasis added).

WOCE settled on a two-year exclusive use "embargo" or "publication rights" period after data collection to "recognize the traditional rights of principal investigators to analyze data they have acquired." (US WOCE SSG, 1988, p.136) In other words, the PIs who

collected that data were to get "first dibs" on that data. But after a certain amount of lead time, in this case two years, that data had to be shared and distributed amongst WOCE colleagues. The data sharing policy emphasized that "a program such as WOCE can only succeed if these traditional practices are modified so that data become available to other scientists within the program more quickly than ever before." (US WOCE SSG, 1988, p. 136)

Although the select primary sources and secondary literature surveyed do not explicitly provide an explanation for why a two-year time period was chosen, what the sources do shed light on is that: first, this time duration was unprecedented as it was recognized as a departure from "traditional" oceanographic practices. Second, this departure from tradition was viewed as necessary. WOCE was not traditional oceanography; it was aiming for global oceanography, and as such it required new research practices. Third, the two-year duration was seen as a good way to balance incentives for WOCE PIs doing data collection on the one hand and for WOCE scientific goals on the other hand. WOCE PIs needed time to conduct fieldwork, verify and analyze data in preparation for publication, and two years would give PIs enough of a head start and was viewed as respecting their "publication rights." (WOCE Report No. 67/91 Rev. 1; WOCE Report No. 67/91 Rev. 2)

For the larger WOCE project, the time duration needed to be short enough to jump-start the modelling phase of the project and avoid data loss as "the more time that passes between the time of data collection and the assembly of the data and metadata at a DAC, the more likely it is that problems or questions about the data will remain undocumented or unresolved," (Lindstrom & Legier, 2001, p. 184). With all the technical, organizational, and infrastructural elements of the data sharing system in place, all that was left was for the data to be shared. The next section presents the data sharing in practice in WOCE.

4.3.3 Data sharing policies in the wild

In later reports, data management and data sharing were presented as an unprecedented "win" (e.g. WOCE International Project Office, 1997); however, the reality was much more difficult. One thing that was done right was that for data management and sharing to happen, it was recognized that a lot of infrastructural building and infrastructural work to make the envisioned goal a reality.

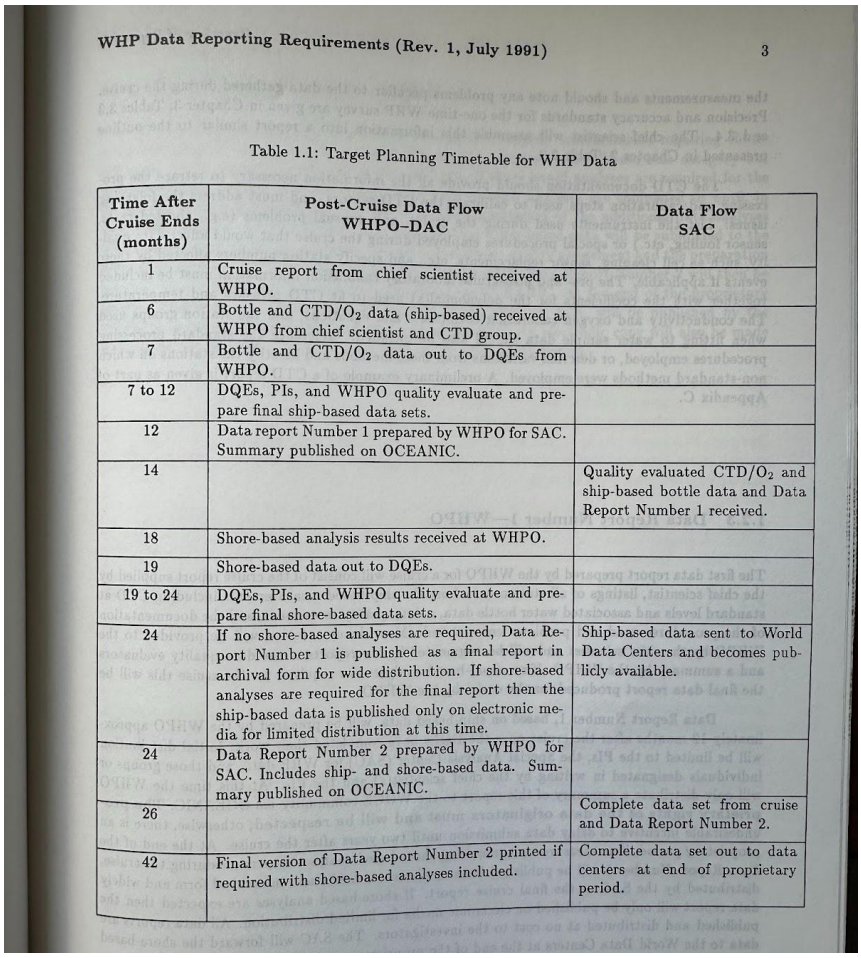
To support data management and sharing in WOCE, seven DACs were established, each tasked with verifying and quality-controlling a specific data type (U.S. Planning Office

for WOCE, 1988). Each DAC was overseen by a domain expert with experience working with that data type. Coordinating these centers was the Data Information Unit (DIU), which functioned as the data management system's control tower. The DIU was responsible for cataloging WOCE data on the web, tracking their status, and serving as a coordination hub for the broader data management and sharing system (U.S. Planning Office for WOCE, 1988).

Launched in 1994, the DIU's website was among the first few hundred on the World Wide Web, with the DAC's websites following later (Thompson et al., 2001). Over time, the DIU's role evolved from serving as a central source for national and internationl planning documents to summarizing experiment's progress and linking to datasets and DACs, and finally to supporting WOCE's concluding phase focused on analysis and modelling (Thomspon et al., 2001).

Infrastructural work, included the various WOCE Operations Manuals (Joyce et al., 1994; WHP Office, 1991)that established standards for collecting and reporting different types of oceanographic data. To name a few, this included developing "quality flags", what "ancillary information" i.e. metadata to document, as well as establishing norms for when different types of data were to be made available and to which organization to share it with (see Figure 14) (U.S. Planning Office for WOCE, 1988).

Figure 14. From WHPO (1991) Showing time for data sharing after collection



WHP Data Reporting Requirements (Rev. 1, July 1991) 3

Table 1.1: Target Planning Timetable for WHP Data

Time After Cruise Ends (months)	Post-Cruise Data Flow WHPO–DAC	Data Flow SAC
1	Cruise report from chief scientist received at WHPO.	
6	Bottle and CTD/O_2 data (ship-based) received at WHPO from chief scientist and CTD group.	
7	Bottle and CTD/O_2 data out to DQEs from WHPO.	
7 to 12	DQEs, PIs, and WHPO quality evaluate and prepare final ship-based data sets.	
12	Data report Number 1 prepared by WHPO for SAC. Summary published on OCEANIC.	
14		Quality evaluated CTD/O_2 and ship-based bottle data and Data Report Number 1 received.
18	Shore-based analysis results received at WHPO.	
19	Shore-based data out to DQEs.	
19 to 24	DQEs, PIs, and WHPO quality evaluate and prepare final shore-based data sets.	
24	If no shore-based analyses are required, Data Report Number 1 is published as a final report in archival form for wide distribution. If shore-based analyses are required for the final report then the ship-based data is published only on electronic media for limited distribution at this time.	Ship-based data sent to World Data Centers and becomes publicly available.
24	Data Report Number 2 prepared by WHPO for SAC. Includes ship- and shore-based data. Summary published on OCEANIC.	
26		Complete data set from cruise and Data Report Number 2.
42	Final version of Data Report Number 2 printed if required with shore-based analyses included.	Complete data set out to data centers at end of proprietary period.

Note. This target planning timetable for WHP Data shows timeframes in months for when different data was to be delivered to the designated DAC or SAC after the end of a research cruise

Overall, the infrastructural support for data management and sharing was quite strong. However, in the early years of the field campaign phase, meeting notes from the WOCE SSG meeting attendees lamented how up to 80% of the cruises missed submission deadlines (WOCE International Project Office, 1992). Far from being an issue limited to cruises, or even a particular data stream, this delay in timely data sharing was rampant across the board. It was continually noted in subsequent reports such as the WOCE report on Data management that PIs were “delinquent” in data sharing (WOCE International Project Office, 1993, p. 16).

In the preface to the WOCE Operational Manual 1994 Rev 2, written 2.5 years after the first Rev. James H. Swift, The Chairman of the WOCE Hydrographic Program (WHP) Planning Committee wrote an impassioned plea to colleagues to manage their data, prepare it for sharing and share it. Swift wrote about the challenges of doing data management to the standards required by the program:

> *"I am a WHP chief scientist, too, and I have grappled with this document. The file types, formats, units, documentation structure, and the role of the chief scientist as chief of all data from the expedition were new to me. It took several iterations to get my files in the correct format, I have not yet completed merging the tracer and hydrographic data files, and I was asked by the WHP Office to reformat the documentation I sent with my data. This has not been easy for me, and so I assume that it is not easy for many others."* (Joyce et al., 1994, p. viii)

It also appears that the issue of data sharing, or a lack thereof, weighed heavily on Swift, who issued an impassioned plea to colleagues to share their data because the success of WOCE depended on it. WOCE was seen as an once-in-a-lifetime opportunity, and that PIs sharing data on their "own good time" was just not good enough:

> *"A much more important issue is the availability of WOCE data to WOCE scientists. One of the major and most central issues we face is that the people who are doing WOCE research must have access to WOCE data. This concept is not difficult, yet there has been a tendency to obscure this basic issue. Yes, it is important that the data be complete and of high quality, that they be documented, and that they be archived, but we are doing this experiment - WOCE - to learn something about the ocean. We can't do that without data. People must not be afraid to share their data… But we must understand that our own good time is probably not soon enough for this big joint project. In order to get the most out of a very large experiment we must work together at least in terms of data sharing ... I urge each scientist to think of how it might be resolved"* (Joyce et al., 1994, p. ix)

Compliance mechanisms to ensure that WOCE's data sharing requirements were followed were few and feeble. Formal compliance mechanisms came down to "peer pressure and funder persuasion for delayed submissions" (Lindstrom & Legler, 2001, p.184). Despite this, in the final 2002 conference in San Antonio, Texas, a presentation titled "WOCE Global Data V3" notes that 10% of collected WOCE data will never be shared (Bindoff & Legler, 2002).

To cut a long story short, WOCE's legacy can still be in U.S. Oceanography today. By chance, one of the PIs I interviewed is a colleague of Carl Wunsch, a central figure and member of the science steering group, who noted the following

> "*And so what the WOCE did is it tried to organize that [data sharing] in some way. It tried to say to investigators when you collect data somebody's paying for it. Often the Federal Government's paying you money to go and collect it. It's your responsibility to make it publicly available to everybody. It's a moral thing, but also a scientific thing, because if you can collect lots of data not just your own, but make use of everybody else's then you get a better view of everything*" (P33).

Of interest to this dissertation, the NSF data management and sharing policies in OCE have carried on this time norm for data sharing and availability, which I turn to next.

4.4 The persistence of WOCE's data sharing time norm in OCE data policies

This section examines how time and timely access to data is discussed in OCE DMP guidance documents. Official NSF guidance for oceanographic data, remarkably, has existed since the late 1980s, more than 20 years before the 2011 DMP mandate. For oceanographers then, whilst writing a DMP as part of the proposal is a relatively new requirement, data sharing is not. OCE policies provide a concrete time for when they expect research data to be made available, "two (2) years after data collection" and can be traced back to the data sharing policies from WOCE (see Table 5.). When compared to the other NSF data policies studied in the larger Data Afterlives project, OCE is the only Division to note a specific time to data availability (Tian et al., 2021).

This section discusses the persistence of the 2-year time to data availability as well as presents an overview of the NSF OCE data policies, their similarities and differences. The first guidance document was established in 1988. The second of such documents, from 1994, titled "*Policy for Oceanographic Data, NSF 94-126*", remains easily accessible today. At the time of writing, OCE produced six guidance documents from 1998 to 2024. Updated guidance documents are introduced rather sporadically. In some instances, a couple of years pass by between guidance documents, other times, a decade passes before an update is made, for example, in the case of the 1994 and 2004 guidance documents. Interestingly, of the six, three were authored pre-DMP mandate era. This raises the question about what data management

practices were already advocated by NSF OCE pre 2010 mandate and which can be attributed to the DMP mandate of the 2010s.

Table 5. Time in WOCE and OCE Data Policies

Year	Author	Document Name	Time to data availability	Document Number
1988	U.S. Planning Office for WOCE	U.S. WOCE Implementation Plan: First Draft	"WOCE data management planning recognizes the traditional rights of principal investigators to analyze data they have acquired, but a program such as WOCE can only succeed if these traditional practices are modified so that data become available to other scientists within the program more quickly than ever before. To ensure that the rights of data collector are not violated, these data should be made available with the following understanding: …. **Publication of analyses based solely on these data may not occur for two years after data collection**... Spurred by the needs of WOCE and other geoscience programs, NSF has been working with other federal agencies to establish a common policy." (p.136 – emphasis added)	Not applicable
1990	U.S. Planning Office for WOCE	U.S. WOCE Implementation Plan 1990	"In accordance with the general data management principles of WOCE … data [will be reported and provided] to the U.S. [National Oceanographic Data Center] NODC for general release **two years after the time of measurement completion.**" (p.17 – emphasis added)	U.S. WOCE Implementation Report Number 2
1993	WOCE International Project Office	WOCE Data management: Data Policy and Practices,	"Any data collected as part of WOCE should be **made publicly available no later than 2 years (the publication rights period) from collection**, unless specifically waived by the	WOCE Report No. 104/93

		Data Assembly and Analysis Centres, Satellite Data Availability and Data Information Unit	[Science Steering Group] SSG and funding agencies." (p.3 – emphasis added)	
1994	NSF	Policy for Oceanographic Data	"Principal investigators are required to submit all environmental data collected to the designated national data centers **as soon as possible, but no later than two (2) years after the data are collected.**"	NSF 94-126
2004	NSF	Division of Ocean Sciences Data and Sample Policy	"Principal Investigators are required to submit all environmental data collected to the designated National Data Centers (Appendix I) **as soon as possible, but no later than two (2) years after the data are collected**."	NSF 04-004
2011	NSF	Division of Ocean Sciences Data and Sample Policy	"PIs are required to submit, at no more than incremental cost and **within a reasonable time frame (but no later than two (2) years after the data are collected),** the primary data, samples, physical collections and other supporting materials created or gathered in the course of work under NSF/OCE grants to the appropriate Data Center."	NSF 11-060
2016	NSF	Division of Ocean Sciences (OCE) Sample and Data Policy	"The Division of Ocean Sciences requires that metadata files, full data sets, derived data products and physical collections **must be made publicly accessible within two (2) years of collection**. This includes software and derived data	NSF 17-037

			products (e.g., model results, output, and workflows)."	
2024	NSF	Division of Ocean Sciences (OCE) Sample and Data Policy	The Division of Ocean Sciences requires that meta data files, full data sets, derived products and physical collections **must be made publicly accessible upon publication, or within two (2) years of collection, whichever comes first.** This includes software and derived data products (e.g., model results, output, and workflows).	NSF 24-124

4.4.1 Pre-2010 DMP mandate OCE data policies

OCE's earliest guidelines for data were issued by NSF in 1988. Although the original document is unavailable, later policies refer to it directly. The 1994 "Policy for Oceanographic Data" (NSF 94-126), following WOCE's data sharing policy, asked that PIs submit data to national centers "as soon as possible, but no later than two years after collection." This timeframe for data availability has persisted across all subsequent policies, including in 2004, 2011, 2016, and 2024. Since at least 1994, metadata has also been expected within 60 days of collection.

OCE policies consistently pair this timeframe for data availability with data infrastructures. Across the two decades, five national archives have remained central and serve as the backbone for ocean data access and preservation.[15]

OCE policies also direct PIs funded under particular Research Programs, such as WOCE, to follow the program-specific data policies. The program-specific data policies reflect OCE's broader guidance and general philosophy but provide more detailed guidance. Over time, the list of such programs changed, with well-funded initiatives like U.S. Global Ocean Ecosystem Dynamics Research (GLOBEC) and U.S. Joint Global Ocean Flux Study (JGOFS) establishing their own Data Management Offices (DMOs). The Biological and Chemical Oceanography Data Management Office (BCO-DMO), now a cornerstone of OCE-funded data sharing, emerged from the merger of these earlier DMOs (BCO-DMO, 2020).

[15] The five National Data Centers in question are: (1) National Oceanographic Data Center (NODC), (2) National Climatic Data Center (NCDC), (3) National Geophysical Data Center (NGDC), (4) National Snow & Ice Data Center (NSIDC), and (5) Carbon Dioxide Information Analysis Center (CDIAC) (NSF 94-126)

The 2004 "Data and Sample Policy" (NSF 04-004) is particularly notable for formalizing several changes. First, it expanded the scope of shared research artifacts to include physical samples and cited the growing complexity of oceanography research, from larger datasets, interdisciplinary teams, and multi-investigator projects, as catalysts of policy evolution.

Second, it shifted the responsibility of determining which data to share onto PIs. While the 1994 policy relied on agencies to determine which data had "high utility," the 2004 policy made data sharing a default expectation of all PIs[16]. The 2004 policy, likewise, highlighted the importance of national data centers, stating that depositing data elsewhere did not relieve PIs of the obligation to archive their data in policy-recommend repositories. This suggests a desire not just for short term accessibility, but for centralized, long-term preservation.

Third, to enforce these expectations, the policy introduced proposal-specific and reporting requirements. Plans for data sharing had to be included in research proposals, and annual project reports were expected to document progress (or lack thereof) on data sharing. Such requirements for annual reporting continue today.

4.4.2 Post-2010 DMP mandate OCE data policies

After NSF's DMP mandate requiring DMPs in all proposals, OCE issued a revised version of its data policy (NSF 11-060). While it closely mirrors the 2004 policy, it introduces three important changes. First, it explicitly acknowledged the new two-page DMP requirement.[17] Second, it further expanded the list of expected shared artifacts to include "other supporting materials." Third, it reiterated the importance of early engagement with repositories, like BCO-DMO, advising PIs to register and submit project metadata upon award, emphasizing that data

[16] In the 1994 data policy NOAA staff and program representatives from funding agencies were responsible for identifying data sets of high utility and asking PIs to submit them for preservation "*NOAA's National Environmental Satellite Data and Information Service staff and program representatives from funding agencies will identify data sets that are likely to be of high utility and will require their principal investigators to submit these data and related information to the designated center.*" (NSF 94-126). Whereas in the 2004 policy, the responsibility for data sharing is on all PIs "*It expects investigators to share with other researchers, at no more than incremental costs and within a reasonable time, the data, samples, physical collections and other supporting materials created or gathered in the course of the work. It also encourages grantees to share software and inventions or otherwise act to make the innovations they embody widely useful and usable*." (NSF 04-004).

[17] Following the NSF Grant Proposal Guide, OCE's 2011 policy also asked researchers to include a DMP in their proposals "*For each proposal submitted to NSF, the NSF Grant Proposal Guide requires that proposals include a supplementary document of no more than two (2) pages that is labeled "Data Management Plan*" (NSF 11-060).

management and planning for data sharing should begin during the active phases of the research[18].

The next policy update, in 2016 (NSF 17-037), introduced new terminology, "derived data products" using examples like model results, outputs and workflows. This addition highlights the growing importance of modeling in oceanography and clarified what constitutes shareable outputs from funded projects. The policy also addressed informal sharing practices by asking PIs to include URLs to data and metadata in annual reports, likely reflecting the practice of using personal and laboratory websites as venues for data sharing.

The 2024 policy (NSF 24-124) adds specific guidance around code, models, and software.[19] It called for model output and code to be made publicly available upon publication and specifies that it must be accompanied by README files and scripts for accessibility. It also emphasized the archival of both new and modified code, with documentation pointing back to the original code.[20]

This increasing detail reflects how OCE policies have adapted to contemporary practices in oceanography, while the expectation for when data should be made available is still rooted in the legacy of WOCE's data sharing policy. Despite expanding the scope of shareable artifacts, the two-year rule remains central. From physical samples to models and scripts, OCE data policy continues to feature a norm born from a decades-old experiment. What was once a situated requirement has now become a one-size-fits-all expectation.

4.5 Conclusion

This chapter traced the origins and persistence of the two-year data sharing norm that is specified in present-day OCE data policy. I show how this temporal standard was not an arbitrary choice,

---

[18] The specific wording in the policy is "*when awards are initialized investigators should immediately contact BCO-DMO/[Inter-Disciplinary Earth Data Alliance] IEDA and register their projects by submitting project metadata*" (NSF 11-060).

[19] Although similar the 2024 in content to the 2016 policy, this most recent policy dedicates greater attention to software, code, and models. Further clarifying what is included in artifacts to be shared, the document specifies: *"this includes software and derived data products (e.g., model results, output, and workflows). Data should be in a format that is easily accessible; data output in a format that is not readily accessible should be accompanied by a readme file and a script to allow reading it."(NSF 24 -124).*

[20] The specific guidance reads as follows: *"Model output used in peer-reviewed publications should be made publicly accessible at the time of publication. Code and scripts should also be archived at the time of publication in a research code repository. If the newly developed code improves and existing numerical model, an accompanying readme file should point to the original code." (NSF 24-124).*

but a historically produced response to the challenges that oceanographers faced during WOCE, a global-scale oceanographic research program. WOCE helped to institutionalize expectations around timely data availability. WOCE's data sharing policy, noted as a departure from traditional scientific practice, balanced the need for rapid data sharing to support modelling activities with PIs concerns about their right to publish the data they collected first. While compliance was uneven, WOCE succeeded in establishing important infrastructure for oceanographic data.

Today this timeframe continues to shape expectations of data availability in oceanography, even as policy expectations expand to include new forms of research outputs like physical samples, models, software, and code. The endurance of this time norm reveals how scientific research programs of the past, and the infrastructures they leave behind, continue to structure the present, informing the policies in present-day oceanographic research.

Chapter 5. Planning for Heterochronous Data Afterlives

5.1 Introduction

Prescriptive research data lifecycle models have a universal quality to them, in the sense that all data are imagined to flow through the same stages, implying a similar trajectory in both the lives and afterlives of data. In a less obvious way, this assumption of equal data availability is also present in Stahlman's researcher-centered data lifecycle model. Stahlman's (2022) model recognizes the nuanced dynamics between data availability, sharing and delay and their relationship to researchers' career lifetimes. It also attends to the affective dimensions that become salient at different stages of a researcher's career. However, one limitation of the model is that it does not explicitly address how different types of research artifacts – whether samples, specimens, models, code, and software – have distinct lifecycles and afterlives, even though they are understood by the OCE data policy under the broad term of research data.

In some ways, this critique of data lifecycle models may come across as the academic version of a cheap shot. As noted by scholars such as Mosconi et al., (2019), visual representations of the data lifecycles do not look anything like the data practices that they seek to represent; that said, the visual representations are representations after all. That said, the same scholars have argued that these lifecycle models help researchers better understand

funders' requirements for data and herein lies the crux of the matter and the topic of this chapter.

This chapter focuses on the divergence between the data lifecycles and afterlives mandated by data policies and those that researchers plan for. As surveyed in Chapter 4. Since the early 1990s, OCE data policies have included physical samples and specimens as artifacts that require management, sharing, and preservation. Since 2004, the OCE data policy has also included software, models, and model derivatives, with increasingly more granular requirements for this category of research artifacts in the most recent 2024 data policy.

Also relevant to this chapter is Joanna Radin's (2015) "planned hindsight" discussed in Chapter 2. To briefly review, planned hindsight describes data practices that are being done at present in service of a future vision. In my study, the DMP requirement, in addition to the data management, sharing, and preservation that go into making publicly available data, are all examples of planned hindsight. This is because oceanographers are asked to plan in their DMPs and carry out data practices that primarily serve future reusers and downstream uses, rather than for their own immediate benefit.

As such, the first section of this chapter uses pertinent examples from DMPs to present how I identify "exhaustive" planned hindsight in my analysis. In the remaining sections of this chapter, I identify, from DMPs, three forms of planned data afterlives: secluded, splintered, and speculative, to describe how material conditions, disciplinary norms, and institutional arrangements shape what data endure, and how. These three forms of planned afterlives offer the following contributions. These futures of data afterlives speak to the limits of future data availability in an era of data policies and open data discourse that imagines future reuse of data as possible for all data for all purposes. This chapter will also show that data should be understood as heterochronous. And just as not all data are created equal, not all data persist equally over time.

5.2 Exhaustive Planned Hindsight and the limits of research data availability

Planned hindsight describes the act of planning data work with a specific future vision in mind. I build off Radin's (2015) work and argue that the OCE policy enacts not just planned hindsight, but a particular form of planned hindsight; an "exhaustive" version. Exhaustive planned hindsight imagines all data, or research artifacts, regardless of type, material

condition, or infrastructural support, as shareable, preservable and useful to future reusers. In this framing, future use and therefore, sharing and preservation, are not just possibilities, instead they are an imperative. This interpretation, is due, in part, to the indeterminacy of current DMP guidance, which leaves room for expansive and unrealistic expectations about what future-oriented data practices should entail.

One salient example of the imperative of preservation for all data is in a DMP where researchers state that they will preserve the fins of killifish after their experiments for the sake of future reuse, writing in their DMP that "upon sacrifice of the brood stock, we will obtain fin clips, which will be archived in case genotyping of killifish becomes of scientific interest" (DMP 220). As in the example here, the PIs plan to preserve their samples after their research project, not for existing use cases, such as for the lab's next project, but instead for speculative future use cases. The use case imagined is one that has no present-day scientific interest as suggested by the phrasing "in case [it] becomes of scientific interest" (DMP 220).

As DMPs reveal, not all data can be preserved, as in the case of using research methods that destroy the phenomena of interest, for example, one research project's DMP highlights this point:

> "*The [lobster] larvae collected will be used in laboratory experiments to assess performance and oxygen consumption, or in physiological assays of thermal stress. In either case the physical samples will be altered or destroyed in the process and will not be available for archiving.*" (DMP 106).

Yet, still, the researchers felt compelled to explain or justify the absence of preservable data. The exhaustive planned hindsight does not acknowledge the limits of materiality and disciplinary norms.

The future imagined by the policy is totalizing, i.e. "one-size-fits-all". As such, there is a disconnect between the policy's imagined data afterlives and the pragmatic, materially, disciplinary and infrastructurally grounded futures planned by researchers.

5.3 Secluded Futures of Physical Samples

The OCE policy is unique among the directorates studied in the larger Data Afterlives project, of which this dissertation is a part of, in that since 2004 it has identified itself as both a data and sample policy. Naming both in the policy's title signals the importance of digital data and physical samples to modern oceanographic research. The imagined future data availability

embedded in the policy assumes that all research outputs – whether digital or physical – can be made equally accessible through data planning. However, this vision diverges from the reality of researchers' plans for physical samples. While physical samples are often preserved, their accessibility and circulation are shaped by contingent access, institutional limitations, and material constraints. Samples may be stored in PI laboratories, made available upon request, or retained in institutions with limited resources and selective accession criteria. These more restricted and contingent trajectories are what I call *secluded futures*: futures in which data are preserved, but where access, discoverability, and in certain cases even longevity are limited.

The futures of physical samples are not unplanned, as researchers anticipate some of these materials to outlast their projects, but the afterlives of physical samples reflect a form of planned seclusion. As such, in DMPs, a form of selective rather than exhaustive planned hindsight emerges, shaped by the constraints of physical matter, localized stewardship, and uneven infrastructural support, in order to circulate and maintain them.

5.3.1 Uneven institutional support for physical samples

Some types of samples have strong institutional support; this is the exception rather than the norm as most sample and specimen types have conditional or selective institutional support. Institutional and infrastructural support after the afterlives of samples and specimens.

One type of sample that has strong infrastructural support, and is discussed frequently in DMPs, is sediment core samples. Sediment cores are a type of sample that captures the sediment layers of the ocean floor. There is also a research program, the International Ocean Discovery Program, that drills into the ocean floor and obtains sediment cores for their research. There are three core repositories associated with this international program including a core repository at the Lamont-Doherty Earth Observatory in New York, U.S., a second in Bremen, Germany and the third and last core repository in Kochi, Japan. For sediment cores, the OCE data policy specifies that these are the recommended archives for sediment core samples.

For other sample types, institutional support exists but the accession criteria might be selective meaning that the samples need to fit the scope of existing features in the first example or be scientifically interesting in the second example. Samples are unique from other

data types found in OCE DMPs in that Museums are listed as potential archives for these data types:

> *"Parasite vouchers will be catalogued in the Smithsonian Invertebrate Zoology Collection, which is home to the [U.S.] National Parasite Collection. The Collections Committee of the Smithsonian Institution's Invertebrate Zoology Collection charges for cataloging new specimens to cover the costs of materials (e.g., jars, labels), storage space, and curatorial labor. Smithsonian has estimated a cost of $5,500 for cataloging the material generated by this project, which we have planned for (see Budget)."* (DMP 329).

From the DMP segment above, we can infer that the PI has contacted the museum in question during the proposal preparation phase for a quote on the archival costs estimated to be "$5,500", which as the segment explains includes the materials, space, and labor involved in cataloging and archiving the parasite samples that can subsequently be reflected in the proposal's budget. Sometimes the selective accession criteria result in the PIs not knowing the data afterlives of their samples at the time of writing the DMP and sometimes not until the end of the project will PIs know if their samples will be accepted for archival and long-term preservation.

In yet other cases, it may not be possible for the PI to know during the proposal preparation and DMP drafting phase whether the samples collected will meet the accession criteria of archival or preservation institutions:

> "*These lineages will be archived twice per year by cryopreserving a subset of culture in the [PI] Lab -80ºC freezer, and will be available to other researchers on request. During Year 3 of the proposed work, PI [Lastname] will contact the National Center for Marine Algae and Microbiota (NCMA) at Bigelow Labs to ask whether she can submit representative evolved lineages to their collection."* (DMP 134).

In this example both disciplinary practices – of archiving only representative evolved lineages – and likely the constraints of funding, labor, and storage for the NCMA combine to leave uncertainty about the sample's long-term archive at the time of writing the DMP. As previewed as well in this quote, many samples and specimens' primary location of archival are in PI's labs and offices, with the PI taking on the role and onus of short-term data sharing and long-term data stewardship. Such arrangements contribute to secluded futures, where physical samples are preserved locally and with care, but their long-term accessibility and

discoverability remain uncertain and highly dependent on individual researchers and uncertain deposition into sample repositories.

5.3.2 PI-driven data stewardship

When there is an absence of dedicated infrastructures such as repositories and archives, PIs and their physical spaces, their laboratories, are the short and long-term archives for physical samples. As such, storage and long-term archival for samples and specimens, depends on the resources, in terms of storage infrastructures and physical space available to the PI in their current institution.

A data manager working at a data repository shared that in their sub-discipline it was common for samples to be found "in some cabinet somewhere" and that this was typical of "small studies," clarifying that he meant those studies with a small research team made up of "one or two professors and their grad students" (P2). In such cases, long-term stewardship is not passed onto a repository, archive, or museum collection. Oftentimes, this may simply be because no archive exists for a particular sample type. Having the PI maintain the role of the long-term steward has its own time horizon, the data manager noted that, "eventually when [the PIs] retire … who knows where it goes after that" (P2).

As such, some PIs comply with the policy by noting their sample collections' availability on personal or project websites: "Molecular material will be archived at Auburn University. Archived material will be listed on the project websites at [University of Houston] and Auburn, and made available, upon reasonable request, to other investigators two years after collection" (DMP 20). However, as Sharma et al., (2023) discuss, websites are unstable long term repositories of this information, as over time, uncertain maintenance of websites often results in broken internal links and "page not found" errors.

Because the afterlives of samples are enmeshed with the policy's logic of exhaustive planned hindsight, I draw on the following DMP segment to illustrate how researchers reinterpret that logic and its implications for how samples are preserved and positioned for future reuse:

> "*Preservation, Archiving and Access: Standard protocols for preservation of the biological samples will be used and documented to allow taxonomic descriptions. Samples of recently settled C. virginica from field experiments will be archived within [building name] on the campus of [State University, PI lastname]. Samples of oyster*

> *larvae used in geochemical tagging experiments will be archived at the UNC-Institute of Marine Sciences, Morehead City, NC [(PI lastname)] and can be provided upon request to the PIs, provided any ancillary study does not interfere with this project's goals.*" (DMP 283).

The first implication I would like to draw from this DMP segment, is that the DMP envisions reuse, not as open-ended to all possible reuses of the biological samples as envisioned by the data policy, but for the specific possible reuse of "future taxonomic descriptions." Aside from diverging from the "exhaustive" reuse possibilities imagined in the policy, as Van Allen's (2023) study on the preservation of butterfly samples show, the imagined future use of physical specimens affects not only *what* is kept, but also *how* they are curated for long-term preservation. Put another way, anticipated future reuse impacts decisions in the present that will enable a select set of reuse, whilst foreclosing other possible forms of reuse.

Turning to *who* the anticipated re-users are. It was quite common in DMPs for projects working with physical samples to note that access to samples are primarily "upon request." This means that the availability of the artifacts hinge on the PI's ability to continue to preserve them, but also to continually manage the logistical work of responding to access requests. While many other DMPs explicitly state that when they write "upon request" they mean this is a broad sense. In this DMP, the PIs are ultimately the arbiter of to whom "upon request" applies to, with the added caveat that the requestors' study cannot duplicate the funded project's goals. The implication of this is that through the added caveat, sample availability, instead of a guarantee, becomes conditionally mediated through the PI.

Another implication is that the PIs are responsible for cataloging and documenting the samples that they plan to preserve in their own collections. As Braun (2023) astutely notes, in the process of preserving samples, the sample is essentially "split" into two in a physical but also epistemic sense. On the one hand there is the sample itself and on the other the sample's metadata – catalog and documentation – as two distinct entities and stored in separate locations. As it pertains to reuse, this complicates matters. For a future re-user to successfully reuse these samples, they will not only need the c.viriginica or oyster larvae samples, but also their documentation. The two will need to be re-joined, reconnected to make the sample "whole" and reuseable. As Braun's (2023) study shows, seemingly trivial issues such as the legibility of handwriting, or the stickiness of a label's adhesive, and how organized the PI and

their research group are will significantly impact the viability of future reuse. The seclusion of physical sample afterlives, then is twofold, not only from potential reuser to the materials themselves. But a seclusion of the materials from the metadata that enables the materials to be reusable.

In contrast, then, to the exhaustive persistence imagined in the policy. In practice, samples may be preserved, such as in a cabinet, but they are neither easily accessible nor discoverable. Instead, their availability is localized, unstable, and as shown in the above example, conditional as well. In some cases, these limitations are contingent on human or institutional factors. In others, the material characteristics of the samples themselves make availability impossible, as I will in the next section.

5.3.3 Material Decay

Long-term maintenance of biological samples is made possible by an assortment of tools such as “glass vials”, “zip-lock bags”, “ethanol”, to specialized substances “RNase inhibitors”. The most commonly mentioned method is cryopreservation. Requiring specialized freezers that can maintain temperatures ranging from “-20°C” to “-80°C”. However, naturally occurring processes such as evaporation, degradation, or contamination, make preservation of physical samples an ongoing challenge, and in some cases not a possibility.

Maintaining long-term access to samples is no easy feat. Data stewards, whether they are individual PIs, archives or museums, aim to prevent or stall naturally occurring processes that beset physical materials. Depending on the type of material in question, keeping physical samples for possible reuse for the typical retention period, a minimum of 5 years, is itself a challenge. Some materials have a shelf life of less than the mandated retention period, as one DMP explained, “solutions for [Inductively Coupled Plasma Optical Emission Spectroscopy] ICP-OES analysis have a shelf life of approximately two years and are disposed of following laboratory safety procedures” (DMP 253). This example shows how material decay creates a planned future of impermanence that runs against the policy’s imagined availability even in short time periods.

Cold storage and cryopreservation, appears to stall or “freeze” material degradation but even then sometimes barely meeting the retention timeframes. For instance, one DMP explains that they will “samples, purified residues … [will be] retained in cold storage”, however the PIs explain that “longer preservation of these materials is not always feasible

given potential for contamination and evaporative losses with time" (DMP 338). In this example, the concern is with the sample disappearing over time, "losses", however another limitation of physical samples over long time durations are questions about the quality of the material and whether it'll remain suitable for downstream uses as a DMP explained, "DNA extracts, RNA extracts, as well as amplified nucleic acid samples will be archived at -80oC at UCSC for a minimum of 5 years, after which their integrity is questionable" (DMP 145). Unlike, the example of the purified residues, DNA and RNA extracts do not face dangers of disappearing, at least in the same way, but rather it's their epistemic value, their "questionable integrity," that is decaying.

The DMP segments here are all examples of exhaustive planned hindsight as PIs are justifying and managing expectations for why it is that their projects may not be able to share data, or if they do for short time periods.

Taking the examples on physical samples presented throughout section 5.3, they show how physical samples are subject to forms of seclusion that challenge the vision of exhaustive availability imagined by the policy. Researchers anticipate these limitations and attempt to account for them in the planning process, justifying through DMPs why sharing may not be feasible or sustainable over time. The secluded futures of physical samples is not simply a logistical issue but, as Van Allen (2023) argues in her study of butterfly specimens, a form of anticipatory curation in which the imagined future uses of a sample shape what is preserved and how. Similarly, Braun (2024) highlights that preservation is always a situated practice, shaped by researchers existing data practices, institutional resources, and the material properties of the samples themselves. The DMP segments presented echo both of these points so what is planned for, then, is not the universal availability of data over time, but a set of constrained and conditional futures, where the act of preservation is less about keeping everything forever, and more about navigating what can feasibly held onto, for whom, what purpose, and for how long.

5.4 Splintered Futures of Models

In the case of models and their outputs, the future imagined by policy is challenged by both scale, disciplinary norms and the epistemic status of model outputs. Researchers plan for futures in which only parts of their research output will be made accessible, while other research outputs are too large or too expendable, e.g. easily regenerated, to justify their long-

term preservation. These splintered futures are marked by uneven accessibility. For example, model outputs, the data generated by models, may be shared on request, stored on local serves, or archived for short periods. Whereas, at present, models themselves are shared and circulate through an array of informal platforms like GitHub and project websites, that curiously sit in a middle-space between formal repositories and alternative informal code-sharing platforms.

Much like physical samples, the futures of models and model outputs are planned for but fractured. Instead of secluded, the planned afterlives of models and outputs are selective, fragmented across various infrastructures, and the longevity of their afterlives is premised on their short-lived relevance or technological obsolescence. Researchers do anticipate future reuse, but the future imagined is provisional, resisting the policy's assumption that all data can or should endure equally.

In response to the challenges of scale, and the fragmented nature of model and model outputs, researchers plan on turning to cloud-based storage and processing, which promises greater storage and processing capacities. This move is framed as a solution, but introduces new constraints around platform dependency, pricing, and other uncertainties.

5.4.1 From large to colossal data scales

DMPs and interviewees shed light on how large datasets, which include model outputs, hydrophone sound wave files, video, and photos, are *"*numerous and large." In other words, these types of data are large both in file size and in volume. Individually, these two attributes, having a lot of files but small in size, or having few large-sized files, appears to not be an issue. A number of DMP segments articulate anticipating not having any problems on this front explaining that the size of the files produced throughout the course of their project, relative to the amount of storage available, is small.

Large data poses an issue for data sharing but principally, moving large datasets takes time. One interviewee recalled how a user wanted to:

> "*transfer all of her video data up to the Microsoft Azure Platform, it took two and a half months…. Another user wanted all of our hydrophone data, the best way was she sent us a disk. And we filled the disk and we sent it back to her. Because moving that much data across the public Internet, it's got nothing to do with our bandwidth. It's difficult when you get to that data size.*" (P21).

The interview excerpt above contains two informative cases to break down. The first case is illustrative of the challenges for PIs when they are the data provider. The time it takes to share data is now on the order of months, instead of days or hours. This has implications for both when PIs need to think about when they begin depositing their data. This in turn has a knock-on effect that impacts the time it takes to close out a research project, as well as obtaining Digital Object Identifiers (DOIs) for manuscript publication as seen in Chapter 6. The second case shows the challenges for PIs when they are the data reuser involved in acquiring large data. At the same time, it presents a possible solution. Which is a return to physical storage mediums, like disks, to move large data. For PIs who find themselves in this situation, they will have to allocate room in their budget to purchase disks, if needed, and take into account the time involved in shipping the disk, as well as copying data to and from the disk.

"Large" data as an analytical category is interesting as there appears to be a concrete threshold where when data reaches a certain size in the order of terabytes (TB), it is universally considered large.[21] However, "large" is also at the same time a relative adjective as what types of data and when data is labelled by DMPs and interviewees as large is shaped by their subdiscipline, their previous data management experiences, and which actor's perspective we are taking i.e. individual project, or data manager at a repository.

The issues highlighted in this subsection are likely to persist and compound as the frontier of model-based research is for greater granularity of temporal and geographic resolution. In other words, not just synoptic view of earth, but in ever finer levels of granularity across both space and time. As a result "this can just really blow up our model data volumes like you wouldn't believe … on of our bigger projects, the total data volume is, over 600 terabytes" (P13). Another interviewee in the same field wondered "if our data sets

[21] This statement comes from a combination of analyzing DMPs and from interviews. A NOAA National Centers for Environmental Information (NCEI) archivist P29 I spoke with noted that there are two separate archiving tools that can be used to send data. Which tool is recommended to researchers depends on the size of the data. If the data is less than 20 Gigabytes (GB) researchers should use Send2NCEI, however if the data is larger than 20 GB then researchers should use the Advanced Tracking and Resource Tool for Archive Collections (NOAA NCEI, 2020). Meanwhile, oceanographers interviewed who work with modelling data will often discuss data in the order of Terabytes (TB). Contrasting the NCEI and modeling data example highlights that there is no universal threshold for what constitutes large data, but that typically interviewees and DMPs will begin to discuss the need for adequate storage technologies and infrastructures when the data is on the order of several TB.

are becoming so huge, how [is] somebody [else] expected to download all this data to their computer and make their analysis?" (P1).

When data scales are on the magnitudes of hundreds of terabytes and petabytes, this scale becomes a barrier to all aspects of model output's afterlives, from sharing, storing, and reuse.

5.4.2 Fragmented Access and Accessibility of Model Outputs

How then do oceanographers plan for the afterlives of their large data? As I will show, the first and second types of splintering occur in the afterlives of model outputs. First, researchers face sociotechnical constraints, such as the difficulty sharing large datasets online and the fact that many repositories are better suited for smaller, observational data. In response, they make strategic decisions about what will be shared and what will not. Since sharing everything is not feasible, model data are split into two categories, those made immediately accessible, and those available only upon request, as the DMP below wrote:

> *"For data sharing, the PI will maintain a web site dedicated to this project. Data subsets such as calibrated raw data, or data products, will be available for direct download from the website… Given the anticipated size of the complete data set (≈ 150Tb for the raw, pre-processed, data products, and fully analyzed data), extensive data sharing will be initiated through a request to the PI who will transfer a copy of the requested data to a storage device (e.g. hard drives, cloud storage) provided by the requestor."* (DMP 201).

In the example above, the PIs plan to curate a meaningful subset of research outputs to be distributed through conventional venues such as data repositories or in supplemental materials of peer-reviewed articles. At the same time, the entire dataset is stored, often through redundant disk arrays, kept in PI's labs. Here we have the second type of splintering, researchers split up their model outputs across various platforms and locations.

For example, DMP 288 stated that only "recent results will be available for immediate download." Whereas "older results will be listed on webpages and files" (DMP 288). A variation of this practice is to share, through a repository, for example in DMP 272, "representative files" whilst the researchers would maintain the entire set of outputs elsewhere. Another popular way to share large datasets involved PIs providing "[Hypertext

Transfer Protocol] HTTP and [File Transfer Protocol] FTP access" for third parties to download their datasets.

5.4.3 Transient Epistemic Utility: Limited lifetimes of model outputs

An additional consideration in the long-term preservation is the limited "shelf-life", for lack of a better term, of model outputs. For example, one DMP states that model outputs "will have a relatively short expected lifetime (less than 1 year) and we have no plans for longer term archival" (DMP 202). Unlike physical samples, where shelf-life may be short due to naturally occurring degenerative processes, for model outputs, the shelf-life is determined by the evidentiary status of "outputs" or "data" within the ocean modeling community.

A data archivist provided some observations for why this may be the case:

> "*I think one of the questions when you have things specifically like model output is, can that model output be regenerated by running the model again? Then, that makes it not as important to hold on to the model output, because you can recreate it. And, or, if it's, like from a really old model, are there models that supersede [it] that can create similar information to the original model data.*" (P29).

If the model output can be easily recreated, then the specific copy that the researchers have is neither unique nor valuable. Both data managers and modelers do not consider model outputs in the same way as they would for observational data as model outputs are not derived from direct observation of the phenomena of study, the ocean, but instead are simulations of the ocean.

One data manager shared that they recently had to deal with the issue of archived model outputs, as their organization transitioned from a tape archive system to an online archive system. During this transition, it struck the data manager just how little thought had been made about *what* to archive and *how long* to archive it for, likening their archived data holdings to "a giant reservoir" (P13). The indiscriminate data archiving was described in a way that was reminiscent of hoarding. When sharing these events, the data manager put themselves in character as a scientist who without pause archived their data:

> *"I can't look at this data now I'll just throw it on tape and I'll look at it, in six months or a year," well, stuff tends to just age off and people forget about it or they move on to other things… There's a tremendous amount of data on there that had been, written once and never read."* (P13).

During this system transition process, the data manager made the active decision to not archive some 10, 15 even 20-year-old model data because, simply, "*it's obsolete*" and no one had looked at it in years as the models that produced the data were, at the time of our interview, 10 - 20 years old. As the science in up-and-coming models advances, results from older models become outdated, outdated models become used and resued less and less.

The divergence between the data policy is that researchers are being asked to share and keep artifacts that they deem to have a limited epistemic utility, whilst the policy treats all data afterlives as valuable and meriting preservation. Instead, model outputs and their limited lifetimes show that some data afterlives are characterized by their expendability.

5.4.4 Frozen code, Fragmented Models

So far in this subsection on Splintered Futures, I have focused on model outputs. But the models themselves, i.e. scripts, software, and codebases, introduce a different divergence with the data policy, and the fourth type of planned fragmentation for model afterlives. Models are understood by my interviewees as separate from model outputs and in some ways the artifact that had greater epistemic value. As a senior oceanographer explained to me:

> "*The models are important as well, because I think that there's a difference between data and model output. Data is a real observation and sometimes modelers talk about data as being the output from their model, which I think is not as important as the real data, so the actual fluid [meaning the ocean]. So I think the most important thing is the model, with the kind of work that I do, rather than the data, because the data is manufactured from the model.*" (P33).

Unlike datasets, which reach a point of stability where they can be deposited and assigned a DOI, models are iterative in nature. Models evolve over time, with the original researchers as well as reusers working with, improving, and updating them. Yet, the data policy assumes these digital artifacts can be deposited as static, finalized artifacts, in essence "frozen" in a similar vein to assigning DOIs to a dataset. This divergence between the policy's assumption and the practices of oceanographers who work with models fragments the way models are shared, maintained, and made available.

Software and code pose an interesting challenge for sharing, as illustrated by the DMP segment that discusses expecting "to issue regular updates as we continue to use and develop

them", as there is no specific point in which they become static. DMPs whose projects include models write about how they plan to share their models on GitHub. As one DMP explained:

> "*The community standard platform for sharing software is GitHub (https://github.com/) … Perhaps the best way for software to have an impact is for it to be highly visible on GitHub. Therefore, our tools will be released to the community as one or more GitHub projects. After the initial release of our tools, we expect to issue regular updates as we continue to use and develop them."* (DMP 32).

This example illustrates that GitHub is considered the "community standard venue" and that GitHub is used as it supports continuous development as the PI expects "to issue regular updates." GitHub allows models to be shared not as stable endpoints but as ever-evolving projects. Modelers use it not only to distribute their code and software, but also to version them and further their development.

This nature of flux that is constitutive of models presents a challenge to their sharing through policy-recommended repositories. PIs who plan on following the mandate closely will have to create a "frozen" version of their model, one DMP noted this:

> "*PI's [Lastname] and [Lastname] will be responsible for ensuring proper freezing of model code and output associated with all project publications and the archiving of code and output on the project websites and with the Biological and Chemical Oceanography Data Management Office (BCO-DMO).*" (DMP 81).

This frozen version of code allows the model to be cited in peer-reviewed publications as well as its deposition into BCO-DMO. Unlike GitHub, however, where active model development and iteration continues, the frozen version of models satisfy policy requirements but may never be used again. These parallel tracks of the frozen model versus the live model splinter the model's availability, creating disconnects between what is archived and what is actually in use.

Unlike sharing data through websites, sharing models through websites appears to be a somewhat stable venue for maintaining access to models. Many PIs who work with models note in DMPs that they will be building off publicly available models that can be found on the model's dedicated website.[22] One senior oceanographer I interviewed is part of a group who

[22] An example of a model that is reused and has its own dedicated website is the MITgcm a model for ocean circulation *"MITgcm and the updated code will be made available at [website URL]"* (DMP 166). Another example

has developed a widely used model that is used in ocean circulation research, they shared with me:

> *"My group has this model, it's called the MITgcm, that is widely used. One of the reasons that it is used is that there is documentation. And we're not perfect. The documentation is far from perfect, but we have had documentation sessions to get it going and to sustain it… So if you type in MITgcm documentation, you'll be able to download thousands of pages of document, or the code … and I think a lot of models are getting into that state." (P33)*.

Dedicated model websites occupy a middle ground between policy-mandated repositories and platforms like GitHub. First, what the quote illustrates is that what makes a model usable and therefore reusable is not just the code, or software, but the ecosystem around it. Documentation, a group who creates it and continually revises and adds to that documentation through "documentation sessions," and a community of users who find the model helpful for their own research.

Second, whilst websites allow sustained access and community uptake. Just like GitHub, however, they fall outside what the policy currently mandates. There are also genuine questions about discoverability and long-term preservation. However, this raises an open question, that is outside the scope of this dissertation, which is, what would long-term preservation and archival accomplish for models if current archival frameworks are premised on an artifact's stability; something that is discordant to models?

Modelers have to deal with a diverse range of artifacts, and not only are model outputs large - which poses issues to sharing and accessibility - but that models themselves are a type of artifact and that these tend to have a range of access points from software distribution and sharing platforms, notably GitHub, but also the Comprehensive R Archive Network (CRAN) for R packages. PIs are also sharing their code in repositories i.e. BCO-DMO, Zenodo, Dryad. As well as model specific websites. This, overall, creates a fragmented landscape of where models and model outputs can be found. But this fragmented landscape is functional. Each platform enables and does different work. For instance, GitHub for continual development, project websites for thorough documentation, and repositories for policy compliance.

---

of an ocean model is the Regional Ocean Modeling System (ROMS) which one project plans to reuse for their research project "*The [ROMS] software is community-supported and developed, and is available freely at [website URL]."* (DMP 203)

Oceanographers who work with models respond to the one-size-fits-all approach to availability implicit in the data policy with splintered access points optimized for different parts of the model's life as well as the purpose at hand.

5.4.5 The promise and perils of cloud computing infrastructures

In response to the challenges of scale, and the fragmented nature of model and model outputs, researchers plan on turning to cloud-based storage and processing, which promises greater storage and processing capacities. This move is framed as a solution, but introduces new constraints around platform dependency, and other uncertainties that come with the involvement of for-profit companies. Put another way, this section is about how new storage solutions are envisioned to introduce new futures and afterlives for model and model outputs.

A move towards the cloud is on the surface about changing where one's data is stored. However, underlying this location change for data has implications for data access and reuse as well as ownership of that data. In line with the concern shared by an interviewee, how are folks supposed to work with such large amounts of data on their machines? The change that the cloud enables is what if one didn't have to bring the data to their machines, therefore circumventing the issues of dealing with a machine's storage capacity. One interviewee explained to me: "If someone wants to look at our published data, instead of bringing the data to the computation, we take the computation to the data. And so that's a paradigm shift" (P1). From this quote we see the involvement of for-profit companies, Google and Amazon, in the storage of publicly funded research data. This interviewee considered this change akin to "a paradigm shift" implying how transformative this development is for the future of data-intensive research.

Apprehension follows any significant change, and this sentiment was expressed by a few interviewees I spoke with. The specific concern with cloud computing and storage was that for all the benefits, it may be a potential Trojan horse. One former repository employee shared:

> "*you have to be very careful because people always say Google and Amazon provide storage for free … Well, that's today, and if they decide to change the pricing policy, you would be caught, So I never trusted the zero cost from Amazon, it was the commercial cost or the academic costs*." (P19).

The above quote conveys the concerns about entrusting the access points of research data to for-profit corporations. The exchange for "zero costs" today, is uncertainty about when the dues owed would be paid. The same interviewee agreed that there are tangible advantages to data

work in a cloud environment but felt that despite NSF wanting this change to occur, the NSF was not committing the appropriate resources commensurate with their stated goal and explained why that was to me:

> "*[The NSF] needs to be like other government agencies, and they need to get an NSF Cloud now … If they want to have all the data generated from NSF projects in the cloud, they should commit resources to making sure those cloud services are available to all data centers*." (P19).

From this interviewee's perspective, the NSF would have more bargaining power, compared to a single data center, in a hypothetical partnership with Amazon Web Services or Google.

Other interviewees were more neutral in their outlook about the cloud development but expressed unease about unknown costs associated with "downloading" data, or data egress, from the cloud. One interviewee admitted that this may lead to a change in their repository's data use and download policies as "in the first go round [we] really did not address [it] tremendously, outside of all data is free … we'll have to start putting some parameters around it, resources are finite" (P21). Another repository was committing more resources by creating their own cloud for the same reason as:

> "*you cannot predict egress charges, the public is our user. Someone could download all of our data. They have the right to do that. It's free to use. But that might hurt us from a budgetary point of view. So we're trying to do at least day one is kind of create our own cloud.*" (P21).

Because of ballooning data storage requirements, many repositories are contemplating or are actively moving towards cloud computing and storage infrastructures. However, there remain unanswered questions about the role of corporate actors, funding for repositories and the costs associated with cloud computing that may impact repository policies.

5.5 Speculative Futures of "Uncommon" Data

For uncommon or unconventional data, their planned afterlives are not characterized by seclusion or splintering but speculation. In some ways then, using the term uncommon, or unconventional data is an ill-fitting term; yet I assign this term as it captures the status of this group of research artifacts from the perspective of data policy with an emphasis on infrastructural absence. For researchers, and possibly entire subdisciplines, these data may be

entirely typical as they may not be rare or epistemically fringe. Instead, they appear "uncommon" and "unconventional" only because they do not fit within the repositories that the data policy mandates. Put another way, the data featured in the examples here can all be understood as orphans of infrastructure, as they lack clear disciplinary repositories and standards that enable their easy deposition into other venues.

These data types are frequently accompanied in DMPs by improvised access, custom-built platforms, or hopes that either existing repositories can accommodate them or new repositories will emerge. Researchers, facing infrastructural absence, engage in a form of speculative planned hindsight where they imagine possible futures in which their data might become discoverable, reusable, or have standards developed for them, but these futures are provisional and contingent at the time of writing the DMP.

Rather than understanding these cases analytically as a failure to comply with the OCE data policy, I suggest instead that these are PIs' pragmatic attempts to plan while faced with infrastructural absence for their data. In other words, DMP authors are interpreting the mandate in flexible ways as they make good-faith efforts to comply with the spirit of the mandate rather than the letter of the mandate. Moreover, these speculative data futures reveal the limits of the exhaustive data availability envisioned in the data policy, when the infrastructures that such planning relies on do not yet exist.

5.5.1 Improvised Access

In the absence of infrastructure, PIs plan on creating access for these data and take on the role of long term stewards. As one DMP explains, they will do for their microstructure data "since there is currently no suitable 'publicly run" repository for these types of data, they will be archived at [website url] data" (DMP 19). Given this, the easiest option is to share uncommon data through websites. Other researchers may go further, by not only sharing data but also documentation and metadata and other ancillary artifacts, which parallels what data reusers could expect to find for datasets that are deposited in repositories. In the case of a research project working with integrated genomic data they explain that there is no "available database" as so part of their project they would "propose to develop a resource that will allow for integration of phage and host genomic data, host range data, and time course data as well as the extensive metadata available for our samples" *(*DMP 58). These two examples are

reminiscent of the PI-led sharing and stewardship of physical samples and models surveyed earlier this chapter, and previously raised issues about discoverability and long-term available apply here too. However, the key distinction for uncommon data is that because there are no alternative sharing options available, PIs are planning on sharing through websites *because* of the OCE data policy. These examples challenge the assumption that sharing data through websites is the path of least resistance.

5.5.2 Community over Compliance

In other cases, repositories do exist for uncommon data, which if available, is more preferable over sharing data through websites. For example, the following DMP explains that "There is no equivalent database in the U.S" (DMP 303) for spectral data. Given this, the PIs have identified a specific database that is maintained by a single researcher and wrote about this in their DMP, "Preservation plan: Collected FTIR spectra will be archived in the PULI database (http://puli.mfgi.hu) maintained by [First name Last name].[23] The purpose of the database is to allow for a quantitative interpretation of acquired spectral data from other researchers" (DMP 303). However, the repositories that DMPs discuss are not listed in the OCE policy and strictly speaking don't comply with the mandate. As it relates to exhaustive planned hindsight, note that the DMP explains that the database is geared towards a specific type of reuse.

Another example of PIs choosing a community infrastructure for uncommon data over policy-mandated ones is in the following DMP that works with marine fish telemetry data:

> "*Our preferred method of archiving and making these data publicly available will be the Hydra hydrophone data repository [website URL]. This repository is specific to U.S. West Coast telemetry data, developed to facilitate data sharing and research coordination among researchers, and publicly available. It can be searched with a web-based map [graphical user interface] or by specific species.*" (DMP 319).

As presented in the segment, the Hydra hydrophone data repository appears to be the to-go repository for this PI's research community. As such, in making this choice, the PIs have implicitly made a choice of prioritizing discoverability of their data and immediate reuse over all possible future uses of their data.

[23] FTIR refers to Fourier-transform infrared. PULI refers to the Pannon Uniform Lithosheric Infrared (PULI) spectral database.

5.5.3 Stabilizing Speculative Futures

Not all uncommon data is doomed to a future of infrastructural abandonment. There is evidence in DMPs that over time new data infrastructures are established, as such this section is about building infrastructures for those data that currently have none. For instance, one DMP writes: "If this website and database are successful in attracting users and other contributors, it is likely that new funding will be requested for continuation after this project" (DMP 149). This is an example of a bottom-up effort to build and develop a community infrastructure, likely morphed from the PIs improvising access to their data.

Whilst this appears to be a more informal and more speculative future as its actualizing depends on attracting users, data contributors, and securing additional funding, others are more concrete. For example, in the case of subsurface float data, which was one of the key data streams in WOCE[24], one DMP wrote:

> "*As of this writing, there is no active data assembly center for subsurface float data. Over the next year, [Co-PI lastname] will be working toward establishing such a center at NOAA's AOML in Miami, FL in collaboration with [AOML Director of Physical Oceanography Division]*." (DMP 214).

At the time of writing this dissertation, there is no formal data assembly center, however, the NOAA Atlantic Oceanographic & Meteorological Laboratory (AOML) websites notes that "subsequent updates will be included as additional appropriate float data, quality controlled by principal investigators is submitted for inclusion." (NOAA AOML, 2024).

As lacking standards for data and metadata can make it difficult for uncommon data to be deposited into infrastructures, a first step towards materializing future availability is the development of such standards. One DMP explicitly identifies this challenge and proposed a solution:

> "*Aside from the LADCP-shipboard [Conductivity, Temperature, and Depth] CTD profiles, there are currently no established standards for archiving or data from many of the fine-scale sensors used in T-Tide… We propose to work with the CPT to evolve*

[24] The NOAA webpage for Subsurface Float Data notes that a copy of the WOCE subsurface float observations have been copied over to NOAA in October 2014, with an update in December 2017 and another in June 2024 (NOAA AOML, 2024)

> *formats for data and metadata suitable for archiving both sensor and (critically) model output from the experiment."* (DMP 74).[25]

This example highlights how researchers plan for an infrastructural intervention, of creating metadata and formatting standards, that precedes and enables future sharing and archiving. While the outcomes of such efforts are unclear, they nonetheless mark a speculative yet intentional attempt to bridge the infrastructural absence of uncommon data.

5.6 Conclusion

This chapter has shown that oceanographic data do not move uniformly through the research data lifecycle, contrary to the assumptions embedded in classical data lifecycle models and policy mandates. By analyzing DMPs, I identify three forms of planned data afterlives that diverge from the futures envisioned by the OCE data policy.

The *secluded afterlives* of physical samples reflect how material fragility and institutional resources enable and limit long-term availability. The *splintered afterlives* of model outputs arise from their scale and short-lived epistemic use. Afterlives of models are splintered due to the iterative nature of model work, compounded by repository requirements that require freezing dynamic code into static forms. The *speculative afterlives* of uncommon data after shaped not by neglect, but by PIs' sincere efforts to plan for sharing in circumstances where no suitable infrastructure currently exist.

Theoretically, this chapter extends Radin's (2015) concept of "planned hindsight." I argue that the OCE data policy promotes a particular mode of planned hindsight, one that is exhaustive. Instead, the examples presented throughout this chapter show how planned afterlives can be better characterized as restricted or selective. Researchers planned futures for their data are shaped by the properties of the data themselves, infrastructural support, disciplinary norms, and practical constraints. As a result, this challenges the data policy's current presumption of universal availability and preservability.

[25] LADCP refers to Lowered Acoustic Doppler Current Profiler. CPT refers to Cone Penetration Testing.

Chapter 6. Practice-time profiles of data management and sharing

6.1 Introduction

The NSF's data policy envisions data sharing as a routine part of research. However, researchers often struggle to make time for the very tasks that would realize this vision. Data management for public sharing is frequently treated as one of the final steps in a research project, something to be completed once the core scientific work is completed. This sequencing of tasks, combined with time pressures and constraints, contributes to the deferral of data sharing work. While the policy presumes a linear progression from data collection to long-term availability, this chapter shows that researchers' experiences diverge significantly from that model. The difficulties of "finding time" for data management is shaped by temporal conflicts that reflect labor constraints and the mismatched assumptions about when, how, and who does data work.

Starting once again with the prescriptive models of research data management and sharing, these models implicitly presume a linear, standardized trajectory from collection to sharing and preservation. Moreover, with the exception of some models, such as the Data Curation Continua (Treloar & Harboe-Ree, 2008) and Stahlman's (2022) researcher-centered data lifecycle model, human actors and the labor required to move data from one stage to the next are absent from these models. It is well known that in practice, researchers struggle to find time for key data management tasks (Borgman et al., 2016; Tenopir et al., 2011, 2015). Based on interviews, this appears to be, in part because researchers conceive of data management for public sharing as something that happens only at the end their projects. Some information scientists contend that time constraints, as time problems are fundamentally information problems (Greyson (2016) in Haider et al., 2022). I show in this chapter that whilst this is at best a partial understanding and that time constraints are also data work and labor problems.

Practices, whether everyday routines like a morning ritual or research-oriented tasks like data management only exist when they are actively performed. As Shove et al., (2009) argue, practices do not merely occur in time, but "practices make time" (p.17). From a productionist understanding of time, as discussed in Chapter 2, this means that practices

require continual effort to be sustained. Put another way, practices must be "actively reproduction and performance" (Shove et al., 2009, p. 19). This understanding of practice and time informs the approach taken in this chapter.

To explore why researchers often find data management and sharing difficult to prioritize, I draw on Shove's (2009) notion of "practice-time profiles" which "refer to embedded conventions of duration, sequence and timing associated with the competent performance of a practice" (Shove et al.,, 2009, p. 25). In other words, this chapter examines how data management work that supports data sharing is shaped by embedded norms around duration, sequence, and timing. The same task, for example, creating documentation, when intended for internal use versus public reuse, takes on a markedly different characteristic. When creating documentation for internal use, it can be quick, informal, and therefore takes less time. Instead, when documentation is intended for public use and reuse, it requires a more thorough approach, and reworking to fit standards, guidelines and requirements articulated by the data policies and data repositories.

This chapter is organized under the following two claims: there is no established practice-time profile for data management and sharing and that knowing the right time to do/reproduce practices is a marker of competence and is necessary for developing practice-time profiles for data management and sharing. I build out my argument throughout the Chapter's four sections.

The first section presents how data management for sharing data publicly lacks an established place in the temporal organization or research practice. The second section describes the consequences of the absence of an established practice-time profile. In particular, the default seems to align with the linear models of research data lifecycles, resulting in an end-loaded practice-time profile for data management. In this second section, I provide examples of how deferring work to the end results in temporal and labor misalignments contributing to minor to substantial delays to data availability. The third section focuses on the temporal dimension of timing and the ways that this matters for data practices. I conclude with three ways that researchers are developing new practice-time profiles for data management. With the caveat that there is no universal or ideal practice-time profile, examples from my study include early planning, creating routines, and automating workflows.

6.2 The Missing Practice-Time Profiles of Research Data Management

The first section of this chapter presents a negative case of a practice-time profile. Given this, I argue that data management for public reuse is challenging because researchers have no established practice-time profiles for this type of data work. Interviewees, when asked, how the NSF mandate had impacted their day-to-day data work, many without missing a beat responded "not at all" (P20). However, in the same breath, they remarked that the "only thing is now I have to spend time making the data so somebody else can use it" (P20). During the first set of interviews I conducted in 2022, I couldn't make heads or tails of how these two statements could be simultaneously true. However, after analyzing the second set of interviews I conducted in 2025, it struck me that this was because interviewees understood metadata, documentation, re-formatting, following standards, as extensions of existing data work, but when researchers find themselves doing data work in preparation for sharing, they find that not to be the case.

As it turns out, data sharing, and managing research outputs to be shareable is far from a trivial task as PIs provided concrete estimates ranging from "10-20 hours of effort", to "30 hours" (P14), as well as more nebulous ones such as that it just "takes forever," which nevertheless got across the point that these tasks are demanding on PIs time and effort. Instead, we can see that the requirement of management and sharing of data policies generates new forms of labor and asks researchers to work outside their established temporal norms and rhythms for research data practices. Working outside of temporal norms and rhythms leads to lacking temporal norms for research data management practices. The next section explores the consequences of this absence of missing practice-time profile.

6.3 End-Loaded Practice-Time Profiles of Data Management

In the second section of this chapter, I show how existing models of research data management, which present data sharing as occurring at the sunset stages of the research project, match some of my interviewee's informal practice-time profiles for data management and sharing. As one PI noted:

> *"So I think we understand that at the end of our project and our output, publications, thesis that we want to … we have to share data. It's making sure [that] at this stage we're still in a position where often, when we've done all the work that we're interested*

> *in, that, we still then have to **do extra time** to make sure the data is in the right formats." (P22 – emphasis added).*

This quote shows that researchers may be aware of what the mandate requires but because projects span multiple years, the task of data management and sharing doesn't get the attention it should be given until the very end. When done at the end, PIs often do not "realiz[e] that data management takes a lot of time, and often dedicated resources and staff" (P25). Instead of incrementally doing data management through the active phases of the research, when left until the end means that all the time and effort that required for data management and sharing is experienced over a short duration. Specifically, I show how this practice-time profile can be characterized as end-loaded and the ways that this end-loading of data sharing makes public data availability vulnerable to delays.

In my data, I did not encounter instances where data sharing did not eventually occur, but this is likely due to sampling procedures and that those who are interested in this subject are more likely to agree to being interviewed. It is likely that this end-loading of data sharing also results in data not being shared at all.

In this subsection I focus on how end-loading practice-time profiles of data management leads to delays. I highlight first constraints internal to the research project by highlighting the role that key personnel play in the dynamics of an end-loaded practice-time profile, as these human actors are more often than not responsible for the labor of data management and sharing. Second, I highlight constraints external to the research project.

6.3.1 Key personnel and the dynamics of data availability and delay

As the research data, per the mandate, is expected to be useable beyond the small circle of researchers to the much larger public, additional data management tasks are required for this move to be made successfully. Common data management tasks necessary during this phase of the project include re-formatting data to open-source formats and ensuring the data has sufficient documentation and metadata. These data management tasks are the intermediate steps that are omitted from the final research outputs and the labor that went into their production is rendered invisible. These intermediate steps in researcher's data work are performed by key project personnel, such as students, technicians and even project staff such as laboratory managers.

This subsection highlights key human actors that, with the exception of a handful of DMPs reported here, are omitted from the documents. Through interviews though it becomes clear how students, technicians, and other non-PI research members, to be referred through this section as personnel, play a significant role in shaping the dynamics of data availability and delay in an end-loaded practice-time profile.

Starting first with graduate and doctoral students, their theses are considered a key research output and one way that PIs write that they anticipate project data being shared. Depending on the graduation timelines and how this lines up with the timing of the project close-out phase, these may further push the availability of data from the policy-mandated time to further down the road. For example, some DMPs write about making their data available according to the mandate but with "the exceptions will be data related to graduate thesis projects, which will not be made available until 12 months following graduation" (DMP 281). One interviewee explained to me how managing both student graduation timelines and data availability timelines is difficult:

> "*It's very often that people, you know, have a grad student write a bunch of code, collect a bunch of data, and then in a hurry when they're close to graduating, they're like, "oh, we gotta do this obligation" and they do it after the fact." (P11).*

This scenario is all too familiar, and even for PIs and graduate students who are proponents of data sharing, it is easy to see how the work required to graduate takes precedence over data management and sharing.

When personnel times are out of sync with the project's lifetime, the timing of data sharing is affected. Personnel times are in sync or fall out of sync as a result of their continuation or rather discontinuation in the PI's research laboratory. Some common examples of discontinuation include graduating or changing jobs. Discontinuation can cause minor to significant delays in data availability. The data management work that key personnel done have a lasting impact, even after the individual in question has moved on to new professional endeavors. One interviewee who described their PI's approach to data management as "hands-off" relied heavily on her predecessor's work which was used as an example for her to follow when producing the documentation required to deposit their data. She shared:

"I was responsible for doing all of the intercalibration reports and. All of that. I would say. At least in our lab. [PI] has a pretty hands off approach in terms of like paperwork and stuff like that. So yeah, I built off of some previous reports from a student who had submitted cobalt data, for I think a different GEOTRACES cruise in the past, which was really helpful." (P28).

The data management and sharing work done by project personnel work can shape data availability even after they have moved on. As the time taken to publish data moves out of sync with personnel career timelines, for example when a doctoral student graduates or when a post-doctoral contract concludes, sometimes years-long delays can happen as the example from P24 below illustrates:

> *"Occasionally somebody involved with the data just gets tied up with things. So, for instance, just a few months ago [in 2024], I submitted some data from an Arctic GEOTRACES cruise that occurred in 2015. And although we had submitted most of my data from that cruise years ago. This particular data, the student who had worked on it … she was busy helping me get the rest of the data ready, and then she got her, Phd, did a PostDoc, got a job. Basically, life intervened. And recently, we were using the data for a paper. So we said, "Okay, we really need to get the data submitted." I don't like being that tardy. But sometimes it's just the way things happen… It was a late publication, and she basically had to still convert from Excel to the right format. And there, BCO-DMO was still very supportive." (P24).*

The purpose of the interview example above is not to single out and place blame on personnel for delays in data sharing and availability. As almost all DMPs note, the PI and co-PIs are ultimately responsible for overseeing and coordinating data sharing. Rather, it illustrates how disruptive it is for data sharing timelines when key project personnel depart from their current role. It is important to highlight that in these cases personnel are being asked to do unpaid work for their previous role in their new role with sometimes many years spanning between the two. On a more positive note, the example shows that even though data sharing may no longer be "timely" and fall outside the timeframes specified by policy, this does not necessarily mean that they are not shared. Publications tend to be produced years after the original funded project has stopped receiving funding. Currently, data sharing is strongly synchronized around publications, I will discuss the implications for data availability in the next section.

6.3.2 “We like to first publish and then share”

Many researchers tend to think about data sharing when they are about to publish, instead of during data collection or other active phases of the project. Likewise, many DMPs do not refer to the policy’s timeframe for public data availability, “two (2) years after collection” and instead substitute the mandate’s timeframe with making their data publicly available “upon publication.” This is for many reasons, such as for fear of being scooped, but of interest to this dissertation is the implications of tying data availability so tightly to publication. The PI whom I quote in the heading shared:

> “*For our work we typically, as many scientists, we like to first publish it and then share. I think the NSF, they have a 2-year time period where they really want to say, after 2 years after production of the data you really have to share it. So typically our challenge is to make sure that within those 2 years that we have open publications ready. And of course, sometimes that takes a little while.” (P22).*

Having data sharing and availability contingent on publication dissolves the concreteness of the time to data availability ascribed by the policy, as the process of publication itself is highly varied and is not guaranteed on the first manuscript submission. Furthermore, synchronizing data management and sharing labor around publication introduces delays for the project and creates time pressure for data managers.

For example, a data managers recounted an incident of when researchers left data sharing until after their manuscript was accepted for publication, mimicking a panicked researcher, the interviewee continued:

> “*We’ve got a paper in the pipeline, “Help! help!” … You can’t wait until the last minute to say “Oh gosh, I need to get this 50 terabytes of data up and published. Because the paper’s being held up because [the data’s] not there yet. That’s bad. I hate that. I basically have to drop everything and go do what I call a fire drill.*” (P13).

The term, “fire drill,” used by the interviewee is apt because from the researcher’s perspective, this is an urgent or emergency situation requiring an immediate response. And much like a firefighter, to address the situation, the data manager has to disrupt their work, rush in, and put out this metaphorical fire. However, reflecting on this “bad” behavior the interviewee admits that “that hasn’t happened too much lately. It used to be a problem several

years ago, but it's gotten way better … our scientists have gotten a lot better about getting me involved early in the process" (P13).

Almost all PIs when asked what they would do differently in their next project replied with some variation of thinking about and doing data management and data sharing "earlier" in their projects. One PI shared that "if we still go through BCO-DMO for future data archiving, I think I would send them data way in advance, before our paper got published. Probably as soon as we submit our manuscript, we will send our data for archiving" (P23).

Timing the public availability of publications and the data associated with the publications is tricky because as mentioned earlier, additional labor is required that are not mere extensions of existing data work. The synchronization of data availability and publications aligns and meets journal requirements for data sharing. For oceanographers, one of the key publication venues is the American Geophysics Union and its journals which require that publications share their data publicly with a DOI. However, tying data to publications further entrenches the end-loaded practice-time profiles of data management and sharing as it signals to researchers that when they publish is when they should be doing the work of data management and sharing. However, researchers are realizing through their past experience how important it is to do data management and sharing at the right time, which is the focus of the next section.

6.4 Making time at "Right Time" for Data Management and Sharing

In the preceding two sections, I presented how PIs understand data management and sharing as separate from the active project time, and that most PIs discussed data management and sharing as something that is tied to the ending stages of a project. In part, this understanding, can be attributed to the OCE policy that mandates that data be made available "2 years after collection and/or acquisition" in a typical three (3) year grant, assuming that data are collected in the first year means that the timeframe for data access coincides with the project ending.

In the third section of this chapter, I move from absence and its consequences towards a focus on the temporal dimension of timing. I also articulate the importance of timing as competence of a practitioner. This is for several reasons. First, interviewees, both researchers and data managers alike, repeatedly emphasized the importance of doing data management "in advance" or "earlier." This suggests that there is both a perceived temporal misalignment between when data management *is* being done and when it ideally *should* be done. This

sentiment also suggests that there indeed is an ideal time, or as we'll see, ideal times, for doing data management.

From a theoretical perspective, timing is one of the eight temporal dimensions of a timescape (Adam, 2008) and became salient as I re-visited the coded DMP segments and interview transcripts in the later stages of my analysis. As it relates to Shove's practice-time profiles, timing is integral and featured in the definition. Shove, drawing on Zerubavel's seminal work in the sociology of time, notes that "we have relatively fixed notions of what constitutes "the proper time" (Zerubavel, 1981, p. 8 in Shove, 2009, p. 25). Practices have quite rigid conventions around timing, depending on the context "taking too long, finishing too quickly or doing things in the wrong order signals incompetence" (Shove, 2009, p. 25). Importantly, for information science scholarship, this subsection argues that it is not only that researchers lack information or even necessarily training on data management and data sharing, but that there is also a temporal misalignment in researcher's current practice-time profiles (end-loaded) missing in the discussion. Making data sharable requires that practice-time profiles be more integrated throughout the entire active phase of the research.

Although almost all researchers interviewed advised others, or their future selves, to do data management "ahead of time," earlier does not automatically mean better, nor does it guarantee that delays won't occur. A competent understanding of the ideal timing conventions for data management also involves knowing when *not* to do things too early. The emphasis on doing data management earlier surfaces repeatedly in my data reflects the current end-loaded data management practice-time profiles. There are ideal moments throughout the lifecycle of a research project that shape whether data management is time-consuming or not, and the intensity i.e. how much time and effort it takes.

A data manager at a large research center expressed that they have seen researchers slowly come to recognize this and is reflected in their data management practices. They said:

> "*the best time to look at [research data management] is when your project is getting started. I love it when I get called to a meeting. They say, we're gonna do a whole bunch of these [simulation] runs, we're going to look at this aspect of the science, how best can we do this? Ask me at the beginning rather than wait until the project is done.*" (P13).

To get to this stage, researchers had to learn through their previous, often stressful, experiences that they should prepare their research data for sharing much earlier than when they expect to share that data.

When this type of work is routinely deprioritized, and out of the normal temporal rhythms of how researchers do their work, how, then are researchers avoiding the last-minute push? Based on interviews and DMPs, researchers are integrating data management and data sharing work throughout the course of the active science portion of their project. In particular, thinking about the desired result of the mandate, i.e. when and where data is to be shared, from as early as upon receiving funding. The ways that the integration of data management for sharing is presented in depth in the next section.

6.5 Towards Integrated Practice-Time Profiles for data management

In the last section of this chapter, I present three ways that researchers are developing practice-time profiles for data management and sharing that suit their data, research aims, and workflows. In other words, the deliberate efforts made towards reconfiguring timing may result in developing a practice-time profiles for data management. The three examples differ in their approach, but what they share in common is that they are preventing an end-loading of data management and sharing work by integrating data management into the active phases of their research and specifically into their intermediary research steps, which are usually made invisible in final research outputs such as peer-reviewed publications and even in datasets.

This section is in line with what Mosconi et al., (2019) identify as pragmatic models of research data management, i.e. that there is no universal or ideal practice-time profile as this differs depending on the researcher, the project, and the data in question, but as noted one common factor is an orientation toward timing. That is, they are all ways to do the right data management work at the right time.

6.5.1 Integrating Data Management into the Intermediary Steps of Research

In reflecting on what they would do differently for their next project; several interviewees discussed the idea of thinking about the end goal of data sharing consistently throughout their project. Researchers who once approached documentation as a final chore are now preparing for sharing in the key phases of the research project. For example, starting from the moment that data are collected as the following DMP explains in detail:

> *"[Co-PI] has developed for highly complex field surveys in Guaymas Basin and the Gulf of Mexico. This system, on a dedicated file server in the [PI/co-PI] labs, now holds sample and site information for three major cruises [cruise identifiers removed], and* ***allows unambiguous retrieval of full sampling contex****t, a digital photo of each core immediately after recovery, destination (the specific lab that has received and analyzed the core) and environmental in situ information (Temperature and geochemical gradients) for every individual sediment core. We are using this system extensively to keep track of samples for analysis, comparison and publication purposes, and enter and circulate updates of newly analyzed data and curated data files.* ***This database has allowed us to provide full core information, geochemistry, sampling site metadata, and in situ and ex situ photographs on request, and to make this information quickly available for our collaborators and their current and future manuscripts.****"* (DMP 132 – emphasis added).

The DMP segment above illustrate that though it takes some extra effort in the moment, in the long run it's still easiest to document and create metadata for data at the time of its creation. Formalizing critical data management tasks, by incorporating them into part of the data collection process, prevents these tasks from becoming that class of tasks that one puts off once and never musters up the time to return, as surveyed in the previous sub-sections.

Another key phase is after data are collected but before data analysis begins. As an interviewee explained, for their next project they planned to have the metadata templates, that many repositories provide, on hand as they analyze their data:

> "*I would potentially make sure I had all of the like [metadata] templates and stuff even before I started processing my data, or running samples, if possible. So that I didn't have to reformat things a bunch of times. I just would prepare them as they needed to be for data sharing right out the gate. I think, that would just be a lot easier to keep everything organized, and I could also do more of the methodology reporting as I go. Rather than having to write it all at once." (P28).*

In the example above, we see a shift in when data management is done, which suggests that creating metadata and documentation are becoming part of the researcher's workflow, rather than a separate task. In doing so, they redistribute the work and labor of data management and sharing across the project's key phases, mitigating the time pressure that typically occurs when all this work is done at the end. As preparing to share data becomes habitual and integrated into the research process, activities, like reformatting data or preparing metadata,

develops into a "routine" one that happens during, not after, data collection and analysis (DMP 132)[26].

6.5.2 Redistributing Data Management Work Through Automation

Modelers, who typically are physical oceanographers, or other researchers who are technically proficient with coding, are automating their data management and sharing workflows. One DMP noted that "We already have scripts in place to generate and submit sequence data to the [National Center for Biotechnology Information] NCBI Sequence Read Archive (SRA) in an automated process. Scripts written during the course of developing our bioinformatics pipeline will be made open source and available upon validation" (DMP 207).

For example, one PI, a self-proclaimed MATLAB user, explained that with the help of their doctoral students, they had "developed some scripts that do exporting from the way we're working internally in MATLAB and exports in [Network Common Data Form] netCDF" (P14), noting that the netCDF format is what public data repositories required. Another PI described a similar situation and explained to me that "the data [our project] produces doesn't actually fit very well into that [find out what is that]. It feels not very useful [for us]… so I wrote computer functions that would write those tables" (P6). In this case, "those tables" refer to additional information required by the repository's data dictionary. A third PI, also discussed adapting their workflow to meet repository requirements, sharing, "this data is literally sitting in a data structure in MATLAB so I wrote a script in MATLAB that would make is so that NCDC likes it" (P9).[27]

Though using technical artifacts to automate work has long promised to save time, scripts do take time to produce. The same PI explained that "it takes a lot of time at the

[26] From DMP 132 the following DMP segment describes how the research team has developed a routine for producing documentation and metadata associated with collecting sediment cores during cruise-based fieldwork, instead of a later point in time. "*Over three cruises, the proponents have developed a shipboard routine where every single sample and sediment core is immediately after retrieval catalogued, photographed, and recorded with full sampling context and time (using Alvin frame grabber images and data). Cores and samples enter the curation procedure after Submersible Alvin returns (ca. 5 pm) and are processed and recorded (digitally and in written lists and backup prints; paper has its place and is less fragile than most computers) before they are divided up for shipboard experiments, geochemistry, microbiology etc. in the daily science meeting after dinner."* (DMP 132) "Alvin", here refers to the Human Occupied Vehicle named Alvin that is part of the Woods Hole Oceanographic Institution's National Deep Submergence Facility (WHOI, 2025).

[27] NCDC refers to one of the policy recommended archives the National Climatic Data Center. Today it is known as the National Centers for Environmental Information NCEI.

beginning, but then the end is so much easier" (P9). In other words, automation involves fronting labor that will make the onerous aspects of data management and sharing less taxing. These practices redistribute time across the research process, transforming what would have been a final bottleneck into a manageable task and one that can potentially be used across multiple research projects, as in the case of converting MATLAB to netCDF files.

Automating data management tasks may be a good solution for some PIs, but it is not a silver bullet. Not only does it take time to code scripts, it may take even more time and experience to know which aspect of data management to automate and to what end as is discussed further in Chapter 7.

6.5.3 Anticipating Future Reuse in the Present

Finally, some researchers are developing integrated practice-time profiles by anticipating the needs of future reusers during the active phases of their own research. Researchers are anticipating and integrating reusability into their data practices not only to reduce their own personal workload but also take into consideration the data practices of future reusers.

For example, one evolving norm in sharing climate models is for researchers to share scripts and programs to run the research data alongside the data itself. In other words, reusable data is no longer just about data, but includes the scripts, software, or setup needed to run that data. As one oceanographer explained:

> *"if the goal is to make the data easily accessible, you don't want the user to then struggle in writing their own program and code. It's a lot more efficient if you host the data and here's the script you need to run to access the data." (P8).*

This example highlights two motivations for changing norms of data sharing in this context, efficiency through avoiding duplication of effort in writing programs and code and equity through making reuse easier for a broader group, not only those who can write their own programs and code. This quote also further shows that some researchers are orienting their present labor toward the needs of future reusers and preemptively integrating their imagined needs into their data practices.

Another example comes from a researcher who repeatedly responded to data reuse requests. The researcher works with history files but shared that researchers who reach out with data requests work with time series data. In the past, this researcher has had to respond to

data reuse requests and “scramble” to convert their data into time series. Having experienced such data requests multiple times, he decided that this was occurring frequently enough that it was worth developing “python-based scripts to do the conversions [from history files to time series]” (P1). Together, these examples reveal how integrated practice-time profiles enable researchers to configure their present-day data practices with future publics in mind.

6.6 Conclusion

This chapter has shown that data management for sharing is not only constrained but actively shaped through time. While data policy and researchers assume that data sharing will be an extension of existing data work, in practice, researcher’s experiences reveal otherwise. Data management for sharing, instead, tends to follow end-loaded practice-time profiles, in which the data work for sharing is deferred until the final stages of a research project. The sunset stages of a project are also oftentimes when key personnel depart or when time is the most limited.

Some researchers are beginning to develop situated ways to manage this by integrating data management tasks into the active phases of their research. These emerging integrated practice-time profiles vary, but a common feature that they share is that data work is being distributed across the project’s timeframe. There is no singular or universal ideal time for doing data management instead there is a range of proper times that are shaped by the project’s data type, scientific goals, researcher’s experiences and disciplinary norms.

Drawing on Shove’s (2009) concept of practice-time profiles, this chapter contributes to a grounded understanding of research data management. By highlighting time and temporality, it argues that time is not just a background constraint but a central and contested factor shaping data management and sharing practices.

Chapter 7. Temporal paradoxes in and of data infrastructures

7.1 Introduction

Time is understood as a base level tension for infrastructure development (Edwards et al., 2007). A classic example, that is also salient in this study, is that between short-term funding decisions and the longer time scales over which infrastructures are thought to grow and

establish themselves (Edwards et al., 2007, p. 8; Ribes & Finholt, 2009). Generally, in cyberinfrastructure scholarship, building, development, repair, and maintenance are key areas of focus as one key assumption is that infrastructures develop into stable and durable configurations of human and non-human actors. For science cyberinfrastructures, these relations are key for ensuring that knowledge and data are maintained and circulated across spatial and time scales. When exploring the temporality of ocean data infrastructures, we see that time is more than a base level tension. Not only does it occur at several scalar registers, i.e. PI - infrastructure, and infrastructure development, but time is less tension and more paradoxical, in the sense understood by Serres (Serres & Latour, 1990) and Radin (2015).

In this chapter, I draw on key concepts from STS and infrastructures studies scholar, including collaborative rhythms, reverse salients, the long now, and convivial decay to develop the argument that the temporal lives of data infrastructures in my field site of oceanography is shaped by a series of paradoxes or contradictions at various scales.

As presented in Chapter 4., the infrastructure for sharing ocean data, albeit uneven, with certain areas more developed than others, overall can be characterized as established and mature as it has its roots in the WOCE. This point of oceanographers being well-positioned to share their research data, and meet the demands mandated by the data policy, was something that was brought up by several interviewees. However, this chapter shows that even when infrastructures appear stable, there still exists friction, uncertainty, or very real possibilities of institutional retirement.

In the first section, I focus on the collaborative rhythms between researchers and data repositories and show the alignments and misalignments between researchers and data infrastructure. The temporal paradoxes at play here are that infrastructures are experienced as frictionless when researchers have established practice-time profiles for data management and sharing. Moreover, large research programs are more likely to receive sustained support. In the second section, I extend Marisa Cohn's (2016) "convivial decay" and present how researchers and data managers anticipate infrastructural demise and make concrete plans for this scenario. Anticipatory convivial decay occurs not only in aging systems but also in relatively new ones as well. Through anticipating and planning for the end, this supports infrastructure and data persistence. In the last section, through examples from WOCE and from discussion with interviewees, I present the ways that semantic work, i.e. metadata data,

standards, documentation, are a recurring reverse salient in the development of oceanography's data infrastructures. The classical understanding of reverse salients is that once these are solved, the infrastructure can continue to grow and develop. However, as I argue using the example of evolving standards for diverse ocean data types, that cyberinfrastructures may be able to develop even with unresolved reverse salients.

7.2 Paradoxical Rhythms of Infrastructural Support

Jackson et al., (2011) have noted that tension arises when collaborative rhythms of different human and non-human actors are misaligned or out of sync with each other. In this section I explore the contradictory dynamics of the collaborative rhythms between PIs and data repositories that the data policies that mandate that they engage with to share their research data.

First, infrastructures are experienced as frictionless when researchers have already established practice-time profiles for data management and sharing. Critically, timing plays a key role in misalignments between PI and infrastructure collaborative rhythms. Large research programs, that are more likely to share their data (Borgman et al., 2016), receive more sustained support. Second, whilst repositories accelerate certain aspects of data management and sharing, they simultaneously delay other aspects. This second, section illustrates that data infrastructures are not experienced in a universal way, as exclusively fast or slow, instead they speed up and facilitate some data management and sharing processes while lagging others. Third, larger research projects and programs that are well-funded are likely to receive more infrastructural support, when it is known that these types of collaborations are already more likely to share their data.

7.2.1 Alignments and Misalignments in PI and Repository Collaborative Rhythms

When PI and repositories timelines align, repositories facilitate data sharing. Many interviewees shared that they had a great experience with BCO-DMO. One interviewee shared, "I don't know whose idea it was to create BCO-DMO, but they deserve a pat on the back" (P24) and described their experience of depositing research data there as "wonderful." Mid and late-stage career PIs who worked in the field pre BCO-DMO had could appreciate how helpful data repositories are:

> "*I think [sharing data has] always been a requirement. At least, since I've been doing NSF work. It was NODC you were supposed to submit data to and their system was kind of impenetrable. I mean, I'm not sure I ever submitted data to NODC. I would try to put it labeled in as an appendix in papers because I did feel like I had a responsibility [to share data]. So as we have now moved to this BCO-DMO system it's become much easier."* (P24).

However, in cases where PI and repository rhythms are misaligned, these infrastructures introduce delays along with support. This section builds on the previous chapter's core arguments that the absence of established practice-time profiles and timing is a central element of the competent performance of data practices. I argue that timing is likewise critical for aligning the collaborative rhythms between researchers and repositories. Without this alignment, even well-funded infrastructures may struggle to effectively support data sharing.

Data managers echo the importance of seeking guidance at specific key moments during the active phases of the project rather than waiting until the final stages. According to one data manager, the ideal sequence of events and their timing is as follows:

> "*[PIs] need to talk to that data repository ahead of time, like when they're starting their project, to get some idea of how they're going to format their data and design their workflow, so that that makes sense. Whereas at the end of the project, if that's when you're trying to wrangle all your data and get it into the right format, it makes it a lot more difficult*." (P2).

This description of an ideal sequence highlights early coordination and reinforces the argument that timing is not just a logistical issue but a structuring force in data management and sharing practices. When researchers defer engagement with repositories until the project's ending phase, they often encounter greater difficulty, even when support is available. In contrast, early engagement allows repositories to play a more active role by providing guidance on ideal workflows or suggesting specific data formats that anticipate submission requirements. This account highlights a temporal misalignment as repositories are design to support sharing, but their ability to do so effectively depends on when researchers first engage with them.

As discussed in greater detail in Chapter 6, a sign that researcher's practice-time profiles may be end-loaded is that they begin to manage and share their research data only

after publication. This temporal orientation is often out of sync with repository workflows, leading to delayed data sharing. For example, one PI shared:

> "*So our paper just got accepted, I think, in December. But it's going to take BCO-DMO a few months to accept the data archiving. So we have to actually find another data archiving place to host our data in order for our paper to be published." (P23).*

Misalignments, between PI and repository, don't just result in a stressful project close-out or mere delays, but actively creates additional work. Here, the PI and their student deposit their data in an alternative repository that could provide a DOI for their dataset within days to meet the journal's submission requirement.

7.2.2 Facilitation and Drag

Data managers I spoke with shared that they manage a lot of these "standard" data through automating the data verification process for these data types. Data verification is what most data managers and curators spend the bulk of their time working on as one interviewee shared:

> *"At [Marine Geoscience Data System] MGDS, we had so many great automated processes for validating and cleaning data. Where I work now, the most time consuming thing is validating the metadata and because we are not the level of MGDS. We do not have the same investment in validation, and so many things that could be and should be automated, they are now still manual. And that is the most time consuming thing. At MGDS, as a data curator, you'd get this submission and 80% of the submission you can validate through these scripts, and you can do it within an hour, depending on the size of the data set. But now a curator has to sit and manually go through everything, which is just kind of a lose-lose. It's time consuming and there's inherent error in that." (P34).*

As illustrated by the quote above, repositories differ in the level of investment into data validation. And as discussed in Chapter 5, data have their own lifecycles and afterlives depending on material conditions, disciplinary norms, and institutional arrangements. Here, I expand further on their institutional arrangements. Depending on the data type, repositories and archives approach their deposition and maintenance differently. For instance, "standard" oceanographic data, such as CTD, underway data, and other "unmanned instrument" data as well as data coming from large research programs have strong infrastructural support.

Instrument data are typically large in volume and collected continuously, making them well-suited for automation. As one data manager explained, the focus of their work involves "creating data pipelines and making sure that those data pipelines are running smoothly [to] make sure that those data are operational and running, and that those data streams are online" (P25). Although automation does not eliminate the need for verification and processing work entirely, it shifts the work of data managers. Rather than manually validating each dataset or engaging closely with individual researchers, their primary responsibility becomes maintaining the "data pipeline" that supports ongoing data availability.

This shift highlights how different data types shape both the temporality and character of data work. For instrument data, automation enables repositories to handle data depositions as routine processes rather than a one-off task. Another data manager described how NOAA's buoy data, transmitted via satellite and automatically archived post-quality control, eliminates the need for meeting individually with scientists:

> "*for example, if it's a NOAA buoy system. The buoys will electronically transfer data via satellite phone back to shore. And those data after they've been quality-controlled, we set an automatic archive up to our system and it automatically gets archived. So you don't need to do the one-on-one with scientists, because it's a routine thing."* (P29).

These examples illustrate how automation introduces a different kind of temporal contradiction; instrument data are well-suited for automation because they are standardized and align with repository workflows. This alignment facilitates routine ingestion and consistent processing.

However, the automation of certain types of data but not others reveals a temporal paradox. While automation has long promised to save time and increase efficiency, the tools that enable it, i.e. scripts and software, require substantial time and labor to develop and maintain. Repositories often must build their own custom scripts as there are no commercial products available. As one data manager put it, "it can take weeks, you know, or months even depending on the type of data. Certain types of data are very easy to work with, like temperature at a fixed point, is really easy to work with. We can do that in a couple of hours" (P25). This example illustrates a tension whereby the benefits of automation are reaped for already-standardized data, which are also the least cumbersome to automate. Meanwhile, non-

standard data, those that would most benefit from automation, require more effort to process. The effort, in terms of time and labor, may be beyond what repositories can afford. As a result, infrastructures built to streamline data sharing may be slowed down by the very processes intended to accelerate them.

When asked how repositories choose which aspects of data management and sharing to automate, interviewees emphasized that some tasks cannot be automated. Biological datasets, especially those that needed to conform with the Darwin Core metadata standards required human processing as "it requires someone really sitting down and looking at the data critically" (P25). Another data manager observed that is "the data sets I get most of the time are not uniform" (P32). In other words, automation is best suited to uniform datasets that already conform to established standards – but those are precisely the data that demand the least verification. To restate the temporal paradox another way, automation most benefits data types that are infrastructurally supported but uncommon, non-uniform, and non-standard data require the most support but resist easy automation.

7.2.3 Aligned infrastructural support: The BCO-DMO - GEOTRACES partnership

Large research programs can secure additional support through official partnerships with data repositories. GEOTRACES is one such research program. As noted in its DMP, "BCO-DMO has agreed to partner with us for all aspects of data-management, as with all other US GEOTRACES cruises, and as lead PI [lastname] will consult with them regularly throughout the planning process" (DMP 11). This partnership provides GEOTRACES with tailored support, including documentation guidance "BCO-DMO will provide a template to illustrate the types of information required for this report" (DMP 11). The report in question is typically the responsibility of the cruise's Chief Scientist and is required for all research conducted on federally owned research vessels. While the DMP does not suggest that BCO-DMO will author the report, it makes clear that repository staff will play an active supporting role in its preparation.

Interestingly, GEOTRACES plans to rely on BCO-DMO to facilitate internal data sharing among its distributed team of PIs. This is significant because internal sharing is not addressed by the OCE data policy which focuses on the public availability of data. While other DMPs also describe internal sharing, they typically position the PI as responsible for it.

In contrast, the GEOTRACES partnership delegates part of this responsibility to BCO-DMO. This suggests a shift in the repository's role from being the steward of data after the research project ends to stewarding data during the research project. In the case of GEOTRACES, internal data sharing to cruise participants is facilitated "in partnership with BCO-DMO, via a password protected web-based server" (DMP 11). A different GEOTRACES DMP, similarly states that BCO-DMO will work with GEOTRACES' Project Management Team to make sure that "as soon as additional lab-generated trace metal data are available … [these] data [will be made] available to collaborators … to enable earliest possible inter-comparison with complementary underway and laboratory-generated data sets*"* (DMP 11). As written, BCO-DMO's involvement then ensures not just the internal distribution of data amongst the distributed GEOTRACE researchers but also its timeliness implied by "earliest possible inter-comparison" (DMP 11).

The shift from moving towards internal distribution of data sets is notable as traditionally we see data repositories playing a key role in the distribution of finalized datasets to the public. In regard to the latter, data repositories play a role in the *timing of data access* by honoring moratoriums and embargoes on data, helping PIs and their teams share their data in "a timely manner" and therefore further helping develop and manage practice-time profiles for researchers who are already likely to have them.

Large programs like GEOTRACES receive highly coordinated, tailored support from repositories like BCO-DMO, including assistance with internal data sharing during the research phase. While in some ways this is an example of the ideal infrastructural alignment, it reveals a temporal paradox that those research groups and programs that have more resources are more likely to already have data management support and share their data. Collaborative rhythms, then, are unevenly synchronized with some types of projects benefiting over others.

In this section I have discussed how temporal paradox manifests between repository and researchers and their projects. In the first example the site of contradiction is between researcher's data practices and repositories, the second site of contradiction is between research data and repositories, the third site of contradiction is between research projects and repositories. In the next section, I will discuss how temporal paradox manifests at a different

site in my data, between the lifespan imagined in data policies for repositories and the lifespan anticipated by the researchers who deposit data.

7.3 Anticipatory convivial decay and the paradox of the long now

Infrastructures are designed for long-term persistence. In practice, interviews and DMPs provide insight into the uncertainties surrounding how long we are talking about when the "long-term" is discussed in this context. DMPs and interviewees alike reveal a vague understanding of the timeframe for data access and preservation that data infrastructures are supposed to achieve, with estimates ranging from 3-75+ years. At the same time, interviews shed light on the anticipatory practices that researchers and repository staff engage in with an aim to plan and prepare for a repository's institutional demise.

7.3.1 The uncertain long tail of public data access

NSF-funded repositories in oceanography are responsible for public access in the short term as surveyed in the previous section in this Chapter and elsewhere in this dissertation. At the same time, they are responsible for maintaining the "long tail of public data access" (P4). Unlike research projects or data, the timeframes that data infrastructures operate on are at a much larger time scales, 5-75 years into the future, depending on the institution in question, its centrality to the discipline it serves, as well as the funding it can procure. Despite the NSF policy mandating that data be deposited in specific repositories, interviewees and DMPs alike were concerned about the actual longevity of data infrastructures. For instance, DMPs routinely mention that repositories will ensure that research data will publicly be accessible for a mere "3 years" into the future. Their worry is driven by the possibility that after passing on the responsibility of data stewardship to repositories and archives, these institutions are unable to fulfill their stated mission before meeting their institutional demise.

Data managers and other interviewees who worked at data repositories discussed the possible short lifetimes of their institutions and shed light on how the organization planned for this unlikely, yet distinctly possible, outcome. Discussions of institutional succession planning in interviews were about what would happen to that repository's data holdings if they were to cease operating. The reason why this possible scenario is given as much thought as it is by

interviewees was due to the uncertainty of available future funding from the NSF as a director for data management explained:

> *"But I think the biggest problem is that there's a mandate to publish the data, and you publish the data in a database that looks fine. And [then] they don't get funding from NSF…what happens to your data? That's a big problem, if you're looking at four- to five-year funding cycles for databases in the U.S. That's not security that your data is going to live long past that funding cycle."* (P4).

Even for well-established repositories associated with long-running research programs, the 5-year funding cycles for them still result in uncertainty about the amount of funding, if any, they will be given the next funding cycle.

7.3.2 Anticipating Infrastructural demise

The uncertainty about funding, or what Karasti et al., (2010) have identified as "project time", leaves interviewees with a feeling of precarity. An open question then is, how are these data infrastructures dealing with the funding uncertainty and what are the planned courses of action? Interviews shed light on the anticipatory practices and draw on Cohn's "convivial decay" to describe the ways that researchers and repository staff engage in with an aim to plan and prepare for a repository's institutional demise.

In this study, researchers and repository staff prepare for convivial decay through making lots of copies and entering formal and informal agreements with archives or other institutions that are perceived to be longer lasting. Likewise, DMPs too, albeit to a lesser degree, reveal anticipatory convivial decay occurring as researchers plan to deposit their data in multiple repositories. Where my study differs from Cohn's study of the geriatric space infrastructures is that convivial decay does not happen only in geriatric infrastructures but in newly formed or "youthful" infrastructures. Paradoxically, end-of-life planning is more likely to ensure the persistence of data as anticipatory convivial decay becomes a future-proofing strategy.

In conversation with interviewees, the subject of the uncertain longevity of data infrastructures was broached frequently. In some cases, PIs have several options to choose from when considering which data repository their data should be deposited in and the lifetime of the data repository can be a factor that PIs considers. For one interviewee, the

longevity of the repository was a key consideration in choosing Zenodo for their data. Explaining their reasoning, the interviewee explained that Zenodo, a European repository, is tied to the European Organization for Nuclear Research (CERN) which hosts the Large Hadron Collider (LHC). The LHC is the largest particle physics laboratory in the world, and as the largest of its kind in the world, the interviewee surmised that CERN was "going to be funded for a while, so the database [Zenodo] will be there" (P4). CERN's funding, the interviewee presumed, and by proxy Zenodo's funding was guaranteed to a certain extent into the future due to the organization's operation of the LHC. The example above highlights that PIs do care about that the longevity of their data's afterlives and may make choices about where they share their data, even if it doesn't follow the OCE policy's recommendation.

In a similar vein, some DMP authors write about a "dual repository plan" for their data. This plan involves PIs identifying two suitable repositories to deposit their data in. For instance, one DMP author wrote "due to uncertainty for the long-term sustainability of the NCBI SRA" they would make their data available also available through a secondary location. The secondary location may be another data repository but oftentimes this is the PI's website. As discussed in Chapter 5. Sharing and depositing through multiple places creates a fragmented landscape for research outputs and has implications for discoverability. At the same time, the example presented here differs in one key way, in that in chapter 5. Different research outputs were shared in different ways, here we are discussing the same data being shared in multiple venues. This duplicated deposition of data is DMP's author's way of planning for anticipatory convivial decay.

Data repositories and infrastructures also engage in the practice of duplication, which my interviewees called "mirroring" but at scale and at a higher frequency. Mirroring involves copying datasets to secondary locations at regular intervals, often with formal institutional backing. One DMP for a large instrument array, for example, wrote that "quarterly snapshots of the data will be archived, with a DOI, at the [University of California San Diego] UCSD library" (DMP 284). This arrangement ensures that data are not only stored at the repository but also backed-up through redundant copies housed at a university library. This is an example of a deliberate strategy for safeguarding data in case of potential infrastructure demise.

Mirroring practices with national archives are even more robust. NOAA, for instance, mirrors both active and retired databases to enable long-term access. The Ocean Biodiversity Information System (OBIS), still up and running, is one such example: One active database currently mirrored at NOAA is the Ocean Biodiversity Information System (OBIS). The data manager who oversaw the process shared:

> "*OBIS have backups and they archive stuff on their own [but] we also have an auto archive of all OBIS data. They approached us because they would like to see this data archived in a U.S. Federal Archive because [then] it's committed for a minimum of 75 years and [NOAA NCEI] is a long-term archive that'll be supported. So that's actually something that, must have been like 2 or 3 years ago, now, I set that up.*" (P29).

This example shows that mirroring is a form of risk mitigation, stewardship of the same datasets are being held across multiple institutions in an effort to preserve these data beyond the lifespan of a single repository. OBIS approached NOAA specifically because its status as a Federal Archive is understood as a proxy for longevity.

Whilst OBIS is still operational, one retired database is the Gulf Science Data Repository (GRIIDC), that supported environmental data collected in the Gulf of Mexico. At the time of conducting the interview in 2025, funding for GRIIDC had ended. One NOAA archivist explained:

> "*GRIIDC [holds data] from a lot of the deep water horizon projects. They had to archive GRIIDC and so, we just pulled their archives into our archives because theirs isn't funded anymore. So it was a good thing because we ingested all that data, and now it can be preserved, because if things stop getting funded, they're not preserved as well*" (P32)

This example highlights the temporal contradiction in anticipatory convivial decay. Where recognizing that infrastructures designed for long-term stewardship may themselves be short-lived and therefore require planning for their own obsolescence. Mirroring and archiving agreements are proactive efforts to extend the afterlife of data beyond the lifespan of the repository and counters potential data loss due to institutional precarity.

As discussed in Chapter 5, the planned afterlives of data vary in their degree of certainty. Some are relatively concrete, for example, plans for sharing physical samples or model outputs, while others, such as those involving uncommon data, are more speculative,

often relying on infrastructures that do not yet exist. Similarly, forms of anticipatory convivial decay fall on a spectrum. They range from formal agreements, which provide explicit commitments and institutional backing, to informal arrangements. Informal arrangements between repositories and archives are akin to promises. Unlike formal agreements, informal arrangements lack enforceability, and just like promises, there is no guarantee that the agreement will be upheld and may not materialize as planned.

One interviewee shared that their organization was in the midst of "talks" with NOAA about having a copy of their data stored there. As in the case of GRIIDC above, if the repository were to be no longer funded or otherwise meet its demise by other means, the issue of migrating the existing data will have been already addressed. However, even with these discussions progressing well, it remained unclear to the interviewee just how long their data would be retained by the national archive. On this last point they mused:

> "*I haven't gotten great answers, because I think it's an area that nobody really understands what would happen. What we are doing though is we are archiving our data, but … we're working on the agreement today to start archiving our data at [NOAA] NCEI. They will keep the data, I believe 75 years post our retirement*" (P21).

Even so, the interviewee shared that the archived data would be on tape, as such, they mused that it would be a problem for accessibility. In other words, whilst long preservation may be achieved in this way, there are questions about just how easily accessible the data will be for future reusers given that the data would be archived on tape.

National archives are not the only institutions that repositories are considering. One interviewee, a data manager working a smaller data repository, noted that they had entered into a formal agreement with the Library at the University that they are based out of. The agreement, similar, to the other examples above, provides some degree of certainty that the data holdings, all the data sets currently overseen by the repository, would be transferred to and hosted by the university library.

Sometimes, the futures of data repositories are less concrete. A former director of a data repository shared that, in the case of their ending, they had similarly planned to transfer their data to a national archive. However, the agreement was described as a handshake deal and speculated "would the [archive] accept it? I don't know, but that was our plan. Nothing signed with the [archive], I think" (P19). Given the informal nature of the deal, it leaves the

matter of the long-term availability of the data repository's data holdings in a tenuous position and speculative even when the institution has engaged in anticipatory convivial decay.

This particular data repository, however, enjoys a certain degree of guaranteed longevity given their central role in their discipline " I know that [discipline] can't be done without [repository] right now" (P19). However, we see here, yet again, the base level tension of time of short timescales of funding versus the longer timescales of infrastructures "NSF could make the decision, say, "Well we don't think [discipline] is important anymore … I wouldn't anticipate that the archive would continue. Without funding, it can't continue" (P19). Despite the centrality of this repository to their subdiscipline, they still felt the need to plan for the end as their institutions lifetime is ultimately guaranteed on the timescales of "project time."

As it relates to the long now, in this moment, we can only speculate whether practices of anticipatory convivial decay, those anticipating the end and planning for convivial decay of their institutions and their institutions data holdings, will make it the time horizon of the "long now", e.g. 200 years. But through the examples in the chapter, I have shown the various ways that this work is being done in order to at least give the data a possibility of doing so. As such, the paradox here is that those anticipating the end, are more likely to achieve longevity.

By anticipating the end of institutions presumed to be durable, a second-level paradox emerges, one that concerns the lifetimes of data infrastructures themselves. Unlike earlier generations of large technical systems, or traditional public infrastructures like electricity, water, or even the internet, cyberinfrastructures may not endure across centuries or geophysical timescales. Instead, their evolution and possible decline may unfold within the span of project timescales. This compression of infrastructural evolution – formation, maintenance, and de-formation onto shorter timescales complicates a core understanding within infrastructure studies literature of infrastructures as long-lived and stable. In the next section, I build on this argument but examine a different temporal paradox occurring between semantic work, important for infrastructure development, and the ever-changing nature of science.

7.4 The Curious Case of Recurring Reverse Salients

It is well understood that semantic work (Karasti et al., 2010) such as standards for metadata, data, reporting requirements, and naming conventions are critical for infrastructure building

and development. In this section, I will discuss the development and evolution of standards as a persistent reverse salient in oceanographic data infrastructures.

Drawing on Hughes' (1993) concept of reverse salients, I argue that infrastructures don't only grow or develop through solving reverse salients. Reverse salients are components of infrastructure that lag behind and inhibit overall progress.[28] Put another way, reverse salients are bottlenecks that prevent infrastructural development and formation. Reverse salients may be technical but can also be social, legal, political, or cultural (Edwards et al., 2007).[29] An example of a social reverse salient in cyberinfrastructure development is standards for data and metadata (Edwards et al., 2007, p. 23). Standards here are broadly defined as they include requirements for data and metadata, naming conventions for variables, units of measurement and data vocabularies. For specific types of oceanographic data such as underway data and hydrographic data, standards were established during WOCE as discussed in Chapter 4. However, many ocean data types do not, as discussed in Chapter 5 and earlier in this chapter. These are social reverse salients because these are sites where infrastructures slow down in order for data to persist over time.

Elsewhere in this dissertation I've discussed how a lack of standards creates challenges for the afterlives of data and can cause misalignments in the collaborative rhythms between PIs and Repositories. In addition to infrastructural absence, and slowing down the timing of data availability, the lack of standards, but interest in acquiring them, may be indicative of an emerging subdiscipline or research community. As one data manager shared that one of the changes she's notice today is that

> "*People are a lot more excited about standards…. there [are] high level philosophies, but there's really no core set of community standards. And it's because some of these*

[28] Hughes (1993) definition of reverse salient is "A reverse salient appears in an expanding system when a component of the system does not march along harmoniously with other components. As the system evolves towards a goal, some components fall behind or out of line. As a result of the reverse salient, growth of the entire enterprise is hampered, or thwarted, and thus remedial action is required… The causes of the lag can arise from within the system; from its environment, or context; or from some complex combination thereof." (p.80)

[29] From Edwards et al., (2007) "Examples of reverse salients relevant to cyberinfrastructure include: generating metadata (this is an unfunded mandate) … using different equipment and data formats; domain specific data sharing and publication cultures; reluctance of modelers who have been working with a given program to shift to a better one if the learning curve is too steep" (p.15)

communities are so niche. And when I go to conferences and I talk to people, they're just aching for standards." (P32).

Another part of this discussion is that over time, standards themselves change. Part of the difficulty of doing this semantic work is that "PIs may not know what standards are out there, because they're constantly evolving" (P25). Central to the argument of temporal paradoxes of this Chapter, while standards indeed enable infrastructural development and growth, new forms of oceanographic data, from new technologies, reintroduce the same critical problem of semantic work. These bottlenecks are not, and likely will never be, solved once and for all as scientific discipline, methods, and data practices all co-evolve. The contradiction, and my argument, is that cyberinfrastructures may be able to develop even with reverse salients – the development and growth experienced will likely be uneven. This understanding enriches our understanding of infrastructuring and potentially highlights the non-linear development trajectories of cyberinfrastructures.

7.4.1 "That's a niche thing": how "newness" reintroduces reverse salients

Seismologists that work with ocean instruments traditionally work with ocean bottom seismometers. These instruments record the earth's movement in bodies of water. Ocean-bottom seismometers, much like their land counterparts, are stationary and have a single location associated with the instrument's data. What happens when the instrument moves, but standards assume it does not? This was the question that one PI and his postdoc ran into "[our] data is really unique because … a normal operation of a seismology experiment is that, you record data at a point over a lot of time, whereas for us, every bit of data that's collected is from a different place" (P11). Similar to other uncommon data without standards repositories will deem the data out of its purview. However, this PI was passionate about data sharing and reached out to the Incorporated Research Institutions for Seismology (IRIS) to discuss how they could do that. This resulted in a 2-year long effort to turn their "uncommon" data into "standardized" data that would be deposited and shared at IRIS. The PI recounted the following:

"*We had a lot of back and forth with [IRIS] before they even would take our data because they wanted to make sure that it's up to their standards. And because they had never received any data like that, we specifically... and that's a large part of the effort*

of my postdoc. We worked with them to develop a standard upon a standard such that that data would be available, consultable, and archivable." (P11)

Creating a standard upon a standard is above and beyond what is required by the data policy but is what it takes to circulate and preserve "unique" data. As if that wasn't enough, the project was faced with the challenge of reformatting their data. In this case, the floating seismometer instruments are from a private instrument manufacturer, as a result the instruments would output data in a proprietary format. Proprietary formats are not compatible with repository requirements for data sharing; instead requiring open formats. Here, the PI and his postdoctoral research scientist worked together with the instrument's manufacturers iterating through "hundreds of versions of code" (P12) to convert the instrument's native format to SEED and miniSEED formats, accepted by IRIS.

In this case, we see an example where even within an established cyberinfrastructure, like IRIS, the introduction of new instruments reintroduces the need for semantic work and development. This involves not only the creation of "a standard upon a standard" but also the development of software to convert proprietary instrument formats into open-source ones. These tasks require substantial labor and coordination between researchers, repository staff, and instrument manufacturers.

Emerging data types in oceanography such as environmental DNA (eDNA) and those from new instruments, similarly, reintroduce the same bottlenecks of standardization and formatting that were once resolved for more mature data types. Establishing what counts as data, how it is named, and what units are appropriate open questions that are answered by researchers in the midst of their scientific work, as was the case for the seven data streams featured in WOCE.

One data manager described the challenges of working with these new data types:

"*In our office, we're starting to work with some newer types of technologies like environmental data … and we have these imaging microscopic instruments they're called imaging flow cytobots. It's a microscope that's deployed in the water and taking real time images of the phytoplankton… [One] team member is charging 60% of her time on that, because it's really complicated, and it hasn't really been done before… also communicating and coordinating with the international community who are working on imaging phytoplankton data management, because that's such a niche thing." (P25).*

In the quote above the interviewee also mentioned environmental DNA as another example of emerging data types. Unlike phytoplankton images from flow cytobots, researchers are beginning to receive guidance from federal agencies, such as NOAA, through guides and manuals which provides recommendations for how to manage and report "-omics" and eDNA data. "Because more and more institutions are collecting this kind of data without knowing how to handle the data, more guidance at the national level … is really helpful" (P25).

This emerging effort is reminiscent of the WOCE Operations Manuals (Joyce et al., 1994; WHP Office, 1991), an example of a reverse salient being explicitly addressed. These manuals included comprehensive, standardized protocols for collecting, formatting and reporting key oceanographic data types. They offered detailed instructions on how variables should be named, which units of measurement to use, how data should be structured and how quality flags should be assigned. This is an example of a reverse salient being solved as data collected by individual WOCE PIs were to be sent to a DAC to be collated into a single dataset. These standards for reporting data, in particular for hydrographic data, is still being cited in DMPs today.[30]

These examples reveal that standards are not one-off problems but persistent reverse salients that recur most acutely at the disciplinary fringes and frontiers of oceanography. Emerging data types like eDNA or images from new instruments become reverse salients because the semantic work needed to integrate them, i.e. developing formats, standards, vocabularies, lags behind the pace of disciplinary change. Infrastructure development slows in these cases, as repositories must "catch up" through standard-setting, software development, or coordination across research communities. Whether it's building a "standard upon a standard" for moving seismometers or coordinating international guidance for images of phytoplankton, such efforts are critical for these data types to circulate and be preserved.

As oceanography evolves through the co-development of new instruments, methods, and epistemic objects, the same infrastructural challenge, the need to establish standards for unstandardized data, reappears. Researchers and infrastructures alike repeatedly encounter this problem in new contexts. This recursive pattern makes sense, as standardization requires

[30] "Major policy requirements that follow those developed for other large scale hydrographic programs such as WOCE … include: Metadata should be delivered as soon as created, from the planning stage onwards, and should be made publicly available immediately; Shipboard data should be submitted within 1 month of collection, or end of cruise; Cruise or project reports should be submitted within 6 months of the cruise end" (DMP 11)

immense labor, and investing heavily in semantic work too early may be premature or impractical before new data gain broader adoption, as is currently the case for -omics and eDNA data. The paradox, then, is that even as reverse salients for more mature data types are resolved, such as hydrographic data during WOCE, the same kinds of challenges re-emerge in new contexts, making semantic work a chronic and necessary feature of infrastructure development and evolution.

7.5 Conclusion

This chapter examines how the availability of OCE-funded data is shaped not only by the presence or absence of infrastructure but by the temporal contradictions embedded in how infrastructures operate over time. I advance the concept of "temporal paradox" to describe these contradictions as they emerge across multiple sites and scales of data infrastructure. Whilst the NSF data policy envisions repositories as long-term stewards of oceanographic data, these institutions operate under short-term funding cycles and uncertainty, making them institutionally fragile. These conditions shape and constrain how infrastructures both support and delay the circulation and preservation of scientific data over time.

I identify three temporal paradoxes that emerge from my data. First, infrastructural support is unevenly distributed and contingent on the temporal alignment of collaborative rhythms between researchers and repositories. Data sharing proceeds most smoothly when PIs already have established practice-time profiles or when large, well-funded research programs are involved. In such cases, repositories can accelerate data availability. Yet for others, the same infrastructures introduce delays and duplicated effort, revealing how infrastructures can simultaneously facilitate and slow down data sharing.

Second, repositories and researchers engage in anticipatory convivial decay that entails planning for infrastructural demise even as they are tasked with ensuring long-term data preservation. These future-oriented practices are not limited to aging infrastructures but also appear in relatively new ones. Through formal and informal agreements with institutions viewed as more stable, such as archives or university libraries, repository staff and researchers work to ensure that the data they steward will outlive the lifespan of data infrastructures. Paradoxically, it is those most attentive to their own institution's precarity that secures the best chance for their research data to endure over time.

Third, semantic work is a recurring salient of cyberinfrastructure. However, paradoxically it is not a one-time bottleneck that is resolved once and for all. Instead, it is a recurring reverse salient that is encountered again and again as instruments, data types, and scientific practices co-evolve. These reverse salients reappear in new contexts and reintroducing the need for semantic work. For example, researchers and repository staff will have to revisit questions about formats, naming conventions and metadata standards. Taken together, these findings suggest that oceanographic data infrastructures do not persist through stability alone, but also through fragility.

Chapter 8. Conclusion

The aim of this dissertation is to explore the ways that the NSF's data management and sharing policy has affected oceanographer's data practices. Specifically, I examine what happens to data, researchers, and infrastructures under the requirements mandated in federal data policies.

8.1 Primary Findings: Temporalities of data management

This dissertation identifies the way that federal data policy mandates intersect with the diverse and often conflicting temporalities of research data, research practice, and data infrastructures, in oceanography. The three empirical chapters and in some ways the background chapter on WOCE all underscore how time, anticipated, structured and experienced, shapes and constrains the availability and persistence of oceanographic data.

8.1.1 Heterochronous data and their afterlives

The first research question this dissertation addressed is: *In what ways do the timeframes specified in data policies shape the management, sharing and afterlives of research data?* I show how oceanographic data do not move uniformly through the research data lifecycle as imagined by classical prescriptive models and by policy. By understanding DMPs as anticipatory discursive infrastructures I identify three forms of planned data afterlives which diverge from the future envisioned by the data policy.

The secluded afterlives of physical samples reveal how material or epistemic decay diverges from policy. The splintered afterlives of models and model outputs are marked by a fragmented landscape. However, models are iterative and do not reach a stable form in the same way that datasets with DOIs do. To be shared through repositories code will have to be frozen

further creating fragmentation. The speculative afterlives of uncommon data may from afar seem like noncompliance, but I argue that the examples provided show that these are PIs earnest attempts to plan to share their research data under conditions of infrastructural absence. These planned afterlives reflect not only the material and epistemic characteristics of the data but also the present and past infrastructures and disciplinary norms that mediate their anticipated circulation and preservation. This finding challenges the presumed universality of data availability imagined in policy mandates and reflected in current DMPs.

The theoretical contribution of this chapter is the extension of Radin's (2015) "planned hindsight." Instead of a general way that future goals are used to justify present-day data practices, I suggest that the future imagined results in a specific mode of justifying present-day data practices, in this case one that is "exhaustive." Through exhaustive planned hindsight, I show how the OCE data policy implicitly presumes that every research output can be and should be shareable and preservable. This effectively amounts an understanding that all data is equally worthy of stewardship. Whilst an ideal goal, this ignores material, epistemic, and institutional limits.

8.1.2 Practice-time profiles of research data management

The second research question this dissertation addressed is: *In what ways do oceanographer's navigate the timeframes specified in data policies in their data management and sharing practices?* In Chapter 6, I show how data management is not merely temporally constrained but is temporally produced. Whilst data management for sharing and reuse is imagined as extensions of existing data work. When researchers do the work of preparing and managing their data for sharing and preservation, they find this not to be the case. This results in "end-loaded" practice-time profiles for data management, where all the work is done at the end of the project. Some researchers having started to integrate data management for sharing into the intermediate steps of research during the active phases of the research project.

There is no ideal, or one-size-fits-all way to integrate data management for sharing into one's data practices, as this is shaped by the researcher, data, aims of the project, and where as well as how they plan to share their data. However some recurring examples from my data includes, creating routines, automating workflows, and anticipating the data practices of reusers. These "integrated" practice-time profiles align better with the desired goal of data sharing as it is

temporally distributing the labor and data work required to make shareable data across the research project instead of at the end.

Theoretically, this chapter contributes to our understanding of research data management for sharing as a practie with distinct temporal characteristics. Drawing on Shove's (2009) "practice-time profiles," I argue that data management for sharing is a site where time is made and resisted as some PIs shared candidly that they do not see data management and sharing as core to the science they are being funded to produce.

Moreover, this dissertation highlights the temporal dimension of timing, i.e. the "when" of data management for sharing, which in my site emerged as a salient and critical dimension, that influences both the sequence of tasks and the duration and intensity of effort required. These findings highlight that there is indeed a "proper time" for managing data for sharing; but that what counts as "good timing" is idiosyncratic in the same way that, as noted by Mosconi et al., (2019) that research data lifecycles are. By idiosyncratic, I do not mean that it is shaped individually but rather that it is shaped and constrained collectively, by researchers, policies, and disciplinary norms. Then in some ways there are a range of "proper times" and in the context of presently "end-loaded" practice-time profiles, the proper time(s) are "earlier" and "in advance" of when researchers currently understand is the time to do data management for sharing.

This chapter also responds to growing calls within information science to pay attention to time and temporality (e.g. Haider et al., 2022). This dissertation contributes to this body of work by showing how data management for sharing practices are temporally structured and distributed. By examining how researchers lack and create practice-time profiles, this research offers an empirically grounded understanding that time is not only background or context, but is a dynamic and contested/resisted dimension of data management practice itself.

8.1.3 Temporal Paradoxes of Data Infrastructures

The third and last research question this dissertation addressed is: *In what ways do data policy requirements, to deposit in designated repositories, shape the capacity of infrastructures to support the short and long-term availability of OCE-funded data?*

Even as data repositories are envisioned as long-term stewards of oceanographic data, they operate under conditions of short-term funding and shifting mandates and are experienced as fragile. This dissertation introduces the concept of "temporal paradox" to describe how infrastructures are both imagined to outlast the projects whose data they help share and preserve,

and yet are themselves subject to institutional demise and precarity. In this context, planning for infrastructural demise becomes a peculiar but necessary form of preservation. In chapter 7, I trace three temporal paradoxes that characterize how data infrastructures in oceanography currently operate. The first paradox is about how infrastructural support may be timely but only if PIs already have established practice-time profiles or there are already exists resources or historical investment. The second paradox is about how infrastructures must imagine and anticipate their own institutional demise which becomes a peculiar but a form of preservation likely to result in the persistence of data through long time periods. The third paradox is about how semantic work are a persistent bottleneck - social reverse salient - in infrastructural development rather than a solved problem.

Theoretically, I build and bring into conversation three concepts from infrastructure studies: Jackson et al., (2011) collaborative rhythms, Cohn's (2016) convivial decay, and Hughes' (1993) reverse salient to examine infrastructural paradoxes and how they unfold, or percolate, across time. Together, these sensitizing concepts help surface the temporal paradoxes that characterize data infrastructures in oceanography do not develop linearly or durably but are temporally complex and uneven arrangements that require constant realignment across multiple scalar registers. This chapter finds that infrastructures do not persist through stability alone.

Through Jackson et al.'s (2011) collaborative rhythms, I show that infrastructural support is unevenly distributed, accelerating some workflows while delaying others. Rather than exclusively speeding up or slowing down the pace of data management for sharing, data infrastructures speed up whilst lagging other aspects, often both simultaneously.

Since infrastructures may meet their end in cared-for ways in the vein of Cohn's (2016) "convivial decay" but also in careless ways. Cohn emphases demise in the context of a geriatric infrastructure, however in my case I find a distinct phenomenon: the ways in which data managers and others working at infrastructures anticipate obsolescence even as they build for longevity. Furthermore, this anticipatory convivial decay is not only in geriatric infrastructures but in relatively "youthful" ones too.

Finally, by revisiting reverse salients with a temporal lens in mind, the chapter shows that bottlenecks in infrastructural development are not resolved once and for all, but recur as new data types, scientific practices, and instruments all co-evolve.

8.2 Future Research Directions

There are many ways that future research, whether in information science, STS, or infrastructure studies can extend the study presented in this dissertation. Here I discuss just three possibilities.

An intuitive next step would be to conduct comparative studies across scientific domains. It is well known, disciplines vary in terms of their data cultures, research practices, epistemic artifacts, infrastructural support and institutional mandates. As such other domains would enrich the typology of data afterlives presented in chapter 5.

Comparative research could also examine how PIs and researchers in other disciplines interpret the data management policies and integrate data management for sharing work during the active phases of the project. Particularly interesting would be comparing oceanography with, for instance, the NSF Division of Social and Economic Sciences (SES) or the NSF Division of Civil, Mechanical and Manufacturing Innovation (CMMI) as managing and sharing data will likely foreground privacy in the former case, and intellectual property in the latter, both areas that were not salient in my data.

Comparative studies will help identify which findings from this dissertation are unique to oceanography, i.e. field-specific, and which are transferable across contexts. Doing so would help to refine the concepts developed through this study e.g. selective planned hindsight, temporal paradoxes of infrastructures.

Another promising direction would be longitudinal studies of researcher practices and infrastructure evolution, or devolution will help better understand how practice-time profiles and collaborative rhythms between researchers and infrastructures evolve over time. For instance, in a similar vein to Stahlman (2022) how do practice-time profiles mature over the lifetime of a PI's career? For cyberinfrastructures, longitudinal studies can examine the durability or ephemerality of cyberinfrastructures. Longitudinal studies do not necessarily have to follow the same research group over the span of multiple years to possibly decades, archival research and historical cases on early examples of cyberinfrastructure could be particularly informative. This would provide additional insight into how data availability over short and long time periods are not just made, but also remade or undone.

Lastly, I encourage future researchers to pay attention to endings and obsolescence. Whilst these are not new topics per se, there is a distinct "productionist" or "inceptionist" lens in current information science and infrastructure studies. There exists an implicit assumption in

scholarship on infrastructural development that durability is better. The emphasis is on beginnings, production, development. Meanwhile, with some notable exceptions (Jackson & Buyuktur, 2014), analytically cases of endings, dismantling, stagnation, and decay are neglected. In other words, focusing analytically and empirically on those cases when data and infrastructures cease to exist.

Endings also sit on a spectrum, for example, in the DMP with biological samples with questionable integrity after 5 years, these data still exist physically but they have lost their epistemic value. Put another way, those biological samples may have ceased to be data in a meaningful way. As other DMP segments as well as Van Allen's (2023) study reveal, what parts of a sample becomes data versus biowaste is not a clearcut answer. As such, future research should also pay attention to when data and possibly infrastructure may still "exist" but have experienced demise of a different mode. For cyberinfrastructures, this would mean looking at historical examples of formal decommissioning, informal neglect, or purposeful deletion. For data, this would mean examining what makes data no longer usable, desirable, in other words evidential?

8.3 From Exhaustive towards Selective Planned Hindsight

If in practice, planned afterlives of some data are not exhaustive as the policy imagines, and can never be, then one possible alternative is to move from exhaustive planned hindsight to selective planned hindsight. In contrast to exhaustive planned hindsight, selective planned hindsight acknowledges that not all data can receive, or perhaps unpopularly, merit equal stewardship. Instead of casting a wide net over every possible data object, this approach foregrounds the following guiding principles: 1) material viability, 2) epistemic prioritization, 3) re-use potential, and 4) considering data reduction.

For material viability, as with examples of physical samples that decay over time due to naturally occurring processes, these may be candidates for partial archival. Extracts, imagines, or digital data obtained from samples may be shared and archived rather than the entire sample collection. For different reasons, large model outputs may not be materially viable due to their scale which makes it difficult to share over the public internet. Storing model outputs on hard disks is of course an option, but researchers and their teams need to plan for physical storage solutions in the form of hard drives and redundant disk arrays. These too have a limit.

Epistemic prioritization reflects disciplinary and community norms. For instance, model outputs are not seen to be valuable for several reasons. They are seen as relatively expendable because the output can be re-run by the model and is therefore neither rare nor unique. Furthermore, oceanographers who work with ocean models consider the model as the more valuable artifacts over the model outputs. Currently these are shared across a series of venues and platforms ranging from informal to repositories. Therefore, not all model derivatives should be shareable and preservable. Potentially an academic NSF version of GitHub is needed.

Re-use potential shifts the mode of preservation and archival from imagining every possible future reuse to a mode that is more targeted and intentional by considering the most likely re-users and reuse for that data.

Considering data reduction. Some artifacts are allowed to lapse once community demand, technical compatibility, or research relevance declines below a threshold. The aim of selective planned hindsight is to recognize the heterochronous nature of data and their afterlives and the finitude of resources. The impact of the principles is aimed at reimagining planning data and their afterlives towards a mode of discernment rather than the totalizing mode of current policy.

8.4 Policy implications

This dissertation offers several policy implications that challenge the totalizing, i.e. "one-size-fits-all" approach to research data. In particular, it calls for a shift from where researchers are expected to preserve all data for all possible reuses to an effort that is aligned with epistemically and materially viable data whilst considering the constraints of labor, resources and infrastructure.

In other words, from exhaustive planned hindsight to selective planned hindsight. Operationalizing selectivity will require adjustments to DMP guidance, what content should be addressed in the DMP, how DMPs are evaluated, as well as better aligning financial resources that recognizes the labor required to produce, circulate, and preserve research data.

With regards to data management and sharing guidance and DMP content, at present DMP authors typically list anticipated data and describe how each item will be stored, shared, preserved and access maintained to them. This may also be a byproduct of using DMP template tools. With selectivity in mind, one alternative might ask researchers to triage by assigning each anticipated dataset or research output into different categories, what these categories are can be addressed by future research. In any case, researchers in DMPs can explain their categorization

using the guiding principles of selective planned hindsight i.e. anticipated user communities, existing data infrastructure support, epistemic value, and preservation feasibility. A shift, such as the one proposed, or an alternative approach would acknowledge the heterochronous lives and afterlives of research data. Importantly, it aims to promote more intentional efforts about what to share and preserve and why.

Just as the guidance and DMP content changes, so must the evaluation and peer-review process for DMPs. Review criteria can be tailored to reward intentionality, clarity, and contextual fit. For instance, whether the proposed sharing and preservation strategies match the scale, type, and format of data.

The last policy implication I would like to discuss relates to aligning funding with the labor required to produce shareable and publicly available data. As discussed in Chapter 6, some PIs in particular those who have received NSF funding before and after the DMP mandate, such as P14 below, see the work mandated by the data policy as separate from the "core science" that they are being funded to do. P14 explained:

> "*It's been hard to make time for it, and not to begrudge the time that is spent on it. You just always feel like it's getting shoehorned into it. And I put two months of my salary on [this NSF project] per year, say, or maybe one month. And that's the same as I used to do when I didn't do really careful data archiving. And now I'm doing the whole science project, and I'm trying to do better and more careful open science data archiving. And yet the amount of time I'm allocated is the same*." *(P14)*

As the quote above from P14 illustrates, this PI essentially feels short-changed. This is because they are being asked to do more work "I'm doing the whole science project, and …careful open science data archiving" and yet the amount of time and money remains the same as before. One option may be to implement funding supplements to properly compensate data management for sharing for datasets that have broad, cross-disciplinary applicability, are epistemically valuable or other criteria(s) as determined by future research. An additional benefit of a grant is that researchers can list it in their CVs and receive professional and social credit for their work.

8.5 Conclusion

By attending to the temporal dimensions of policy, data, practice, and infrastructures, this dissertation has contributed to a better understanding to why data availability remains uneven,

fragile, and contingent. Rather than treating time as an abstract constraint or background to everyday scientific work, this study foregrounds and centers temporality at the center of oceanographic research.

Bibliography

Adam, B. (1998). *Timescapes of modernity: The environment & invisible hazards*. Routledge.

Adam, B. (2008). *OF TIMESCAPES, FUTURESCAPES AND TIMEPRINTS*.

Adler, A. (2019). *Neptune's Laboratory: Fantasy, Fear, and Science at Sea*. Harvard University Press.

Aldiabat, K., & Le Navenec, C.-L. (2014). Philosophical Roots of Classical Grounded Theory: Its Foundations in Symbolic Interactionism. *The Qualitative Report*. https://doi.org/10.46743/2160-3715/2011.1121

Arzberger, P., Schroeder, P., Beaulieu, A., Bowker, G., Casey, K., Laaksonen, L., Moorman, D., Uhlir, P., & Wouters, P. (2004). Promoting Access to Public Research Data for Scientific, Economic, and Social Development. *Data Science Journal*, *3*, 135–152. https://doi.org/10.2481/dsj.3.135

Baker, K. S., & Chandler, C. L. (2008). Enabling long-term oceanographic research: Changing data practices, information management strategies and informatics. *Deep Sea Research Part II: Topical Studies in Oceanography*, *55*(18–19), 2132–2142. https://doi.org/10.1016/j.dsr2.2008.05.009

BCO-DMO. (2020). *Abount BCO-DMO*. https://www.bco-dmo.org/about

Bennett, A., Sutherland, W., Tian, Y., Finn, M., & Acker, A. (2021). Pathways to Data: From Plans to Datasets. *2021 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, 254–257. https://doi.org/10.1109/JCDL52503.2021.00077

Bindoff, N., & Legler, D. M. (2002, November). *WOCE Global Data V3*. WOCE and Beyond, San Antonio, TX. https://www.ncei.noaa.gov/data/oceans/nodc/woce_conf2002/bindoff.ppt

Bishop, B., Oliver, E. C. J., & Aporta, C. (2022). Co-producing maps as boundary objects: Bridging Labrador Inuit knowledge and oceanographic research. *Journal of Cultural Geography*, *39*(1), 55–89. https://doi.org/10.1080/08873631.2021.1998992

Borgman, C. L., Golshan, M. S., Sands, A. E., Wallis, J. C., Cummings, R. L., Darch, P. T., & Randles, B. M. (2016). Data Management in the Long Tail: Science, Software, and Service. *International Journal of Digital Curation*, *11*(1), 128–149. https://doi.org/10.2218/ijdc.v11i1.428

Bowen, G. A. (2009). Document Analysis as a Qualitative Research Method. *Qualitative Research Journal*, *9*(2), 27–40. https://doi.org/10.3316/QRJ0902027

Braun, V., Lafuente-Funes, S., Lemke, T., & Liburkina, R. (2023). Making Futures by Freezing Life: Ambivalent Temporalities of Cryopreservation Practices. *Science, Technology, & Human Values*, *48*(4), 693–699. https://doi.org/10.1177/01622439231170557

Burgess, A. (2010). Doing time: An exploration of timescapes in literacy learning and research. *Language and Education*, *24*(5), 353–365. https://doi.org/10.1080/09500781003633170

Campbell, R., Javorka, M., Engleton, J., Fishwick, K., Gregory, K., & Goodman-Williams, R. (2023). Open-Science Guidance for Qualitative Research: An Empirically Validated Approach for De-Identifying Sensitive Narrative Data. *Advances in Methods and Practices in Psychological Science*, *6*(4), 25152459231205832. https://doi.org/10.1177/25152459231205832

Carlson, J. (2017). *An Analysis of Data Management Plans from the University of Michigan*. http://deepblue.lib.umich.edu/handle/2027.42/136230

Charmaz, K. (2014). *Constructing grounded theory* (2nd edition). Sage.

Chignard, S. (2013, March 29). *A brief history of Open Data*. Paris Tech Review. https://www.paristechreview.com/2013/03/29/brief-history-open-data/

CHIPS and Science Act, Pub. L. No. 117–167 (2022). https://www.congress.gov/117/plaws/publ167/PLAW-117publ167.pdf

Cohn, M. L. (2016). Convivial Decay: Entangled Lifetimes in a Geriatric Infrastructure. *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, 1511–1523. https://doi.org/10.1145/2818048.2820077

Conway, E. M. (2006). Drowning in data: Satellite oceanography and information overload in the Earth sciences. *Historical Studies in the Physical and Biological Sciences*, *37*(1), 127–151. https://doi.org/10.1525/hsps.2006.37.1.127

Currie, A. (2021). Stepping Forwards by Looking Back: Underdetermination, Epistemic Scarcity and Legacy Data. *Perspectives on Science*, *29*(1), 104–132. https://doi.org/10.1162/posc_a_00362

Darch, P. T., & Borgman, C. L. (2014). Ship space to database: Motivations to manage research data for the deep subseafloor biosphere: Ship Space to Database: Motivations to Manage Research Data for the Deep Subseafloor Biosphere. *Proceedings of the American Society for Information Science and Technology*, *51*(1), 1–10. https://doi.org/10.1002/meet.2014.14505101056

Data.gov admins. (2013, April 4). *Open Data: A History*. Data.Gov. https://data.gov/blog/open-data-history/

Dietrich, D., Adamus, T., Miner, A., & Steinhart, G. (2012). De-Mystifying the Data Management Requirements of Research Funders. *Issues in Science and Technology Librarianship*, *70*. https://doi.org/10.29173/istl1556

Dumit, J. (2004). *Picturing personhood: Brain scans and biomedical identity*. Princeton University Press.

Economist Impact. (2023, March 28). *From data sets to data flows: Making the case for open-source ocean science*. https://impact.economist.com/ocean/ocean-health/from-data-sets-to-data-flows-making-the-case-for-open-source-ocean-science

Edwards, P. N. (2003). Infrastructure and Modernity: Force, Time, and Social Organization in the History of Sociotechnical Systems. In T. J. Misa, P. Brey, & A. Feenberg (Eds.), *Modernity and Technology* (pp. 185–225). MIT Press.

Edwards, P. N. (2010). *A vast machine: Computer models, climate data, and the politics of global warming*. MIT Press.

Edwards, P. N., Jackson, S. J., Bowker, G. C., & Knobel, C. (2007). *Understanding infrastructure: Dynamics, tensions, and design*.

Evans, D. L., Alpers, W., Cazenave, A., Elachi, C., Farr, T., Glackin, D., Holt, B., Jones, L., Liu, W. T., McCandless, W., Menard, Y., Moore, R., & Njoku, E. (2005). Seasat—A 25-year legacy of success. *Remote Sensing of Environment*, *94*(3), 384–404. https://doi.org/10.1016/j.rse.2004.09.011

Garforth, L. (2012). In/Visibilities of Research: Seeing and Knowing in STS. *Science, Technology, & Human Values*, *37*(2), 264–285. https://doi.org/10.1177/0162243911409248

Glaser, B., & Strauss, A. (1967). *Discovery of grounded theory: Strategies for qualitative research*. Aldine Transactions.

Gonzales, L. (2020, January 14). NSF Geosciences Directorate Funding by Institution Type. *American Geosciences Institute Geoscience Currents*.

Greyson, D. (2016). Evolution of information practices over time. *Proceedings of the Association for Information Science and Technology*, *53*(1), 1–8. https://doi.org/10.1002/pra2.2016.14505301052

Haider, J., Johansson, V., & Hammarfelt, B. (2022). Time and temporality in library and information science. *Journal of Documentation*, *78*(1), 1–17. https://doi.org/10.1108/JD-09-2021-0171

Halfmann, G. (2018). *Seafarers, silk, and science: Oceanographic data in the making* [Doctoral dissertation]. University of Exeter.

Hess, C., & Ostrom, E. (2007). *Understanding knowledge as a commons: From theory to practice*. MIT Press.

Hey, T., Tansley, S., & Tolle, K. (Eds.). (2009). *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research.

Hudson-Vitale, C., & Moulaison-Sandy, H. (2019). Data Management: Plans A Review. *DESIDOC Journal of Library & Information Technology*, *39*(06), 322–328. https://doi.org/10.14429/djlit.39.06.15086

Hughes, T. (1993). *Networks of power: Electrification in Western society, 1880-1930*. JHU press.

Interagency Working Group on Digital Data. (2009). *Harnessing the power of digital data for science and society* [Report of the Interagency Working Group on Digital Data to the Committee on Science of the National Science and Technology Council].

Jackson, S. J., & Buyuktur, A. (2014). Who Killed WATERS? Mess, Method, and Forensic Explanation in the Making and Unmaking of Large-scale Science Networks. *Science, Technology, & Human Values*, *39*(2), 285–308. https://doi.org/10.1177/0162243913516013

Jackson, S. J., Ribes, D., Buyuktur, A., & Bowker, G. C. (2011). Collaborative rhythm: Temporal dissonance and alignment in collaborative scientific work. *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work*, 245–254. https://doi.org/10.1145/1958824.1958861

Joyce, T., Corry, C., & WHP Office. (1994). *WOCE operations manual, volume 3: The observational programme, section 3.1: WOCE hydrographic programme, part 3.1.2: Requirements for WOCE hydrographic programme data reporting* (WHP Office report; WHPO 09-1; WOCE report; no. 67/91).

Karasti, H., Baker, K. S., & Millerand, F. (2010). Infrastructure Time: Long-term Matters in Collaborative Development. *Computer Supported Cooperative Work (CSCW)*, *19*(3–4), 377–415. https://doi.org/10.1007/s10606-010-9113-z

Kern, S. (2000). Time and Medicine. *Annals of Internal Medicine*, *132*(1).

Kintisch, E. (2015). A moment of truth arrives for U.S. ocean science. *Science*, *347*(6221), 463–463. https://doi.org/10.1126/science.347.6221.463

Kitchin, R. (2014a). Big Data, new epistemologies and paradigm shifts. *Big Data & Society*, *1*(1), 205395171452848. https://doi.org/10.1177/2053951714528481

Kitchin, R. (2014b). *The Data Revolution*. SAGE Publications.

Kitchin, R. (2023). *Digital timescapes: Technology, temporality and society*. John Wiley & Sons.

Lehman, J. (2018). From ships to robots: The social relations of sensing the world ocean. *Social Studies of Science*, *48*(1), 57–79. https://doi.org/10.1177/0306312717743579

Lehman, J. (2021). Sea Change: The World Ocean Circulation Experiment and the Productive Limits of Ocean Variability. *Science, Technology, & Human Values*, *46*(4), 839–862. https://doi.org/10.1177/0162243920949932

Leonelli, S. (2018). The Time of Data: Timescales of Data Use in the Life Sciences. *Philosophy of Science*, *85*(5), 741–754. https://doi.org/10.1086/699699
Leonelli, S., Rappert, B., & Davies, G. (2017). Data Shadows: Knowledge, Openness, and Absence. *Science, Technology, & Human Values*, *42*(2), 191–202. https://doi.org/10.1177/0162243916687039
Levin, N., Leonelli, S., Weckowska, D., Castle, D., & Dupré, J. (2016). How Do Scientists Define Openness? Exploring the Relationship Between Open Science Policies and Research Practice. *Bulletin of Science, Technology & Society*, *36*(2), 128–141. https://doi.org/10.1177/0270467616668760
Lindstrom, E. J., & Legier, D. M. (2001). Chapter 3.5 Developing the WOCE Global Data System. In *International Geophysics* (Vol. 77, pp. 181–190). Academic Press.
Marcial, L. H., & Hemminger, B. M. (2010). Scientific data repositories on the Web: An initial survey. *Journal of the American Society for Information Science and Technology*, *61*(10), 2029–2048. https://doi.org/10.1002/asi.21339
May, J., & Thrift, N. (Eds.). (2003). *Timespace: Geographies of temporality*. Routledge.
M'charek, A. (2014). Race, Time and Folded Objects: The HeLa Error. *Theory, Culture & Society*, *31*(6), 29–56. https://doi.org/10.1177/0263276413501704
Memorandum on Transparency and Open Government. (2009, January 21). Whitehouse.Gov. https://obamawhitehouse.archives.gov/the-press-office/transparency-and-open-government
Miksa, T., Simms, S., Mietchen, D., & Jones, S. (2019). Ten principles for machine-actionable data management plans. *PLOS Computational Biology*, *15*(3), e1006750. https://doi.org/10.1371/journal.pcbi.1006750
Mirowski, P. (2018). The future(s) of open science. *Social Studies of Science*, *48*(2), 171–203. https://doi.org/10.1177/0306312718772086
Mosconi, G., Li, Q., Randall, D., Karasti, H., Tolmie, P., Barutzky, J., Korn, M., & Pipek, V. (2019). Three Gaps in Opening Science. *Computer Supported Cooperative Work (CSCW)*, *28*(3–4), 749–789. https://doi.org/10.1007/s10606-019-09354-z
National Centers for Environmental Information. (2002, December 20). *NODC Standard Product: World Ocean Circulation Experiment (WOCE) Global Data Resource (GDR), versions 1-3, on CD-ROM and DVD*. https://www.ncei.noaa.gov/access/metadata/landing-page/bin/iso?id=gov.noaa.nodc:NODC-WOCE-GDR
National Oceanic and Atmospheric Administration. (2021, February 26). *How much of the ocean have we explored?* How Much of the Ocean Have We Explored? https://oceanservice.noaa.gov/facts/exploration.html
National Oceanic and Atmospheric Administration. (2023, January 20). *What does the ocean have to do with human health?* National Ocean Service National Oceanic and Atmospheric Administration. https://oceanservice.noaa.gov/facts/ocean-human-health.html
National Research Council. (1982). *Data Management and Computation: Volume 1: Issues and Recommendations* (p. 19537). National Academies Press. https://doi.org/10.17226/19537
National Research Council. (2000). *50 Years of Ocean Discovery: National Science Foundation 1950-2000*.
National Research Council. (2011). *Critical Infrastructure for Ocean Research and Societal Needs in 2030*.

National Science Board. (2005). *Long-Lived Digital Data Collections Enabling Research and Education in the 21st Century* (NSB-05-40; pp. 1–87). National Science Foundation. https://www.nsf.gov/pubs/2005/nsb0540/

National Science Foundation. (2010, May 10). *Scientists seeking NSF funding will soon be required to submit data management plans [Press release]*. https://www.nsf.gov/news/news_summ.jsp?cntn_id=116928

National Science Foundation. (2011, January). *Grant Proposal Guide: Chapter II.C.2.j*. https://www.nsf.gov/pubs/policydocs/pappguide/nsf13001/gpg_2.jsp

National Science Foundation. (2015). *Today's Data, Tomorrow's Discoveries: Increasing Access to the Results of Research Funded by the National Science Foundation*.

Neang, A. B., Sutherland, W., Beach, M. W., & Lee, C. P. (2021). Data Integration as Coordination: The Articulation of Data Work in an Ocean Science Collaboration. *Proceedings of the ACM on Human-Computer Interaction*, *4*, 1–25. https://doi.org/10.1145/3432955

NOAA AOML. (2024, June). *Subsurface Float Data*. Subsurface Float Data. https://www.aoml.noaa.gov/phod/float_traj/data/index.php

NOAA NCEI. (2020, June 14). *Archive: Submitting Your Data*. National Centers for Environmental Information (NCEI). https://www.ncei.noaa.gov/archive

Nowlin, W. (1985). WOCE/TOGA Data Management. *Eos, Transactions American Geophysical Union*, *66*(19), 436–436. https://doi.org/10.1029/EO066i019p00436

Ocean Literacy Network. (2013). *Ocean Literacy: The essential principles of ocean sciences for learners of all ages*. National Oceanic and Atmospheric Administration. http://oceanliteracy.wp2.coexploration.org/?page_id=164

Ocean Observatories Initiative Facility Board. (2021). *Ocean Observatories Initiative (OOI) Science Plan: Exciting Opportunities Using OOI Data*. Ocean Observatories Initiative Facility Board. https://doi.org/10.23860/ooi-science-plan-2021-01

OECD. (2007). *OECD Principles and Guidelines for Access to Research Data from Public Funding*. OECD. https://doi.org/10.1787/9789264034020-en-fr

Open Knowledge Foundation. (n.d.). *The Open Definition—Open Definition—Defining Open in Open Data, Open Content and Open Knowledge*. Open Definition. https://opendefinition.org/

Oreskes, N. (2021). *Science on a Mission: How Military Funding Shaped What We Do and Don't Know about the Ocean*. University of Chicago Press.

OSTP. (2022). *Ensuring Free, Immediate, and Equitable Access to Federally Funded Research*. White House Office of Science and Technology Policy (OSTP), Executive Office of the President of the United States. https://doi.org/10.5479/10088/113528

Parham, S. W., Carlson, J., Hswe, P., Westra, B., & Whitmire, A. (2016). Using Data Management Plans to Explore Variability in Research Data Management Practices Across Domains. *International Journal of Digital Curation*, *11*(1), 53–67. https://doi.org/10.2218/ijdc.v11i1.423

Pasek, J. E. (2017). Historical Development and Key Issues of Data Management Plan Requirements for National Science Foundation Grants: A Review. *Issues in Science and Technology Librarianship*. https://doi.org/10.5062/F4QC01RP

Pendleton, L., & Sorensen. (2021, April 14). *The hidden downside to ocean data and how to make it more sustainable*. World Economic Forum.

https://www.weforum.org/agenda/2021/04/10-ways-to-make-ocean-data-more-sustainable/
Pinfield, S., Salter, J., Bath, P. A., Hubbard, B., Millington, P., Anders, J. H., & Hussain, A. (2014). Open-access repositories worldwide, 2005–2012: Past growth, current characteristics, and future possibilities. *Journal of the Association for Information Science and Technology*, *65*(12), 2404–2421.
Priego, L. P., & Wareham, J. (2020). *The stickiness of scientific data*.
Qian, C., Huang, B., Yang, X., & Chen, G. (2022). Data science for oceanography: From small data to big data. *Big Earth Data*, *6*(2), 236–250. https://doi.org/10.1080/20964471.2021.1902080
Radin, J. (2015). Planned Hindsight: The vital valuations of frozen tissue at the zoo and the natural history museum. *Journal of Cultural Economy*, *8*(3), 361–378. https://doi.org/10.1080/17530350.2015.1039458
Rainger, R. (2000). Science at the Crossroads: The Navy, Bikini Atoll, and American Oceanography in the 1940s. *Historical Studies in the Physical and Biological Sciences*, *30*(2), 349–371. https://doi.org/10.2307/27757835
Ralph, N., Birks, M., & Chapman, Y. (2014). Contextual Positioning: Using Documents as Extant Data in Grounded Theory Research. *SAGE Open*, *4*(3), 215824401455242. https://doi.org/10.1177/2158244014552425
Reddy, M. C., Dourish, P., & Pratt, W. (2006). Temporality in Medical Work: Time also Matters. *Computer Supported Cooperative Work (CSCW)*, *15*(1), 29–53. https://doi.org/10.1007/s10606-005-9010-z
Ribes, D., & Finholt, T. (2009). The Long Now of Technology Infrastructure: Articulating Tensions in Development. *Journal of the Association for Information Systems*, *10*(5), 375–398. https://doi.org/10.17705/1jais.00199
Saldaña, J. (2013). *The coding manual for qualitative researchers* (2nd ed). SAGE.
Serres, M., & Latour, B. (1990). *Conversations on science, culture, and time*. University of Michigan Press. https://doi.org/10.3998/mpub.12087
Sharma, S., Wilson, J., Tian, Y., Finn, M., & Acker, A. (2023). The New Information Retrieval Problem: Data Availability. *Proceedings of the Association for Information Science and Technology*, *60*(1), 379–387. https://doi.org/10.1002/pra2.796
Shove, E., Trentmann, F., & Wilk, R. (Eds.). (2009). *Time, consumption and everyday life: Practice, materiality and culture*. Berg Publishers.
Smale, N. A., Unsworth, K., Denyer, G., Magatova, E., & Barr, D. (2020). A Review of the History, Advocacy and Efficacy of Data Management Plans. *International Journal of Digital Curation*, *15*(1), 30. https://doi.org/10.2218/ijdc.v15i1.525
Stahlman, G. R. (2022). From nostalgia to knowledge: Considering the personal dimensions of data lifecycles. *Journal of the Association for Information Science and Technology*, *73*(12), 1692–1705. https://doi.org/10.1002/asi.24687
Steinhardt, S. B., & Jackson, S. J. (2014a). Material Engagements: Putting Plans and Things Together in Collaborative Ocean Science. *2014 47th Hawaii International Conference on System Sciences*, 1505–1514. https://doi.org/10.1109/HICSS.2014.194
Steinhardt, S. B., & Jackson, S. J. (2014b). Reconciling rhythms: Plans and temporal alignment in collaborative scientific work. *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, 134–145. https://doi.org/10.1145/2531602.2531736

Steinhardt, S. B., & Jackson, S. J. (2015). Anticipation Work: Cultivating Vision in Collective Practice. *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, 443–453. https://doi.org/10.1145/2675133.2675298

Suchman, C. (2022). *Directorate for Geosciences (GEO)*.

Tedersoo, L., Küngas, R., Oras, E., Köster, K., Eenmaa, H., Leijen, Ä., Pedaste, M., Raju, M., Astapova, A., Lukner, H., Kogermann, K., & Sepp, T. (2021). Data sharing practices and data availability upon request differ across scientific disciplines. *Scientific Data*, *8*(1), 192. https://doi.org/10.1038/s41597-021-00981-0

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., Manoff, M., & Frame, M. (2011). Data Sharing by Scientists: Practices and Perceptions. *PLOS ONE*, *6*(6), e21101. https://doi.org/10.1371/journal.pone.0021101

Tenopir, C., Dalton, E. D., Allard, S., Frame, M., Pjesivac, I., Birch, B., Pollock, D., & Dorsett, K. (2015). Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide. *PLOS ONE*, *10*(8), e0134826. https://doi.org/10.1371/journal.pone.0134826

Thompson, B. J., Crease, J., & Gould, J. (2001). Chapter 1.3 The origins, development and conduct of WOCE. In G. Siedler, J. Church, & J. Gould (Eds.), *International Geophysics* (Vol. 77, pp. 31–VIII). Academic Press. https://doi.org/10.1016/S0074-6142(01)80110-8

Tian, Y., Bennett, A., Sutherland, W., Ferguson, A., Ford, M., Li, J. N.-K., Yarbrough, E., Finn, M., & Acker, A. (2021). *An Analysis of NSF Data Management Plan Guidelines*.

Tracy, S. J. (2013). *Qualitative Research Methods: Collecting evidence, crafting analysis, communicating impact*. Wiley-Blackwell.

Traweek, S. (2009). *Beametimes and Lifetimes: The World of High Energy Physics*. Harvard University Press.

Treloar, A., & Harboe-Ree, C. (2008). *Data management and the curation continuum: How the Monash experience is informing repository relationships*.

U.S. Planning Office for WOCE. (1990). *U.S. WOCE implementation plan 1990* (No. 2). U.S. WOCE Office, Dept. of Oceanography, Texas A & M University.

U.S. Planning Office for WOCE, U. S. W. S. S. C. (1988). *U.S. WOCE implementation plan, first draft*. U.S. Planning Office for WOCE, Dept. of Oceanography, Texas A & M University.

Van Allen, A. (2023). Entangled Timelines. Crafting Types of Time Through Making Museum Specimens. *Centaurus*, *65*(2), 291–312. https://doi.org/10.1484/J.CNT.5.135353

Van Keuren, D. (2000). Building a New Foundation for the Ocean Sciences: The National Science Foundation and Oceanography, 1951-1965. *Earth Sciences History*, *19*(1), 90–109. https://doi.org/10.17704/eshi.19.1.c531h01m58j324q6

Webster, F. (1985). WOCE/TOGA Data Management Working Group Meeting Report. *Bulletin of the American Meteorological Society*, *66*(7), 853–854. https://doi.org/10.1175/1520-0477-66.7.853

Whitmire, A., Boock, M., & Sutton, S. C. (2015). Variability in academic research data management practices: Implications for data services development from a faculty survey. *Program: Electronic Library and Information Systems*, *49*(4), 382–407. https://doi.org/10.1108/PROG-02-2015-0017

Whitmire, A., Carlson, J., Westra, B., Hswe, P., & Parham, S. W. (2015). *The DART Project: Using data management plans as a research tool*. https://doi.org/10.17605/OSF.IO/KH2Y6

WHOI. (2025). HOV Alvin—Woods Hole Oceanographic Institution. *Https://Www.Whoi.Edu/*. https://www.whoi.edu/what-we-do/explore/underwater-vehicles/hov-alvin/

WHP Office. (1991). *WOCE operations manual, volume 3: The observational programme, section 3.1: WOCE hydrographic programme, part 3.1.2: Requirements for WHP data reporting* (WHP office report WHPO; 90-1; WOCE report no.; 67/91). WOCE Hydrographic Programme Office.

Willinsky, J. (2005). The unacknowledged convergence of open source, open access, and open science. *First Monday*. https://doi.org/10.5210/fm.v10i8.1265

Witze, A. (2015). US ocean sciences told to plot fresh course. *Nature*, *517*(7536), 538–539.

WMO, I. (1986). *Scientific Plan for the World Ocean Circulation Experiment* (WMO/TD-No. 122).

WOCE International Project Office. (1992). *World Ocean Circulation Experiment, Scientific Steering Group: Report of the eighteenth meeting: WOCE-18: Texas A & M University (TAMU), Moody Campus, Galveston, Texas, USA, 12-14 May 1992.* (WOCE report; no.94/12). WOCE International Project Office.

WOCE International Project Office. (1993). *WOCE Data Management: Data Sharing Policy and Practices: Data Assembly and Analysis Centres: Satellite Data Availability and Data Information Unit* (WOCE Report No. 104/93). WOCE International Project Office.

WOCE International Project Office. (1997). *World Ocean Circulation Experiment International Project Office, August 1997, Ocean circulation and Climate World Ocean Circulation Experiment, WOCE Report No. 154/97* (WOCE Report No. 154/97). WOCE International Project Office.

WOCE International Project Office. (2002). *WOCE observations, 1990-1998: A summary of the WOCE global data resource* (WOCE report; no. 179/02). WOCE International Project Office.

Woods Hole Oceanographic Institution. (n.d.). History of Oceanography. *Dive & Discover*. Retrieved May 16, 2023, from https://divediscover.whoi.edu/history-of-oceanography/

Woods, J. D. (1985). The World Ocean Circulation Experiment. *Nature*, *314*(6011), 501–511. https://doi.org/10.1038/314501a0

Wylie, C. D. (2019). The plurality of assumptions about fossils and time. *History and Philosophy of the Life Sciences*, *41*(2), 21. https://doi.org/10.1007/s40656-019-0260-3

Wylie, C. D. (2024). Timing Science: The Temporal Role of Scientists in the Construction of Data. *Philosophy, Theory, and Practice in Biology*, *16*(2). https://doi.org/10.3998/ptpbio.5646

Appendix A. Interview Protocol For 2022 Interviews

Introductory Protocol

Thank you again for agreeing to participate in an interview for this project. Through this interview process, we hope to learn more from you about the role that data management plans play in shaping data sharing, re-use, archiving and access issues related to your project and more broadly in oceanography. In a moment I will start the Zoom recording and ask for your verbal consent. Whenever you are ready I will start the recording.

[Begin recording]

Have you had a chance to read the consent form?
Do you have any questions for me?
Do you consent to participate in this study?

A. Data and scientific work (10 mins)

1. Could you tell me about your research?
2. What types/forms/formats of data do you work with in your research?
    a. *Follow-up*: Are these "typical" data for your discipline?

B. Project-specific questions: (10 mins)

1. What was the funding process like for this grant?
2. Could you walk me through the data collection process for your project?
    a. *Follow-up:* Where did your data come from?
    b. *Follow-up:* What format was it in?
3. Could you tell me about who worked on data-related work on your project? By data-related work I mean data collection, cleaning, analysis, archiving.
4. Are there other documents that you generated in the lab to teach students/collaborators about data management or data cleaning?

C. Data Management Plan evaluation: (10 mins)

1. Has the way you write DMPs changed over the course of different NSF-funded projects? If so, how?
2. Could you tell me about how data-related work in your projects has changed after the 2011 NSF DMP policy mandate?

D. Data Management Plan: (15 mins)

1. I've noticed that a lot of research funded under OCE deposit their project data in BCO-DMO, could you tell me about your experience(s) working with BCO-DMO to archive your research data?
    a. Are there other repositories that you worked with to deposit and archive your research data?
2. What role did the DMP play in your everyday data management practices?
3. After submitting your project proposal, how often, if at all, did you refer to and/or edit your DMP?

4. *[If not answered above]* Could you walk me through the process of how you wrote this DMP?
    a. *Follow-up:* Did you use any resources, e.g. tools or templates, to help you draft this DMP?
5. What else should I know about DMPs?

E. Wrap up: (5 mins)

1. How have data collection and dissemination methods changed since you entered your discipline?
2. Are there particular challenges your discipline faces when working with data?
3. How is open access to data being discussed in your discipline?
4. Who on your NSF project team could I talk to to learn more about your project?

Thank you for sharing your expertise and experience with me today.

Appendix B. Interview Workflow Checklist

Interview Implementation Details

- The recommended interview length is approximately 1 hour.
- The intended participants for this interview are researchers in oceanography who have experience working on a NSF funded research project. Specific job titles, roles, responsibilities may vary depending on the interviewee's discipline, stage of their career, and nature of the funded project.
- The interview can be conducted in-person or using a video conference platform. If both are feasible, the recommendation is to ask the participants for their preference.
- The recommended data recording strategy includes audio-recording the interviews and transcribing the interviews verbatim.

Interview Preparation Workflow Checklist

- Review the DMP submitted by each interviewee and their reply to our email campaign, if on file
- Review award metadata for the research project in question
- Locate and review interviewee's academic and professional background. This can be done through visiting their profile on their institution's website, personal website, google scholar page, and/or skimming their CV.
- Locate and review, if applicable, the research project website for which we have the DMP

Interview Day Materials Checklist

- A copy of the interview guide
- A copy of the informed consent form
- A laptop to conduct the interview (if applicable)
- An audio recording device and batteries (if applicable)
- A tool to take notes during the interview (e.g., a physical notebook, Google Doc, Word)

Appendix C. Interview Protocol for 2025 Interviews with PIs, graduate students, and research scientists

Introductory Protocol
Thank you again for agreeing to participate in an interview for my project. Through this interview process, I hope to learn more from you about the ending stages of your most recent NSF-funded project, how you shared research artifacts, what challenges you encountered, if any, and what made this stage of the project easier for you. I will start the recording and ask you for your verbal consent if you're ready.

[Begin recording]

Have you had a chance to read the consent form?
Do you have any questions for me?
Do you consent to participate in this study?

A. Researcher background

1. What are the primary research questions that you, or your lab, investigates?
2. Please tell me about your NSF project that just ended

B. Project-close out

1. How does your lab approach data sharing, including decisions about time, personnel, and processes? (*some teams have one expert, some it is something everyone does; some teams do work weekly others do bootcamps*)
2. What parts of the research did you share? (e.g. code, software, samples)?
    a. *Follow up:* Where did you share it? OR Who did you share it with?
        i. Can you describe your experience depositing your data in [location name]?
        ii. Did you work with anyone to help deposit your data?
3. What format(s) is the data in?
    a. *Follow up:* Were the data collected in this format?

C. Lessons learned

1. What did you do to prepare your data for sharing?
2. Approximately, how much time did you spend preparing your data for sharing?
    a. *Follow up:* How did your actual time spent on data sharing compare to your initial expectations?
3. What challenges did you encounter during data sharing, if any?
4. Looking back, what factors made data sharing easier, if any?
5. What would you do differently next time (to make data sharing easier)?
6. In your experience sharing data, would you consider the process you described typical?

D. Reflection and wrap up

1. Based on your experience are there any changes you would make to your data management approach for your next project?

2. If a colleague were to come to you for advice, what would you tell them about data sharing?
3. What skills or practices do you think make researchers good at data sharing?
4. Have other researchers contacted you about reusing your data for their research?
    a. Have you re-used others' data in your own research?
5. Is there anything else I should know about the ending stages of an NSF project?
6. Who would you recommend I speak with about data sharing in oceanography?
    a. Would you be comfortable introducing me or may I mention that you recommend them?

Thank you for your time and for sharing your experience with me today!

Appendix D. Interview Protocol for 2025 Interviews with data managers

Introductory Protocol
Thank you again for agreeing to participate in an interview for my project. Through this interview process, I hope to learn more from you about data management and data sharing in oceanography. In particular, what challenges you encounter when doing this work, if any, and what makes doing data management and sharing easier for you. I will start the recording and ask you for your verbal consent if you're ready.

[Begin recording]

Have you had a chance to read the consent form?
Do you have any questions for me?
Do you consent to participate in this study?

A. Interviewee background

1. Could you tell me about yourself, and specifically your education and professional background?
2. If applicable, did you receive training or certifications for research data management?

B. Information about work

1. Could you describe what a typical work week looks like for you?
2. How does your team approach data sharing, including decisions about time, personnel, and processes?
3. What tools or technologies do you use to do data management and data sharing?
4. Which stakeholders do you work with, i.e. data providers, end-users, colleagues?
    a. *Follow up:* At what point in a project do they approach you or your team for help with data management and sharing?
5. Have stakeholders contacted you about reusing [your organization's] data for their research?
6. Do you work with other data repositories and archives in your work?
    a. *Follow up:* Could you describe your experience working with other data repositories/archives?

C. Data management and sharing

1. In your experience, what areas of data management and sharing do researchers, PIs, and data providers find challenging?
2. What aspects of data management and sharing do you and your team find challenging?
3. What aspects of data management and sharing do you and your team find the most time-consuming?
4. What makes data management and sharing easier for you and your team?

D. Reflection and wrap up

1. If a colleague or data provider, were to come to you for advice, what would you tell them about data management and sharing?

2. What skills or practices do you think make researchers good at data sharing?
3. In your time working in the field, what do you think has changed the most about how people discuss data management and sharing?
4. Who would you recommend I speak with about data sharing in oceanography?
    a. Would you be comfortable introducing me or may I mention that you recommend them?

Thank you for your time and for sharing your experience with me today!

Appendix E. Codebook Used For Second Cycle Coding

Code	Sub Code	Definition	Examples
Durational Time		Data-related activities that have an explicit time duration associated with them.	"This will require that we broadly disseminate our data in a timely fashion (i.e. beginning no later than ~3-4 years from project outset) during project analyses and manuscript preparation." (DMP 317)
Durational Time	Time-related to data	Time inherent to the data themselves i.e. time-series data or related to how they were produced i.e. sampling rates.	" One, are time series, we measure waves every half a second for example. And we do this kind of continuously for a month at different locations, so we have a time series of wave height, waves, water going up and down. We also measure the currents that are associated with those waves, so we have time series again at 2 hertz, sometimes 16 hertz, depending on what we're doing." (P1) "Shipboard analyses will include salinity, alkalinity, and sulfate concentrations, which will be reported at 2-5 days resolution." (DMP 348)
Ambiguous/ non-durational time		Data-related activities that have no explicit time duration associated with them. Frequently about data availability.	"Soon after the completion of the model simulations, a subset (see Table 1) of the generated output will be contributed by N. Lovenduski to BCO-DMO." (DMP 168) "Our reprocessed seismic reflection sections will be submitted to the Academic Seismic Portal and made immediately open to all upon

			acceptance of results for publication in a peer-reviewed journal. Our expectation is that this could happen late in Yr2 or during the following year." (DMP 242)
Routines		Data practices that occur regularly over a time period. May be either on a pre-defined time period or have no explicit time related to the routine. This code is only for the practice-level not for data-level routines.	"Fieldnotes are reviewed each evening and entered into an electronic notebook (e.g. Evernote), specimen databases, and workflow software (e.g. Kepler)." (DMP 317)
Delay		Descriptions of reasons or data activities that hinder or postpone the accessibility/availability of data. Frequently about moratoriums and embargoes.	"For data that cannot be immediately published (e.g., intellectual property data: IV), data will be embargoed for up to two years through the Biological and Chemical Oceanography Data Management Office (BCO-DMO; www.bcodmo.org/), or until publication." (DMP 207) "There's, you know, we try to protect our students. If we think that, "God! Maybe we should wait a year for the student to publish before we make this available." But for the most part, we're not too worried about that stuff." (P1)
Time-consuming data work		Aspects of working with data that are laborious, difficult, and/or otherwise described as taking a lot of time. Unlike, time investment code there is no explicit idea of making activities/tasks less time consuming at another point in time.	" It's hard to write notes about every one hour run, every sensor when we have 50 days with 25 sensors, and each sensor has to each channels of data. So we encourage people, if they're gonna use our data in that kind of way that they, we don't need we're not asking people to make us co-authors. We just encourage people to chat with us a little bit,

			tell us what they're doing, and that doesn't happen often enough but we try, [laughter] we try." (P1) "So after the NSF mandate asking people to make their data available, did that impact the way that you work with data in your projects at all? No, not at all. We're gonna do the project. The only thing is now I have to spend time making the data so somebody else can use it. Think about your stuff, no else is ever gonna use whatever model you have going your own it works for you, but it's not yet ready to put on the Internet to market. But other than that, yeah, we don't do anything different whether we're gathering data, whether it's a data management plan or not." (P1)
Time investment		Descriptions of work or actions done to help alleviate time-consuming aspects of data-related work. This code is about past or present work/activites. In addition to work, also includes spending money on tools/technologies.	"Anyway and that's just, so now we have this raw data and it's not in good shape. All kinds of things happen. One, things that are happen no matter what, we're in the surf zone, you mentioned bubbles. And I mentioned bubbles, bubbles reflect acoustics. And if there's too many bubbles, we send out a signal and it just gets stuck in the bubbles and never comes back. So that current rather than being 10 centimeters a second it's gonna be some weird number, maybe 200 centimeters a second or maybe minus 75 we... Anyway, so we have tech, we've written lots of computer codes to try to look at the data and clean it up." (P1)

Policy standards		Specific periods of time that are due to the DMP policy or university policy. Most frequently appears in discussions of data availability and data retention.	“The PIs retain the right to use the data garnered during the project before making them available for wider use. However, in accordance with NSF policies for data sharing, all data will be made available upon publication or no later than two years after the project is completed.” (DMP 264)
Anticipation work		Data-related practices that cultivate and channel expectations of the future (Jordan and Jackson, 2015). This code captures planning activities and planned work. ALSO includes discussions of work that should/could have been done but was, i.e. lack of anticipation work.	“Observational data will be archived at NCEI. Immediately following the grant award we will negotiate an archiving agreement with NCEI with agreements on format, media, method of transmittal of the data and output.” (DMP 107) “Well the same thing with data management, it's gonna change, but the fact that you thought about it and you're aware, that to me is a really good, good idea. I'm supportive of NSF's data management plan thing. And it's not long, it's not hard to write, but you have to think about it and that's good.” (P1) Negative case “ If it's a first time, someone's been coming, if not... We try to train people to be able to do it on their own. So the idea is that we have computer systems where we set up their measurement routines and their workflow, so that in the end, it's quite simple for them to contribute to our data repository. But that takes quite a bit of time, usually it depends on the technical sophistication of the

			person that's going to be archiving the data." (P2)
Change		About change over periods of time. Includes changes at the structural level (policy, institutional, disciplinary) and micro level changes at the PI/lab unit over time e.g. developing data practices over the course of a project or many projects.	"They're making terabytes of images a day. And then from these images, we can do different things. We can try to track the phone that's moving around and get estimates of currents. So we might make a map of what do the currents look like, at least as far as our cameras and our algorithms are concerned, every 10 minutes, here's a 10 minute average, another 10 minute average.We do that. That's kind of a new thing last, just couple last few years that we're learning how to do this. We're learning that you cannot buy enough disc space, it doesn't exist to store all this stuff and back it up. I mean we really are talking like terabytes a day of data that come from the cameras." (P1) "has the way you write data management plans changed over the course of different NSF funded projects? I don't think it really changes other than what are the data involved? So if we're proposing to NSF to do something that involves gathering data at NSF's expense. Kind of have a standard, I'm gonna say a standard, but we say, "Hey, we're gonna make the data available the way we always make the data available", we give... We include some citations for how we do quality control, and people can go look up in the literature, "Oh, this is how they interpolate over bad data or whatever." We often say,

			"Well, make the data available." (P1)
Money		Discussions about the projects' "revenue" i.e. grants, awards, funding. ALSO about project expenses, i.e. costs, salary. ALSO includes discussions about the lack of money i.e. unpaid labor.	"But I think the biggest problem is that there's a mandate to publish the data, and you publish the data in a database that looks fine. And they don't get funding from NSF. And then what happens to your data? That's a big problem, if you're looking at four to five, five-year funding cycles for databases in the US, that's not security, that your data is gonna live long past that funding cycle, and that I remember from that initial open access workshop the NSF people who were there, got sent that message loud and clear, but I haven't really seen any change, and that was so good. 10 years. And 10 years ago, probably. " (P4)
Affect		Strong feelings or emotions expressed by PIs/interviewees.	"Well anyway, what funded this project that I was doing with a UW student was partially funded by those guys, partially funded, I mean they had NSF money, the NSF didn't care, but the Journal cared, this is my biggest pet peeve with data sharing. I'm happy to share data, I do it as much as I possibly can. But the journals AGU being... Has been an actor as any, now says, "You gotta publish your data. You wanna publish an art journal, put your data online." So I had a big fight with Brooks Hansen, he's the publications guy at AGU, I've known him for... Sort of known him for a long time. "Brooks, you want me to spend money that's not been funded, my time or my engineer's time, put this data on

			the internet. So that's costing me money that I don't necessarily have." 0:44:26.9 Steve: Put it on the internet so somebody can get it for free, write a paper, publish it in JGR and then you're gonna charge me $35 to read the paper. What the hell kind of business model is that" (P1)
Wow		Captures notable snippets, passages, examples.	"So one of the things that has been missing for a very long time in our program prior to the NSF data management plan is most of the science that happens after the expedition happens from people collecting those samples and then taking them back to the labs, working on a specific project. Before there was data management plans that data would sometimes just disappear and you'd never see it in particular if somebody published on that and didn't publish a data table, that means that data is not available for anybody else. And so to me, that was a real problem for the whole premise of that program is for putting data out there to advance science right." (P4)