Complex Adaptive Systems Theory Applied to Virtual Scientific Collaborations: The Case of DataONE

A Dissertation Presented for the

Doctor of Philosophy

Degree

The University of Tennessee, Knoxville

Arsev Umur Aydinoglu

August 2011

Copyright © 2011 by Arsev Umur Aydinoglu

All rights reserved.

Dedication

I dedicate this study to my Mîr, my lives would not have a meaning without his teachings.

Acknowledgements

I would never have been able to finish (maybe even able to start) my dissertation without the guidance of my committee members; help from my friends; and support from my family.

I would like to express my deepest gratitude to my advisor, Dr. Suzie Allard, for encouraging me to think outside the box and to try new approaches in my doctoral education. Without her guidance, intellectual contribution, practical solutions, and sense of humor, writing this dissertation would be painful at best. Her patience towards my wild ideas deserves special thanks.

I would like to thank Dr. Carol Tenopir, who supported me since the beginning of my graduate education and provided me with the opportunities to be involved in first class research.

I would also like to thank to Dr. Heather Douglas who contributed not only to my dissertation but also my intellectual development with her infinite knowledge on the philosophy of science.

Dr. Virginia Kupritz's experience and knowledge on mixed method studies helped me to find my way in the labyrinths of abundant data I collected.

I am grateful to the members of DataONE. They have accepted me to their meetings, agreed to be interviewed by me, and shared their thoughts and passion on one

of the biggest problems of today's science –data. Without their contribution this dissertation would not happen.

My very old friend Emine Yetgin for directing me to pursue a PhD and my colleagues Dr. Lei Wu and Dr. Katerina Spasovska for preventing me from derailing to pursue a PhD deserve a huge thank you.

Finally, I would like to thank my wife, Pinar. She was always there cheering me up and stood by me through the good times and bad. She also took very good care of my daughter, Neva Isik, when I was immersed in doctoral work and unable to help her.

.

Abstract

This study is the exploration of the emergence of DataONE, a multidisciplinary, multinational, and multi-institutional virtual scientific collaboration to develop a cyberinfrastructure for earth sciences data, from the complex adaptive systems perspective. Data is generated through conducting 15 semi-structured interviews, observing three 3-day meetings, and 51 online surveys. The main contribution of this study is the development of a complexity framework and its application to a project such as DataONE. The findings reveal that DataONE behaves like a complex adaptive system: various individuals and institutions interacting, adapting, and coevolving to achieve their own and common goals; during the process new structures, relationships, and products emerge that harmonize with DataONE's goals. DataONE is quite resilient to threats and adaptive to its environment, which are important strengths. The strength comes from its diversified structure and balanced management style that allows for frequent interaction among members.

The study also offers insights to PI(s), managers, and funding institutions on how to treat complex systems. Additional results regarding multidisiplinarity, library and information sciences, and communication studies are presented as well.

.

Table of Contents

Chapter 1 Introduction 1

Chapter 2 Literature Review 10

Scientific Collaborations 10

a. Scientometrics 11

b. Qualitative Case Studies 14

Complexity Theory 28

Emergence 51

Chapter 3 Background for DataONE 54

About DataONE 54

NSF the Office of Cyberinfrastructure 55

The Fourth Paradigm: Data-intensive scientific discovery 56

Sustainable Digital Data Preservation and Access Network Partners (DataNet) 57

Data Conservancy 58

The Data Observation Network for Earth (DataONE) 58

Chapter 4 Methods 68

Case study 68

Rationale for qualitative inquiry 70

Sampling 71

Rationale for selecting DataONE 72

Data collection 74

1. Semi-structured interviews 75

2. Naturalistic observations 79

3. Survey 81

Data Integration, Evaluative Criteria, and Analysis 81

Chapter 5 Results 86

1. Large number of components and counteracting forces: 88

2. Variation and Diversity 91

3. Connectivity, interdependence, and interaction 103

Interdependency & Working Groups 105

Communication behaviors 113

4. Feedback 116

5. Unpredictability 116

6. The edge of chaos 117

7. Self-organization, emergence, and strange attractors 117

8. Adaptation to environment (context) / pattern recognition / learning 122

a. The management 122

b. The Principal Investigator (PI)..... 125

c. Data Lifecycle 126

d. Working group structure 127

e. Other Changes 132

9. Historicity and path-dependence..... 135

10. Coevolution..... 139

Chapter 6 Additional Results Regarding Library & Information Science and Communication Studies 142

1. Library and Information Science 142

2. Communication Studies 147

1. Geography..... 149

2. Software capabilities..... 150

3. Cyberinfrastructure team vs. Community engagement team..... 153

4. Intimidation..... 157

3. Bridging Role 160

Chapter 7 163

Conclusion & Discussion 163

1. Summary 163

2. The NSF's 2nd year review 170

3. Contributions of this study 171

Contribution 1. A complexity framework for virtual scientific collaborations has been developed 171

Contribution 2. The complexity framework to virtual scientific collaborations has been applied to real case. 172

Contribution 3. It would be possible to have comparable results that increase our understanding of scientific collaborations. 173

Contribution 4. There are some lessons for the management/Principal Investigator (PI). 174

Contribution 5. There are some lessons for funding agencies. 176

4. Discussion on Multidisciplinarity, Communication Studies, and Information & Library Sciences 180

List of References 187

Appendix 204

Appendix A Interview Guide 205

Appendix B Survey Instrument 207

Appendix C Coding scheme of interviews for the complexity framework 212

Appendix D Coding scheme of interviews for other themes 217

Appendix E Tables 222

Vita 229

List of Tables

Table 1 – The development of coauthors patterns in selected fields (1980-1998) as reflected by the mean cooperativity. 12

Table 2 - Characteristics/Principles of Complex Adaptive Systems 34

Table 3 - Complexity Framework 87

Table 4 – Working Groups in DataONE 109

Table 5 – The reasons that led the termination of SSC 179

List of Figures

Figure 1– Basic concepts in complex adaptive systems 42

Figure 2 – Complex Adaptive System 50

Figure 3 – Coordinating nodes, member nodes, and candidate member nodes in the U.S. as of February 2011 62

Figure 4 – Organization chart for DataONE as of February 2011..... 65

Figure 5 – Data lifecycle as adapted by DataONE as of February 2011 67

Figure 6 – Involved or interested institutions in the U.S. as of February 2011. 90

Figure 7 – Involved or interested institutions world wide as of February 2011 90

Figure 8 – Data lifecycle as adapted by DataONE as of February 2011 93

Figure 9 – Subject disciplines in DataONE according to the responses to the survey 95

Figure 10 – Career ages of DataONE members according to the survey 99

Figure 11 – Organization chart for DataONE adapted as of 2010..... 106

Figure 12 – The number of individuals in working groups according to the survey..... 112

Figure 13 – DataONE adapted to complex adaptive systems figure 125

Figure 14 – Secondary working group membership according to the survey 130

Figure 15 – Organization chart that was submitted in the grant proposal as of 2009..... 134

Figure 16 – Organization chart for DataONE as of February 2011..... 135

Figure 17 – Information channels used to seek information regarding DataONE matters 147

Figure 18 – Communicating with own working group members and other working group members..... 148

Figure 19 – Complex adaptive systems framework..... 166

Chapter 1

Introduction

Literature and film have commonly depicted scientists as focused on individual pursuits. For example, there is the mad scientist working in his château to resurrect Frankenstein or the absentminded professor with his assistant to discover Flubber. However, today such individual efforts –of course aiming at saner scientific achievements– are not as prevalent in the scientific arena as they have been in the past and they receive limited funding. In reality, research has moved from individual effort to collaborative effort for the last 70 years. Simply, because that 'many hands make light work' and the need for experience, a combination of diversified skills, and expensive equipments to conduct these studies. Joining forces and resources increase efficiency and productivity. This became obvious after World War II because technology and scientific advancements decided the winners –specifically the discovery of radar, penicillin and atomic bomb (Guston, 2000, p. 114; Douglas, 2009, p. 16) through systematic funding by government.

Science and technology are an important aspect of current civilization. Benefits include higher living standards, increased life expectancy, new jobs and products with growing economies, decreased poverty and social inequality, and help in environmental

sustainability (Bush 1945; Steelman, 1947). In the U.S., as the leading country in research, the amount spent in research & development (R&D) expenses are approximately $333 billion in 2007, which was 2.68% of gross domestic product and there are 1.38 million researchers (UNDP, 2008). Given the size of the R&D expenses, number of researchers, and influences in our daily lives, efficiency in the functioning of scientific research is crucial. However, the efficiency of government funding in research has been in question since the 1970s. Detailed accounts of discussions and prescriptions to increase the efficiency can be found in science and public policy literature (Guston, 2000; Guston & Keniston, 1994); Kitcher, 2001; Pielke, 2007).

Today, the world faces ever more complex scientific challenges such as climate change –impact on land-based and ocean ecosystems– (IPCC, 2007); energy problems – increasing demand, climate change, fossil fuels– (IEA, 2009); space programs – permanent moon base and manned mission to Mars– (NASA, 2006); research on subnucleic particles –Large Hadron Collider experiments – (LHC, 2009); and destructive pandemics –AIDS, swine flu, and malaria– (WHO, 2009). These challenges are like multi-faceted problems. Each discipline is dealing with only one facet; therefore, failing to respond to all of them as the activities of these individual disciplines are not coordinated with each other and most of the time their solutions are contradicting with each other. A new strategy, a holistic approach is needed because the problems reside in multiple disciplines. Successful negotiation of these challenges relies on multidisciplinary scientific collaborations, an emerging model that suggests new organizational forms and

new patterns of work –that is multidisciplinary scientific collaborations. They are new because the researchers involved in them have to get out of their disciplinary comfort zone, process and integrate data and information generated by other disciplines, communicate them results. This is a new relationship, a new workflow, a new structure, a new organization, indeed a new model for scientific research. In addition to scientific challenges, changes in the funding environment (for instance the increasing involvement of the private sector through corporate social responsibility programs and intermediary role of non-governmental organizations in directing funds), the changes in the public perceptions of and expectations from science, and developments in communication and information technologies (globalization and the Internet) have changed the functioning of scientific collaborations. In order to increase the efficiency of the research activities and make a better use of public money, the new developments should be taken into consideration in the discussions. This study contributes to the literature by examining scientific collaborations from a complexity theory perspective with a focus on the role of communication and information behaviors in this complex system.

To begin two terms need to be explained: scientific collaboration and emergence. First, a scientific collaboration is a purposeful working relationship between two or more people, groups, or organizations in order to research phenomena, to develop a scientific instrument or technology, to build a facility, and/or to publish a study. There are different reasons to collaborate but simply put, collaborations form to share expertise, credibility, material and technical resources, symbolic and social capital.

Historically, the investigation of collaborations in the scientific arena had started as co-authorship studies by de Solla Price (1963; 1977) and Garfield (2009). The increase in the number of co-authored articles in 1960s made de Solla Price (1963) argue that sole authorship would be extinct by the 1980s. Time proved that Price was wrong about the extinction of sole authorship; however, research has become a collaborative activity over the last seventy years as mentioned above. The number of co-authored papers and the number of citation rates per papers have increased in all fields (Glanzel, 2001). Scientific collaborations have grown bigger in every dimension (size, budget, resources, and magnitude) and become the primary way scientific research is being conducted. It is called big science (de Solla Price, 1963). Big science is large-scale projects, which needs vast resources, funded by national governments or groups of national governments. The term was popularized by Alvin M. Weinberg's (1961) article in *Science* in which he compared current efforts in big science such as space research and particle accelerators to the glory of pyramids and Notre Dame Cathedral; however, he also pointed out the financial burdens on the budgets (1961, p. 161-4). Today, it has been called mega science (Bodnarczuk & Hoddeson, 2008, p.510); however, the term 'mega' emphasizes too much the size (budget, personnel, etc.) and overlooks the complexity of the collaboration, which is the main difference of today's collaboration and the main key (and also challenge) to the efficiency. The focus of this study is in between these two extremes (a simple co-authoring activity between two scholars and mega science).

Second, complex adaptive systems are systems “that have a large numbers of components, often called agents, that interact and adapt or learn” (Holland, 2006, p. 1). This study perceives scientific collaborations as complex adaptive systems and elaborates their features accordingly. For instance, because of the counteracting forces acting inside them (each agent/collaborator either individual or institutional has its own agenda), collaboration’s behaviors become complex. A basic protocol (such as an internal newsletter for communication) could easily lead to a complex system due to multiple interactions it triggers among agents. Interactions encourage or discourage (feedback) certain actions, behaviors, and communication and information flows, which puts the system into a dynamic equilibrium state or simply makes it adaptive. As a result, non-linear relationships dominate the collaboration and transform them into complex adaptive systems. The advantages of operating as a complex system are being open to learning, ability to adapt change, resilience to external and internal threats, being cost effective, and being innovative. Therefore, the assessment of a scientific collaboration is crucial in deciding allocating limited resources to the one that has the maximum potential to be successful. However, complexity theory comes with its own shortcomings that challenge the very basics of ‘good’ science: prediction and control. According to complexity theory long-term prediction is not possible; likewise the manipulation of the variables in the system. What is possible is short-term prediction and encouraging/discouraging certain behaviors in the system. It is retrospective and strong in explaining.

The second term is ‘emergence’, one of the most important concepts of complex adaptive systems. It is used on purpose because the terms ‘formation’, ‘establishment’, or ‘creation’ do not fully cover what it is happening in today’s scientific collaborations. These words imply an external force in the occurrence of the collaborations; however, today more and more collaborations are ‘self-organizing’ themselves. Their formation is not dictated from a higher authority but the individuals feel the urge to do something for various reasons and organize from the bottom-up. Emergence, a system that results from the actions of its interacting agents, makes more sense than the other terms mentioned above because it explains how individual researchers come and work together around a phenomenon or a problem of interest to offer new knowledge. In addition, the emergence concept is related to the “the whole is bigger than the sum of its parts” rule of systems that was mathematically proved by Poincare (Waldrop, 1992). Examples are everywhere. For instance, in neurology, the brain might be composed of cells but its functions such as thinking and memory are beyond the capabilities of these cells. The communication among cells creates something new, something that does not exist before (Mitchell, 2009). Another example is music. When individual instruments in a jazz band play altogether, the melody emerges. In these examples, memory and melody are ‘emergent’ properties. Emergence is something that occurs between the lower-level and higher-level properties (Sawyer, 2005, p. 3) and the two-way interaction among them. For instance, individuals (lower-level) affect the economy (higher-level) by their individual decisions but individuals are also affected by the economy such as after a crash in the stock market.

In this study, individual researchers and the practices they have are lower-level properties and the collaboration itself (with products, outcomes, relationships etc.) are the higher-level properties.

In conclusion, scientific collaborations are an important element of modern civilization. Bertrand Russell (1961) once said “Almost everything that distinguishes the modern world from earlier centuries is attributable to science” (p.20) and scientific collaborations are an important part of it. More has to be learned about them in order to have a better functioning research system, and thus, better lives. Current studies employ linear models and have limited power to explain the dynamics of the research process. This research studies a new paradigm that may be capable of being employed to overcome these limitations and suggests that complexity theory can be a tool for the new paradigm.

This study posits that if scientific collaborations behave like complex adaptive systems, they should demonstrate basic features of such systems. Therefore, the research question is: “How can the emergence of DataONE –a multidisciplinary, multinational, and multi-institutional scientific collaboration– be explored from a complex adaptive systems perspective?”

The outline of the study is as follows. In the second chapter, the literature review is presented. There are three bodies of literature of interest for the study. The literature on scientific collaborations primarily consists of bibliometric studies and case studies, and

this literature review identifies gaps in the literature. Second, complexity theory and complex adaptive systems are defined and summarized. Third, the emergence concept is reviewed. This chapter introduces the terminology and concepts that can fill the gap that exists in the current literature.

The third chapter provides the background information for DataONE: the history of computational research in the U.S., the NSF's data vision and the DataNet Solicitation.

The fourth chapter presents the methods used in this study. This chapter starts with the statement of the research question which is followed by the introduction and the rationale of the method –case study– to answer the research question. Afterwards, the selected case – DataONE (Data Observation Network for Earth) and the rationale for selecting it are explained. The chapter then explicates the data collection process by reviewing the three data collection methods used in this study: semi-structured interviews, naturalistic observations, and online survey. This chapter ends with how to integrate the mixed methods and how to conduct analysis on the data coming from different sources.

Chapter 5 introduces the complexity framework that is developed for the assessment of scientific collaborations. The results reported in this chapter are analyzed according to the framework. Ten concepts in the framework are tied to the findings that are generated through interviews, observations, and online survey.

The findings of particular interest to scholars of library and information science and communication studies are discussed in Chapter 6.

The final chapter summarizes the findings and discusses their implications for the future research directions.

Chapter 2

Literature Review

Three bodies of literature are examined in order to provide a background for this study. These bodies of literature focus on scientific collaborations, complexity theory, and emergence. These bodies of literature need to be explained in order to understand which gap the research question fills and how. The first part of the literature review is about scientific collaborations. The methodologies and the topics that are covered by scholars so far regarding scientific collaborations are presented and the contribution of this study is discussed. In the second part, the complex adaptive systems theory and emergence concepts are introduced so that the readers can follow how the complexity framework is developed. The complexity framework (and this study) is the first step of a developing a tool to assess scientific collaborations.

Scientific Collaborations

Scientific collaboration is a family of purposeful working relationships between two or more people, groups, or organizations in order to research phenomena, to develop a scientific instrument or technology, to build a facility, and to publish a study (Hacket, 2005). There are different motivations to collaborate. According to Maienschein (1993) the three reasons to collaborate are that (i) individuals need help and division of labor will increase efficiency, (ii) collaboration increases credibility through its members own credentials and acceptability, and (iii) collaborations could attract more resources. In a

nutshell, collaborations form to share expertise, credibility, material and technical resources, symbolic and social capital.

Scholars have different classifications which are based on the methods to study. Vasileiadou (2009) adds surveys as well. Wagner (2002) divides the literature on international collaborations into three parts: i) scientometrics; ii) social studies of science (descriptive, historical, or qualitative studies about collaborative networks); and iii) policy studies and official government publications (p.13-4).

For the purpose of this study, the research on scientific collaborations is grouped into two categories (a) scientometrics (de Solla Price, 1963, 1977; Garfield 2009; Vasileiadou, 2009) and (b) case studies using qualitative methods focusing on military & scientific community, organizational features, multidisciplinarity and other studies (Cloud, 2001; Harper, 2003; Agar, 2006; Shrum, Genuth, & Chompalov, 2007). There are a limited number of studies utilizing surveys and one comprehensive study employing mixed methods done by Shrum, Genuth, and Chompalov (2007); however the sample for the latter covers collaborations that had been active before 1990s. Even though that study offers valuable insights on many topics, the advances in information and communication technologies has changed the rules of the game regarding communication and data/information behaviors of collaborations, and thus, new studies are needed.

a. Scientometrics

Studies on scientific collaborations started with scientometrics in 1960s. Scientometrics is “a quantitative, statistical method of analysis using bibliometric data to

describe existing patterns of linkages among scientists" (Wagner, 2002, p. 13). Price and Garfield were first scholars to study citation patterns. Other scholars investigated co-authorship patterns for different scientific fields for example using Science Citation Index.

The number of co-authored publications has been increasing (Glanzel & De Lange, 1997; Ding et al., 1999). For instance, Glanzel's (2002) study shows the patterns in biomedical research, chemistry and mathematics (see Table 1).

Table 1 – The development of coauthors patterns in selected fields (1980-1998) as reflected by the mean cooperativity, "the average number of authors contributing to one paper" (Glanzel, 2002, p.465).

	1980		**1986**		**1992**		**1998**	
Subject Field	**Papers**	**M**	**Papers**	**M**	**Papers**	**M**	**Papers**	**M**
Biomedical research	64501	3.47	74360	3.96	86544	4.57	98793	5.13
Chemistry	66576	3.07	69703	3.27	80083	3.50	94600	3.82
Mathematics	14385	2.22	11892	2.30	13362	2.36	18729	2.59

Some of the most common studies that are related to co-authorship and collaboration are the ones focusing on a country or region or discipline or sector. Collaboration in Central African countries (Boshoff, 2009), citation patterns of the publications of South African scientists by the type of collaboration they operate in (Sooryamoorthy, 2009) collaboration between China and G7 countries through publications (He, 2009), collaborations in India to chemical sciences (Sangam, 2009),

collaboration in epidemiology and public health (Navarro & Martin, 2008), collaboration in social sciences in Turkey (Gossart & Ozman, 2009), and cooperation patterns in neuroscience (Braun, Glanzel, & Schubert, 2001) are to name a few areas that are studied.

Today the pressure of 'publish or perish' has been higher than ever and the competition is so fierce in some sectors, such as pharmaceuticals, a new type of collaboration has been born in order to extract the maximum amount of scientific and commercial data and analyses through carefully planned and prepared papers (Sismondo, 2009). In this new type of collaboration "clinical research is typically performed by contract research organizations, analyzed by company statisticians, written up by independent medical writers, approved and edited by academic researchers who then serve as authors, and the whole process organized and shepherded through to journal publication by publication planners" (Sismondo, 2009, p. 171).

Even though, scientometrics is an important field and provides valuable insights to the field, its contribution, in regards to the dynamics of scientific collaboration is limited for various reasons. First, as it was pointed by Cronin (2001), is the issue of hyperauthorship –articles authored by more than 100 authors, which is a common practice in particle physics and biomedical fields. The dynamics of co-authoring and collaborating between two authors is different from the dynamics of co-authoring and collaborating among 100 authors; however, through a bibliometric analysis, the researchers have a limited understanding of this difference. Second, is the inclusion of

honorary co-authors to increase the credibility of studies and the chance of getting published (LaFollette, 1996). Physics and medicine are the fields which received most of the funding. Through bibliometric analysis, understanding the dynamics of collaboration is again limited as the honorary authors' contribution to the final product is merely a name. Third, Katz & Martin (1997) argue that only certain roles in the collaborations are awarded by authorship. For instance, someone who actually collected the data in the field might not get credit in the article. Again, the bibliometric analysis fails to tell about the dynamics of the collaboration. Finally, in Subramanyan's (1983, p.35) hypothetical but feasible example "a brilliant suggestion made by a scientist during casual conversation may be more valuable in shaping the course and outcome of a research project than weeks of labour-intensive activity of a collaborating scientist in the laboratory." These reasons demonstrate the limited explanatory power of such studies in explaining the dynamics of collaborations, because the contributions of the collaborators (researcher, scholar, author, data collector, technician, analyzer, etc.) are not always reflected in the final product. A bibliometric analysis is a powerful tool yet it can only reveal what is in the final product. If the contributions are not in the final product, which might be the case due to various reasons summarized above, it has limited power.

b. Qualitative Case Studies

1. <u>Military & scientific community collaborations</u>

Another line of research on scientific collaborations is qualitative case studies. The first collaborations were established between the military and the scientific

community and studied in detail by scholars. These collaborations are important for two reasons: the military's direct support and paving the way for government support. Although collaborations existed in the U.S. (could be traced back to the Civil War) or Europe (for instance the Kaiser Wilhelm Society in Germany in 1911), it has become 'structured' after the World War 2. Since 1945 military has provided more than 50 percent of federal R&D expenditures (Guston & Keniston, 1994, p. 16). World War 2 has been the worst thing that humanity ever faced and had many negative impacts on everything but science. Guston (2000, p.114) notes that the development of radar, penicillin, and the atomic bomb was crucial to the victory of the Allies. Additionally, "…the importance of science for American survival and prosperity were amply illustrated during the war. The stunning successes of radar, penicillin, and most dramatically, the atomic bomb, made apparent to the country how powerful an ally science could be" (Douglas, 2009, p. 33). President Roosevelt's science advisor Vannevar Bush (1945) made this relationship official by mentioning the importance of science in achieving "national security" (p.17). The second important outcome from these collaborations included paving the way for government support. Politicians and scientist realized that the technologies and science developed in the war time could also be very useful in peace time. Therefore, again with Vannevar Bush's vision (health & public welfare), government support of scientific research has become indispensible and constantly growing. With the support from military and government, with the former focusing on applied research and the latter focusing on basic research, bigger scientific collaborations

and big science projects have become feasible. Moreover, a collaboration culture was born.

There are many studies that examine the collaboration between the scientific community and the military. For instance, van Keuren's study (2001) is about the US Naval Research Laboratory between 1948 and 1962. A satellite, which could work as an electronic intelligence satellite and astronomical observatory, was built together with civil astronomers and military personnel. In another similar project, CORONA, the first American enterprise for secret photography from space, was later used for earth sciences by researchers (Cloud, 2001). Benefits of military support have been obvious in other disciplines too. The collaboration working on Project Vela Uniform, which was a research program in seismology to have a better detection and identification of Soviet underground nuclear-weapon tests, had transformed seismology from a small academic discipline to a large academic-military-industrial enterprise (Barth, 2003). In the case of asteroid studies, astronomers and planetary scientist initiated collaboration by promoting asteroid collision mitigation studies in order to receive funding (Mellor, 2007). However, things were not always smooth in military – scientific community collaborations. In meteorology, for example, there were tensions between the military and scientific community in collaborated studies after World War 2 when the military retained the control of meteorological research funding (Harper, 2003). Such tensions were experienced in Soviet Russia as well as in computing (Gerovitch, 2001). Sometimes it took decades to have tangible results from such collaborations such as the case of the Air

Force which funded molecular electronics in 1950s and the Naval Research Laboratory did the same in 1980s. Their efforts led to nanotechnology in the last decade (Choi & Mody, 2009). The collaborations between the military and the scientific community are well documented and studied.

The examples above demonstrate how different stakeholders (military, government, and scientific community) can collaborate and what can be the impact of the collaboration. The diversity of the stakeholders provides both challenges and opportunities to both sides then cannot be thought of before. The diversity concept plays a key role in scientific collaborations which is going to be explained in detail in further chapters. Furthermore, it provides a brief history of government-funded basic research.

2. Organizational features

Scientific collaborations behave like organizations because they have the five basic features of organizations identified by Scott (1981): (1) Social structure, which could be normative, cultural-cognitive, or behavioral; (2) participants, who are individuals who contribute to the system to gain something in return; (3) goals, which are the desired ends; (4) technology, which is everything that is produced by the organization; and (5) environment, that is the context that organizations exist in physically, socially, technologically, and culturally and are open to impacts from there (p.18-24).

If these features are adapted to scientific collaborations then (1) scientific collaborations have different *social structures*, for instance Shrum, Genuth, and Chompalov (2007)[1] identified four organizational types – bureaucratic, leaderless, non-specialized, and participatory (p. 129); (2) [participants] researchers provide their knowledge and technical expertise to collaborations and in return receive many things – career boost, learning from seniors, satisfaction to work on their passions, access to equipment and funds, or just fame; (3) [goals] raison d'être for a collaboration to come together – develop a technology or knowledge, build a facility or equipment, etc.; (4) *technology* is produced in most of the collaborations and other spin-offs; (5) collaborations are open to the social, politic, cultural and technological climate – which leads to opportunities and challenges such as the increase in the funding of defense-related research in Cold War era (Guston & Keniston, 1994) or high performance computing and communications related research in the last decade (NSF, 2006).

As scientific collaborations behave like organizations, their organizational features are studied by scholars. For instance Hong's study (2008) is about the sources of authority, reasons for conflict, and group dynamics in an isotope lab at a Chinese

---

[1] It has to be mentioned that Shrum, Genuth, and Chompalov's study is not only a qualitative case study but a mixed methods study using in-depth interviews and surveys of 53 cases (collaborations).

university. Shrum, Genuth, and Chompalov (2001) examined the relationship between trust, performance, and conflict in 53 physics and related sciences' collaborations. They identified three types of conflict: (i) conflict between project teams, (ii) conflict with project management, and (iii) conflict between scientist and engineers. Jeffrey (2003) did an ethnographic study as a participant observer in a multidisciplinary research group as an intermediary person to smooth the waters between social scientists and simulation modelers in a collaboration that was investigating the desertification in South East Europe. Conflicts are inevitable in social groups including scientific collaborations; thus, knowing how to deal with them becomes crucial if a scientific collaboration is going to function properly and even survive.

Another important topic for an organization; and also for a collaboration, is the identity. Hackett (2005) studied how a research collaboration establishes identity and the tensions in them such as autocracy vs. democracy, varieties of risk, role conflicts, openness vs. secrecy, competitive cooperation, and balancing continuity and change. For most of the members, the scientific collaborations are on the side. They have their tenure-track jobs in the academia and work for maybe a couple of projects simultaneously. Therefore, they wear different hats. The borders between the projects and institutions might get blurred if there is uncertainty (generally there is). The researchers have to juggle with the different roles/hats they have and juggling brings tensions.

A recurring theme for conflict is the one between the researchers and engineers or technicians –especially when the collaboration is formed to build something. In Shrum,

Genuth, and Chompalov's study (2007) this conflict is significant when autonomy is low and interdependency is high (p.173-4). The role of engineers and technicians could be quite important in scientific collaborations. Horning (2004) described the importance of technicians' and engineers' roles with the problems of formal training for them. Timmermans (2003) argued for an analysis of the process of crediting people for their scientific accomplishments when he studied the life of Vivien Thomas, a black technician in surgical research, as Thomas did not get any credit for his studies due his profession and race. The role of technicians and engineers are overlooked, especially in scholarly works.

Shrum, Genuth, and Chompalov's study (2007) not only covers the types of organizations mentioned above but also hierarchy and decision-making in scientific collaborations. Sims' study (2005), in which he examined a pulsed-power facility at the U.S. Los Alamos National Laboratory, is also about hierarchy, social order, and norms of conduct in scientific collaborations. The safety procedures at the lab become rituals and contribute to the social order in the collaboration and have an important role in defining the organizational culture.

Productivity is the main concern for scientific collaborations as for most of the organizations; hence, it has been studied by many scholars. Allison (1980) discusses the disciplinary differences in the distribution of productivity and the functional relationship between productivity in scientific collaborations. According to a quantitative study done by Lee and Bozeman (2005), scientists who collaborate more, publish more. According

to their study (2005) scientists who collaborate more do publish more papers, but when the count of publications is adjusted for the number of authors per paper, the influence of collaboration falls below the level of statistical significance. Scientific collaborations from developing countries also get their share of research. Through a study done in the Institute of Biomedical Research of the National University of Mexico, developments in international visibility, participation in invisible colleges, increase in productivity, and increase in horizontal collaboration were observed (Lomnitz, Rees, and Cameo, 1987). Wagner (2008) presents a new model of collaboration for developing countries through complexity theory and discusses the policy issues related to funding of scientific research. The relationship between collaboration and productivity, and developments in information and communication technologies in Africa and India were also studied by Duque et al (2005). The quality of research is as important as the productivity. Presser (1980) found a small relationship between the research performed collaboratively and the quality of scientific research, whereas Hart (2007) "found no evidence to support the superiority[2] of co-authored articles" in the discipline of academic librarianship. These results might seem odd but it should be kept in mind that the scholarly productivity is not the only reason for researchers to collaborate. There are different motives to collaborate.

---

[2] By 'superiority' Hart means quality and impact which he measures through citation count.

In addition, a collaboration might have an impact on the scientific community beyond its life span such as a telescope or facility built that serves for tens of years. In addition, these results are contradicting with the previous studies (Glanzel, 2001; Ding et al, 1997).

3. Multidisciplinarity and other studies

Scientific collaborations do have different features than organizations as well. One of them is the multidisciplinary structure (except the unidisciplinary collaborations of course). Cummings and Kiesler (2005) examined what kinds of problems occur in multidisciplinary and multi-institutional collaborations by surveying the principal investigators of 62 collaborations who received grants for their projects. Their study revealed that the multi-university projects were more problematic than the multidisciplinary projects because of the coordination issues that were brought by distance. Mazur and Boyko (1981) studied the success and failure of five big science oceanographic research projects and found that the source of origin of the research problem, the quality and interest of participating scientists, the presence of a persistent scientific leader, and the degree of project independence from the funding agency differentiates a successful project from a failed one, whereas formality of collaboration, social friction among researchers, and communication problems did not have a significant effect.

Another area is how innovation occurs and the role of scientific collaborations in it. Cambrosio, Keating & Mogoutov's study (2004) mapped collaborative work in

biomedicine and examined innovation in the field through antibody reagent workshops, which are indeed collaborations to identify and classify reagents. Mirowski & van Horn's study (2005) is about innovation in and commercialization of scientific research through contracts in the biopharmaceutical sector.

Brunn and Sieda (2008) argued that collaborations tend to select different kinds of knowledge networking strategies, depending on the perception of the problem they work on. Well defined problems were studied with modular and translational networking, whereas ill-defined problems were studied with integral knowledge networking[3].

Like technicians and engineers, computer scientists have also become important for the success of scientific collaborations in the last two decades. The National Science Foundation have been funding high performance computing since 1960 (NSF, 2006, p. 30) but it was not until 1991 when the Congress passed the act for high performance computing and communications (Computer Science and Telecommunications Board, 1995, p. 89) that the importance of cyberinfrastructure had been realized in the success of a scientific collaboration; and thus, related literature has started to develop. For instance,

---

[3] Modular knowledge networking refers to activities in which tasks are modularized and distributed to autonomously working agents; in integral knowledge networking tasks are handled as a joint effort; and translational knowledge networking is a combination of both.

Hine (2006) argues that the use of databases in research leads to changes in work practices, communication regimes, and knowledge outcomes, all of which are very important in the functioning of a scientific collaboration. WikiProteins is a project to create a Wikipedia-like single portal to access biomedical data and resources, and make it maintained by the biomedical research community (Lopresti, 2008). Agar's study (2006) focuses on the effects of computers on the first generation of scientists who used them. Data sharing issues have been a problem for researchers interested in detecting gravitational waves for almost a century (Collins, 2004). Laser Interferometer Gravitational-Wave Observatory (LIGO) collaboration is one the biggest scientific collaborations with 60 institutions in 11 countries (LIGO, 2010). The nature of the phenomenon provided an additional challenge in the forming of this collaboration. Data, the waves that hit the detectors, comes with 'noise' that needs to be eliminated. Different research groups have different methods and calculations to eliminate the noise. Eliminating the noise is so crucial that data cannot be shared without it. However, research groups want to integrate their data with the others if the noise in others' data is eliminated with their method and calculation. It was one of the biggest obstacles in forming LIGO which took decades and is examined in great detail by Collins (2004). A similar problem occurred for data in Antarctic science (Dean et al 2008). This time governments and politics were involved in the negotiation of sharing data among scientific institutions.

A recent area of study about scientific collaborations is ethics. Hedgecoe and Martin (2003) focus on development of pharmacogenetics (conventional small molecule drugs) and the social and ethical issues it brings; Rasmussen (2004) examines the collaborations between pharmaceutical companies and laboratory-based researchers in universities; and Montgomery and Oliver (2009) investigate how guidelines for ethical scientific conduct for government funded projects are created.

Above, the literature on scientific collaborations that uses qualitative case studies were summarized by their approach: the studies focusing on the collaboration between military and scientific community, the collaboration as organizations and organizational features, and the structure of collaborations such as forming, disciplinary structure, etc. However, even though they provide crucial insights on how things are done in collaborations, these case studies are far from producing comparable results. As Vasileiadou (2009) argues "What they all have in common is an understanding of the practice of collaboration as an inherently more "messy" process, with the risks, tensions and local contingencies it entails. … they all lack a systematic approach which could help compare those results in different settings."

4. Virtual Research Collaborations

The final body of literature that is examined for this study is the study of virtual organizations (or distributed organizations). A virtual organization is "a group of individuals whose members and resources may be dispersed geographically and

institutionally, yet who function as a coherent unit through the use of cyberinfrastructure" (NSF, 2011). The two key elements of virtual organizations are having an organizational structure without sharing a physical space and using computer-mediated communication to function (Cogbern, Santuzzi, & Velasquez, 2011).

Today, the problems that scientists deal with require different resources (human, technology, and equipment) and having these resources in one single place is not possible; hence, virtual scientific collaboration has become a necessity and almost a norm to conduct scientific activity. However, virtual organizations have their challenges that could be categorized into three groups: "(1) logistical problems, such as communicating and coordinating work across time and space, (2) interpersonal concerns, such as establishing effective working relationships with team members in the absence of frequent face-to-face communication, and (3) technology issues, such as identifying, learning, and using technologies most appropriate for certain tasks" (Straus, 1996).

Research on virtual organizations is relatively new but addresses a wide range of dimensions. According to a study conducted by Powell, Piccoli, and Ives (2004) using a life cycle model, studies on virtual organizations focus on four general categories: (i) input (design, culture, training); (ii) socio-emotional processes (trust, cohesion, relationship building); (iii) task processes (communication, coordination, task-technology fit); and (iv) output (performance, satisfaction).

The network characteristics (centrality, hubs, and incoming/outgoing links) of virtual organizations is another hot topic. The relationship between them and the performance (Cronin & Meho, 2006; Haythornthwaite, 2009), and information flow and team dynamics (Panzarasa, Opsahl, & Carley, 2009) are studied.

The increasing number of multidisciplinary research projects has increased the number of studies on the diversity of virtual teams as well. For instance, when there is too much diversity researchers establish cliques, stop communicating, and even disrupt other's efforts (Adamic & Glance, 2005; Stvilia, Twidale, Smith, & Gasser, 2008).

Furthermore, scholars have investigated the performance of virtual teams heavily (Aubert & Kelsey, 2003; Janicik &Bartel, 2003; Kacen, 1999; Ancona & Caldwell, 1992). Performance is based on scholarly and non-scholarly production (such as patents) and adherence to budget and deadlines.

However, there are two short comings of the literature on virtual organizations. First, the studies focus on small teams or groups when applied to scientific research context. Considering big sized organizations, only commercial organizations have been studied so far. Although, there are some similarities between profit-based (commercial) and non-profit based (research); they are actually different kind of organizations because of their raison d'être: profit and answering a research question respectively. Second, which is also valid for every kind of scientific collaboration (virtual or not). The studies tend to treat research teams as traditional organizations. However, most of them are

complex organizations in the sense of complex adaptive systems which is going to be explained in the next chapter. Briefly, complex systems are based on nonlinear relationships among components and they are non-reductionist (Mitleton-Kelly, 2003). The literature on virtual teams uses linear theories to explain their behaviors. A nonlinear system's behavior cannot be explained through a linear equation.

In summary, in this section, the literature on scientific collaborations is presented in two categories: (a) scientometrics and (b) case studies using qualitative methods focusing on military & scientific community, organizational features, multidisciplinarity and other studies, and virtual organizations. There are a limited number of quantitative studies that are mentioned in the text when they are relevant; however, they are not many generalizable findings regarding scientific collaborations.

Complexity Theory

The second body of literature relevant to this dissertation focuses on complexity theory. It is hard to argue that a unified theory of complex systems exists (Mitleton-Kelly, 2003, p.1; Mitchell, 2009, p.14). Complexity theory has close ties with chaos theory and other concepts from biology, physics, and chemistry such as catastrophe, autopoiesis, chaos, dissipative structures, autocatalyctic process, attractors, multi-agent systems, thresholds and transformational processes, fractal geometry, fuzzy logic, and systems theory (Salem, 2009; Smith & Jenks, 2006; Mitleton-Kelly, 2003); however, in this study the focus is on social sciences.

The discussion of complexity theory begins by outlining the difference between linear and non-linear models. Since Descartes, linear modeling has dominated the scientific world because of its freshness, competence and convenience for calculations. Linear systems are simple and deterministic, and therefore, variables in linear systems could be manipulated (at least theoretically) and are definitely predictable. It was revolutionary and became an important tool for science, because if a phenomenon could be modeled linearly, it could be controlled (such as the acceleration of something through applying force – Newton's first law) or foretold (the orbits of the planets in the solar system). The main hypothesis behind this view is that a phenomenon is the aggregation of its components, which are variables, so it should be broken down to its smallest units and they should be studied in order to understand it.

However, many phenomena in life are neither linear nor can they be reduced to its simplistic units or both. In mathematics, the inability of linear analysis to explain non-linear systems was proved by Poincaré at the beginning of 1900s; which could be translated as a non-linear system is more than the sum of its parts (Waldrop, 1992). In non-linear systems small inputs can have large system effects (or vice versa) and there is sensitivity to initial conditions which makes prediction almost impossible.

The problem was working with non-linear systems was beyond human computational ability. When non-linear relations are realized, the related data were not preserved and/or the non-linear relations cannot be measured/calculated due to their complexity. This happens because "modeling the nonlinear outcomes of many interacting

components has been so difficult that both social and natural scientists have tended to select more analytically tractable problems" (Anderson, 1999, p. 217) which produces deficient and incomplete reflections of reality. Thus, scholars end up having a discipline that is not connected to life and it does not help us to control or predict phenomenon as a result of its dependency on linear modeling.

Sometimes non-linear relationships were simply disregarded by unrealistic but more tractable feasible assumptions. For instance, in economics it is assumed that 'there is equilibrium in markets' despite all the opposite evidence (Waldrop, 1992, p. 255) or in archaeology social and economic systems were assumed to be in equilibrium (Bentley & Maschener, 2007, p. 15-1-2); both of which contradicts the reality.

Due to the messiness, the study of non-linear systems had not got much interest until the 1960s when, with the development of computers, computational power has increased enormously; and thus, solving non-linear equations have become easy (Gleick, 1987). Consequently physicists, meteorologists, economists and chemists adapted non-linear models to their disciplines.

The main difference between linear and non-linear systems is the center of attention given by researchers: interaction. Instead of focusing on units, in complexity theory, researchers focus on interactions. Interaction is an intricate relationship among units or variables and is generally short ranged (Cilliers, 1998, p. 4). For example, information is generally received from immediate neighbors. As the information travels

unit through unit, it can be enhanced, suppressed or altered in many ways, such as the telephone game. Positive and negative feedback loops exist in interactions; hence, some actions are encouraged and some discouraged. Everything that is related to the system could be found in interactions and the level of analysis becomes interactions in complexity theory. As Nobel chemist Prigogine (1997) argued in his book 'The End of Certainty', this new paradigm is interested in instability, disorder, diversity, and non-linear relationships rather than the traditional mechanistic Newtonian view which dealt with stability, order, equilibrium and linear relationships.

There is not a unified complex adaptive systems (CAS) theory but in definitions there are some concepts that are indispensible such as agents, interaction, co-evolution, and emergence. Here two definitions are offered:

- "The theory of complex adaptive systems (CAS) originated in the natural sciences and articulates how interacting agents in systems adapt and coevolve over time, and who, through their interactions, produce novel and emergent order in creative and spontaneous ways (Webb, Lettice & Lemon (2006)."
- "A complex adaptive system consists of a large number of agents, each of which behaves according to some set of rules. These rules require the agents to adjust their behavior to that of other agents. In other words, agents interact with, and adapt to, each other (Stacey, 2003, p.37)."

According to Kauffman (1993) when the relationships are simple, the system's behavior is easy to understand, explain, and predict, which is what is done in linear modeling. In the other extreme, when immeasurable nonlinearity dominates the system, it looks random and chaotic[4]. Complexity, sometimes called as 'order in disorder', is between them, not easy to understand but not impossible either.

Complexity theory focuses on "organizing rather than organization" (Weick, 1979) and prescribes that "…chaos is a science of process rather than state, of becoming rather than being" (Gleick, 1987, p. 5). It is continuous recreation of interactions and relations between units, which also results in dynamic equilibrium. It is this continuous recreation, redefinition and emergence that makes it harder to understand, predict and equalize.

According to Holland (1998) in complex systems overall patterns are greater than the sum of the parts –as Poincare pointed out earlier about non-linear systems and all complex systems are non-linear– and also such systems may act coherently without domination by a central source, which means the system cannot be localized to its subsets. This approach suggests bounded rationality principle. The units cannot know the

---

[4] Chaotic used here in the sense of lacking order and neatness. Technically speaking chaotic systems are studied and modeled mathematically.

big picture due to lack of information and their limited information processing ability. They can only know about their immediate neighbors. Thus, they position themselves according to them. It is very common in explaining survival and extinction in habitats in evolutionary biology. No creature knows what is going on in this planet but position themselves (such as developing camouflage skills to hide or long legs to run faster) to their prey and hunter. The whole habitat is in a state of dynamic equilibrium tied to each agent. This is called coevolution (Waldrop, 1992, p. 259-60) or coevolution to the edge of chaos (Pascale, Millemann, & Gioja, 2000).

Complexity theory has its own challenges that oppose the basics of science: prediction and control. Due to the sensitivity to initial conditions, it is not possible to make predictions. If there is no prediction, there is no control. Complexity theory becomes retrospective and used to explain past events. However, an infinite number of different explanations of past events can be constructed –as McKelvey (1999) suggests it is not different from witchcraft: "…without a programme of experimental testing complexity applications … will remain metaphorical and if made the basis of consulting agendas … are difficult to distinguish from witchcraft' (p. 21)." Experimenting is not easy in complex adaptive systems. As a result, for human systems, many of the results come from computer simulations not from empirical observations (Houchin & MacLean, 2005). On the other hand, short term prediction might be possible: "the impact of an incremental change can be predicted in the very short term" (Thietart & Forgues, 1995, 26)

In addition, according to some scholars it is not clear whether it is a theory, merge of theories or a framework, and a common terminology exists (Morel & Ramanujam, 1999). A single, unified theory of complexity or complex adaptive systems does not exist because complex systems or complex adaptive systems can be found in different systems, inorganic or organic, and at different levels from molecular level, cellular level to population level. These systems have been studied by scholars from different disciplines and the introduction of a single unified theory has not been possible so far. Yet, there are basic features or characteristics or concepts that have been acknowledged in the literature (although some argue that circularity exists among key concepts (Houchin & MacLean, 2005)). Using the seminal articles in the field of organizational studies, Table 2 identifies the most important ones. It does not refer to other disciplines as in the previous section it was made clear that collaborations indeed behave like organizations. These features or concepts are used to have a better understanding of complex adaptive systems through a framework (see Table 2).

Table 2 - Characteristics/Principles of Complex Adaptive Systems

Features of Complex Adaptive Systems	**Thietart & Forgues (1995)**	**Anderson (1999)**	**Mitleton-Kelly (2003)**	**Benbya & McKelvey (2006)**
Large number of components/agents	X			X
Variation and diversity				X
Connectivity and interdependence and interactions		X	X	X

Feedback		X	X	
Unpredictability and nonlinearity	X	X		X
Far-from equilibrium/edge of chaos	X		X	
Emergence / Self-organization / Strange Attractors	X	X	X	X
Space of possibilities / adaptation to environment (context) / learning			X	X
Historicity and path-dependence			X	X
Co-evolution			X	
Multidimensional / Scale free / Fractal	X			

It is common to use analogies in qualitative studies to explain complex and abstract processes and avoid long and monotonous texts. In this section, I am using a meal analogy *(in Italic)* at the end of each concept/feature of complex systems in order to explicate the process of the emergence of a complex adaptive system. A complex system is like a meal; it needs various ingredients; follows certain processes; taste and smell are emergent properties; etc. Readers should keep in mind that my analogy, like every analogy, is not the actual thing itself and has limitations; on the other hand, it is useful.

1. Large number of components: For a system to be considered as a complex system, there has to be multiple components (or agents) interacting with each other. These agents have different vectors or agendas and they try to pull system accordingly.

As Thietart and Forgues (1995, p.25) state: "Proposition 1. Organizations are potentially chaotic[5]. 1a. The greater the number of counteracting forces in an organization, the higher the likelihood of encountering chaos. 1b. The larger the number of forces with different patterns, the higher the likelihood of encountering chaos" For an outsider, it is a messy, chaotic bunch, that does not have a purpose or make sense. Some examples of such systems are the immune system, nerve system, brain, slime mold, ant colonies, and markets –all have countless agents operating or working for themselves, without knowing the big picture. Unfortunately, there is not a number in the literature to argue that 'this amount is sufficient to have complex system' (Mitchell, 2009). Different systems have different number of agents. For instance, in a jazz band, 10 people might be enough to a complex melody emerge; on the other hand, for consciousness to emerge, the human brain needs 90 billion neurons.

Analogy: In order to prepare a meal I would need certain amount of ingredients.

---

[5] Chaos and complexity were often used reciprocally –such as the study cited here. They are both non-linear systems but they are not the same. According to Baranger (2000) the constituents of chaotic systems are not 'interdependent' and chaotic systems are not 'emergent' –two of the prominent features of complex systems. Also it has to be kept in mind that not every non-linear system is complex or chaotic. Complex and chaotic systems are subsets of non-linear systems.

2. Variation and diversity: A large number of components is necessary but not sufficient. If there is no variation and diversity among agents, if they are all the same, it would just be a predictable, linear system that consists of huge number of agents. For instance, gas molecules in a container are a chaotic system, not a complex one or a refrigerator is a linear simple system, not a complex one. "In each system, each agent is different from the others (diversity), and its performance depends on the other agents and the system itself, each of which can influence the other's behavior" (Benbya and McKelvey, 2006, p.18). This diversity at certain conditions results in an emergent property. A system's behavior cannot be reduced to a single agent's behavior because diversity and variety gives each agent a different role (or vector) (Holland, 1995). There is no single dominant vector in the system.

Analogy: We generally need more than one type of ingredient to prepare a meal unless the meal is going to be boiled eggs. Let's assume I am going to make Noah's pudding to serve at my dissertation defense to the committee members. According to the story, when Noah's Ark came to rest on Mount Ararat, Noah prepared this special dish with what was left in the ship's kitchen. I would need a variety of ingredients: wheat, rice, barley, chick peas, beans, sugar, dried fruits, and nuts.

3. Interaction / connectivity / interdependence: The problem with the container full of gas or the refrigerator is the lack of interaction (or very limited interaction) among agents –single gas molecules or different parts of refrigerator. For instance, there is a thermostat inside the refrigerator set to a value, if the measured value is higher, the cooler

starts working until the temperature decreases. It is simple, predictable, and linear –no room for surprises or changes. However, "Complex behaviour arises from the inter-relationship, interaction, and interconnectivity of elements within a system and between a system and its environment." (Mitleton-Kelly, 2003, p. 4). As in Axelrod and Cohen's (1999) complex system definition "(a system is complex when) …there are strong interactions among its elements, so that current events heavily influence the probabilities of many kinds of later events." The action of one agent has an impact on other agents and even on other systems. The impact does not have to be equal on others, some might be affected more –which makes sense because agents do not know the big picture; they are affected by their immediate neighbors and like in the Chinese telephone game the impact is disturbed by each agent; thus, a uniform impact on each agent almost never happens. It is like a domino effect[6], each agent is dependent to the slightest change –a single flick– in its neighbor. "Complex patterns can arise from the interaction of agents that follow relatively simple rules (Anderson, 1999, p.218)". However, very different from the domino effect, the results are unpredictable.

---

[6] A limitation of the analogy reveals itself here. Dominoes are very predictable and prepared by a central planner. However, here a different feature is highlighted. A single domino interacts with only one domino (or maybe a couple of dominoes): the one that hits it and the one that it hits.

Analogy: The interaction between these components happens when these ingredients are put into a pan and heated together. The tastes, the smells, and everything else merge into another yet it does not become a uniform paste; some ingredients preserve their individual existence such as single chick peas, rice, nuts, and dried fruits. The amount of one ingredient I am going to use depends on the amount of the other ingredients.

4. Feedback: Positive (amplifying effect) and negative (dampening affect) feedback loops are typical of nonlinear and complex systems and one of the main reasons of unpredictability. In addition, these feedback mechanisms or processes are the main reasons of that scholars cannot isolate a variable and study it isolated –which results in the sum of a system is greater than the sum of its parts. As Anderson (1999, p.218) puts it: "… complex systems resist simple reductionist analyses, because interconnections and feedback loops preclude holding some subsystems constant in order to study others in isolation." Due to transfer of energy or information among agents impacts lose their proportion. The strength of feedback process is often determined by the degree of the connectivity (Mitleton-Kelly, 2003, p. 16).

Analogy: In a complex system, feedback occurs between the components yet the ingredients are far from processing this information in this analogy. I, as the cook[7]*, on the other hand, regularly check what is going on inside pan. I smell it; check the consistency, color, even taste –if possible; add more ingredients; stir or stop stirring; increase or reduce the temperature, etc according to my observations.*

5. Unpredictability and nonlinearity: These feedback loops and nonlinear relationships, create a condition called sensitivity to initial conditions –which results in unpredictability. The butterfly effect –a butterfly in Amazon flaps its wings and causes a tornado in Texas– is the famous example of sensitivity to initial conditions. "… the behavior of complex processes can be quite sensitive to small differences in initial conditions, so that two entities with very similar initial states can follow radically divergent paths over time. (Anderson, 1999, 218)." For instance, meteorologist Lorenz was too lazy to type .506127 into the computer while he was working on a climate model trying to predict weather, so he typed .506 (Gleick, 1987, 16). The results were so

[7] Here is the limitation of analogy. A cook contradicts with the idea of complex system as the cook being the one and only central planner and the controller of the system. The system's behaviors could be reduced to the cook's behaviors. Complex adaptive systems have neither central planners nor controllers.

different and unexpected that he started to work on this strange event and became one of the founding fathers of chaos theory[8].

Analogy: Trying to imitate famous cooks, I do not follow a recipe. Deciding on the amounts is an eyeball estimate at best. In addition, sometimes I substitute an ingredient or two. For instance, I use molasses instead of sugar if I do not have any sugar left at home. The result is a different taste and consistency in my Noah's pudding each time. The end result is in general parameters, it is sweet and pudding but the rest is unpredictable.

---

[8] It has to be mentioned that some scholars such as Bennet or Freimuth attribute 'sensitivity to initial conditions' as a feature of chaotic systems, not complex systems.

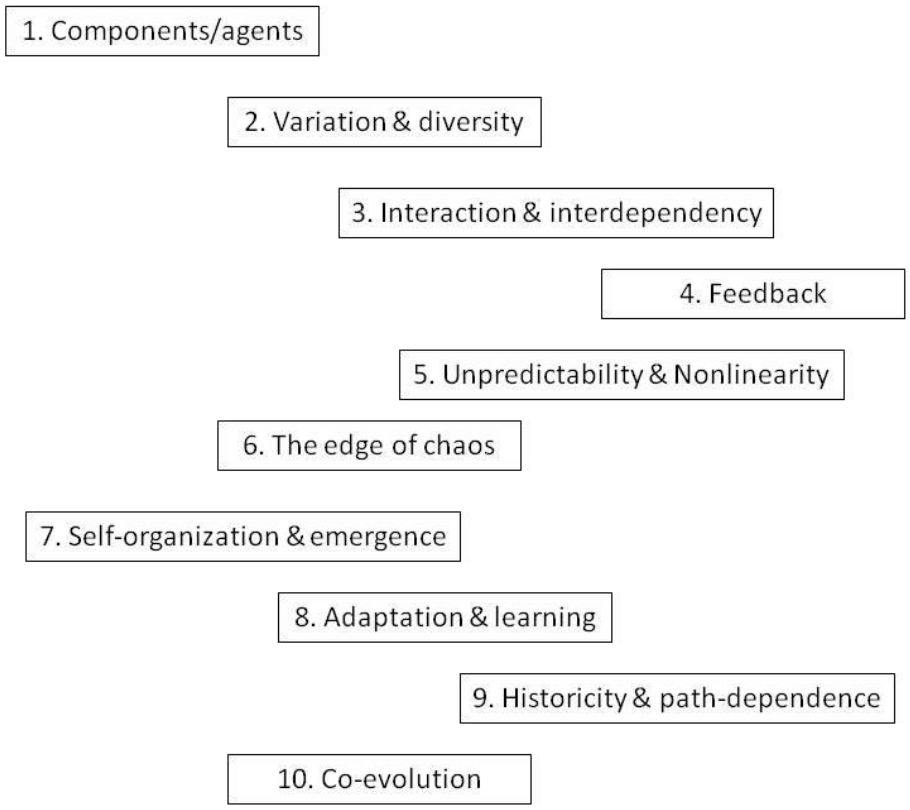


Figure 1– Basic concepts in complex adaptive systems

6. The edge of chaos/far from equilibrium: Systems do not stay in equilibrium forever. They react to internal and external (environmental) factors and equilibrium changes. They can exist of fluctuate between three states: stable, chaotic, and in between (Lewin, 1992; Thietart & Forgues, 1995; Anderson, 1999, Benbya & McKelvey, 2006). The 'in between' phase is actually when the system behaves 'complex'; however, different scholars named that phase different: Kauffman – melting zone, Cramer – critical complexity; McKelvey – region of emergent complexity (Benbya & McKelvey, 2006,

p.17). It is also known as the edge of chaos. This is where action takes place. In this zone, according to Mitleton-Kelly (2003, p. 10) "open systems exchange energy, matter, or information with their environment and which when pushed 'far-from-equilibrium' create new structures and order." Here higher levels of mutation and experimentation happen, which could become critical in a system's resistance or response to external threats (Pascale, Milleman, Gioja, 2000). Being away from equilibrium gives the system a chance to come up with a better configuration that increases the likelihood of its survival.

Analogy: In order for a complex system to emerge, certain environmental and structural criteria have to be met. In cooking, it is mostly the heating, the duration for the heat exposure, and the structure of the ingredients (cut in small pieces or grated). I use medium fire until it reaches a certain consistency.

7. Emergence, self-organization, and strange attractors: When the system receives energy, matter, or information, it absorbs until it reaches the critical point – which is the edge of chaos. At this point excess energy, matter or information generates something –a form, pattern, behavior, structure, etc. This is called emergence or self-organization. The emergent structure is neither planned nor predicted. As Anderson (1999, p.218) puts it "… complex systems tend to exhibit "self-organizing" behavior; starting in a random state, they usually evolve toward order instead of disorder." This does not contradict the second law of thermodynamics because of the excess energy (or information) the system received. Benbya and McKelvey (2006, p.16) summarizes this

happening referring to Kauffman, Cramer, and McKelvey: “In other words, new behavior patterns appear as consequences of agent interaction. No single program or agent completely determines the system’s behavior, despite the fact that each of the heterogeneous agents holds some common schemata. These systems self-organize when they find themselves in the “region of emergent complexity” at the “edge of chaos” (Cramer, 1993; Kauffman, 1995; McKelvey, 1999).” Each agent contributes to the emergent property differently; thus, it is unpredictable.

These forms, patterns, structures emerge around the excess energy, matter, or information –strange attractors(Anderson, 1999). “When in a chaotic state, organizations are ‘attracted’ to an identifiable configuration. a. When in a chaotic state, organizations are more likely to adopt a specific configuration than a deterministically “random” pattern. b. The greater the openness of an organization to its environment, the more likely is the ‘attraction’ by the organization to a given configuration (Thietart & Forgues, 1995, p.26).” A new order (equilibrium) is reached. In human systems generally it creates irreversible structures or relationships (Mitleton-Kelly, 2003). For instance, the idea of ‘minimum wage’ or ‘school’ as an educational institution are irreversible structures that have emerged in our civizilation.

Analogy: After enough stirring with the right temperature, the ingredients reach a critical point, and the right taste emerges. Noah’s pudding is ready to serve.

8. Space of possibilities / adaptation to environment (context): The emergent property, although it is a new equilibrium, is an adaptation. The system cannot continue as it was and through generating new patterns, forms, behaviors, relationships, and structures it adapts to the new conditions/environment. Just one strategy, one kind of agent is not desired, even though the basic economics (which relies on linear equations) or biology tells us to find the optimum to maximize, because when the conditions change that strategy or agent might not be optimal or suitable (Pascale, Millemann, Gioja, 2000; Mitleton-Kelly, 2003). This will result in annihilation. Thus systems do not work 'optimally' and instead try to have diversity and variation which builds in resilience. For instance, the immune system has multiple mechanisms, not one, to respond to pathogens. Or, companies invest in R&D or training to be available to respond to changing market conditions. McKelvey (2001) defines this process as 'adaptive tension.' If systems do not explore these 'space of possibilities' they become fragile.

The natural laws for molecular systems or DNA in organic systems, or consciousness, or rules or relationships in human systems are actually schemas for the actions of agents –their actions are bound to these schemas. "The existence of these shared schemas, together with the agents' individual schemas (diversity), opens up the possibility of changes to these rules, or in other words, evolution and learning (Benbya & McKelvey, 2006, p.19)." These schemas can change, that change is adaptation, that change is learning –crucial to survival.

Analogy: The fantastic thing about Noah's pudding is that there is not one recipe. Noah's Ark is acknowledged in many cultures and the recipe is adapted to local resources and tastes. People use different dried fruits, some add cinnamon, some add rose water, some use pecans or almonds instead of walnuts, etc.

9. Historicity and path-dependence: To explain these concepts Arthur's (1994) 'increasing returns' concept must be explained. Simply, increasing returns are positive feedbacks in the system. General economic theory envisages negative feedback and argues that systems (market) will come to equilibrium at the optimum point yet in reality it does not have to. For instance, the QWERTY keyboard was introduced to slow down typists because the typewriters got jammed when typed fast. People learned how to type on the QWERTY keyboard, demanded more QWERTY keyboards, more QWERTY keyboards become available in the market and used more, more people learned how to type … And the cycle continues. Today we still use the QWERTY keyboard; however, today we do not have a jamming problem. We could use a more efficient keyboard but it does not happen because of the latent cultural knowledge we have in using the QWERTY keyboard. Arthur (1994) calls this 'lock-in'. Many companies have inefficient workflows but they do not change it because it has been like that forever. It does not change until an external force threatens the system. Past events affect future events. There is a sequence of events that limits the possible actions in each step until it reaches inertia or lock-in (Schreyögg, Sydow, & Holtmann, 2011). This happens due to the interdependency. The

end result is not predicted and might endanger the survival of the system. Complexity theory is not used to predict or manipulate but to explain past, it works retrospectively[9].

Analogy: During the preparation, let's assume I accidentally put more sugar than needed. Whatever I do, at the end it will taste sweeter than it is supposed to be. I could add more cinnamon to break the sweet taste but it can work only so much. The end result/taste is bound to previous actions.

10. Co-evolution: The adaptation, and thus the evolution, is not alone but together –including the environment which is a collection of systems with other agents. Every agent in the system is interconnected to each other; hence, a change in one creates a change in another; that one in another. It continues like that until every agent repositions (changes or mutates) themselves. In Stacey's (2003, p.2) definition of complex adaptive system this feature becomes clearer: as "A complex adaptive system consists of a large number of agents, each of which behaves according to some set of rules. These rules require the agents to adjust their behavior to that of other agents. In other words, agents interact with, and adapt to, each other." If an agent or a group of agents cannot adapt, they do not survive; they become extinct or die or leave the system.

---

[9] It should be remembered that short-term prediction is possible. For instance, tomorrow's weather can be predicted (to a degree) but not next week's weather.

Analogy: Noah's pudding is a system that exists among other systems. It is a dessert, in a dinner meals system, it exists with the entrée and the main dish. For instance, in Turkey, it is not served when the entrée is barley soup and the main dish is seafood because it has already barley in it, it becomes too much and its taste is not considered suitable for seafood. However, for instance, it co-exists with lentil soup and meatballs because it complements the tastes of these meals.

11. Multidimensional / fractal: There is not a single unit of analysis in complex systems because of fractal structure. "When in a chaotic state, organizations, generally, have a fractal form. a. When in a chaotic state, similar structure patterns are found at the organizational, unit, group and individual levels. b. When in a chaotic state, similar process patterns are found at the organizational, unit, group and individual levels" (Thietart & Forgues, 1995, p.27). Thus the effects are contagious and also similar. For instance, an individual's decision to sell stocks in the market might be represented at the market level –which means everybody is selling that same stock. (The emergent property is the decline in that stock's price). Moreover it is contagious among different type of systems. "Complex systems are multidimensional, and all the dimensions interact and influence each other. In a human context the social, cultural, technical, economic and global dimensions may impinge upon and influence each other" (Mitleton-Kelly, 2003, p.5). For instance, the Internet, which is basically a military technology, has changed so much in our economic, educational, social, etc. life.

In conclusion, these concepts and the Figure 2 (below) help us to understand complex adaptive systems better. The nonlinear interactions (including feedback processes) among various agents result in an emergent property in an open system. The agents and the emergent property impact each other. Also there is interaction among other systems. For instance, on a local level, teenagers interact with each other through their cell phones and use shortenings in their SMSs. In turn, a global SMS language emerges. The lower level interactions are the cause of unpredictable higher level order (SMS grammar if it can be said) that emerged. The higher level order then dictates the local interactions –the teenagers who want to communicate with others use the SMS language. It could also spread into other systems, for instance into instant chatting environment (MSN, Skype, etc.).

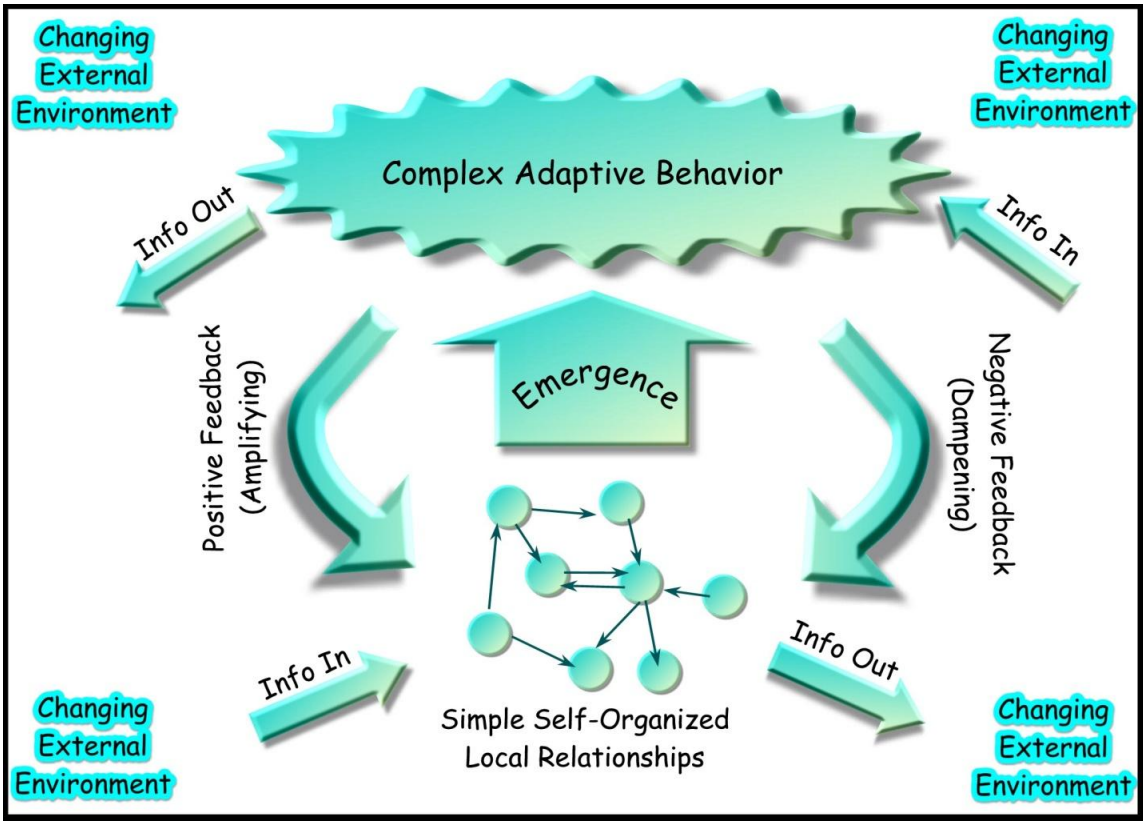


Figure 2 – Complex Adaptive System

These concepts can be found in every complex system including organizations and scientific collaborations; and thus, constitute the backbone of complexity theory. Complex behavior is explained through these concepts. Scholars conducted many studies using these ideas. Arthur's study (2009) is about how technology develops and evolves. Cilliers (1998) writes on complexity theory and postmodernism. Salem (2009) explains the applications of complexity theory in human communication. Wagner (2008) investigates the relationship between scientific collaborations, developing countries, and complexity theory and makes science and technology policy recommendations to

developing countries. Sandole (2006) applies complexity theory to conflict resolution whereas Clemens (2006) uses it to explain the ethnic conflict in Post-Soviet-Eurasia. Hoffman (2006) investigates the ozone depletion with complex adaptive systems theory. In a compilation edited by Bogg and Geyer (2007) complexity theory and its reflections in sustainability, education, health, international relations and development, philosophy, politics and policy, and social theory were examined.

In a nutshell, according to complexity theory, the relationship between the units is nonlinear and these systems cannot be reduced to its parts and units cannot be isolated. Feedback loops cause unpredictability in the long term whereas some patterns or forms might be observed for short periods.

Emergence

The third body of literature relevant to this study is emergence, which is actually a feature of complex adaptive systems (CAS) but not the central concept. Emergence is the process whereby the global behavior of a system results from the actions and interactions of agents (Sawyer, 2005: 2).

Sawyer (2005) explains the development of social system theory in three waves. First wave is Parson's structural functionalism; second wave is general systems theory from 1960s to 1980s; and third wave is the complex dynamical systems theory developed in 1990s at Santa Fe Institute. Third wave is what we call complexity theory or theory of

complex adaptive systems today. Waldrop (1992) describes the establishment of Santa Fe Institute and the development of complexity theory in detail in his popular study.

However, according to Sawyer (2005) the emergence concept is crucial in understanding social systems and it could be helpful to position it in the center of complexity studies (p.21-6). Without it, explaining social systems is impaired. Natural systems could be explained easily because they are less open, relatively easily quantified, and not subjective. Due to its complexity, language, for instance, is an emergent structure. Salem (2009) describes the process of information and communication as a socially emergent process as well. According to Sawyer (2005) "relatively simple higher-level order 'emerges' from relatively complex lower-level processes"(p.3). His example is language shift; lower level consists of individual speakers, whereas language is the higher level. The rules of language are understandable (grammar, lexicon and else) yet the relation or communication among individual speakers and how they come up with new words or phrases cannot be known. Language shift or slang is an emergent property. When this principle is applied for instance to scientific collaborations, a simple research network emerges from the complex relationships among researchers.

Sawyer (2005) believes that "the most important missing element is the sophistication of human symbolic communication" (p.22-3) in explaining social emergence. Given the complexity and impact of symbolic communication on social life, it is an important aspect to disregard. Complexity theory focuses on interactions not

variables. The symbolic communication is where interaction happens; therefore, it becomes essential to understand and explain emergent properties.

In summary, this chapter provides the literature review on scientific collaborations and complex adaptive systems. The studies on scientific collaborations are examined in two subsections: scientometrics and case studies. Scientometrics studies are not very useful in explaining the dynamics of scientific collaborations as they deal with the scholarly outcome. Case studies, on the other hand, do not provide generalizable findings. In regards to understand the emergence of a scientific collaboration, these shortcomings can be overcome by a different approach: complex adaptive systems perspective. Thus, the basic features and principles of such systems are explained along with the emergence concept in the rest of the chapter in order to develop a framework to assess virtual scientific collaborations.

Chapter 3

Background for DataONE

In this chapter the background information which resulted in DataONE is provided. DataONE, as a system, exists with other systems and is influenced by them. Therefore, the history of the environment that DataONE exists is important in understanding how it emerged.

About DataONE

DataONE (the Data Observation Network for Earth) is focused on enabling data-intensive biological and environmental research through cyberinfrastructure. Funded by the National Science Foundation (NSF), DataONE is a multi-institutional, multinational, and interdisciplinary collaboration working on developing an organizational structure that will support the full information lifecycle of biological, ecological, and environmental data and tools to be used by researchers, educators and the public at large. According to the official website, it "will ensure the preservation and access to multi-scale, multi-discipline, and multi-national science data" (DataONE, 2009). It is not a surprise that a project that addresses data issues would emerge now, because we are now in data-intensive research era (Hey, Tansley, & Tole, 2009) – more details are provided below.

NSF the Office of Cyberinfrastructure

The NSF's support to cyberinfrastructure dates back to 1960's in which campus-based computational facilities were funded. (NSF, 2006, p. 30). In 1980's the NSF initiated Supercomputer Centers program (NSF Office of Cyberinfrastructure). Simultaneously "academic-based networking activities also flourished" (NSF, 2006, p. 30) which led to an increase in the efficiency of researchers and educators. In 1991 Congress passed the High Performance Computing and Communications (HPCC) Act to use them in forecasting severe weather events, cancer gene research, predicting new superconductors, aerospace vehicle design, earth biosphere research, simulating and visualizing air pollution, energy conservation and turbulent combustion, and microelectronics design and packaging (Computer Science and Telecommunications Board, 1995, p. 89). "These HPCC projects joined scientists and engineers, computer scientists and state-of-the-art cyberinfrastructure technologies to tackle important problems in science and engineering whose solution could be advanced by applying cyberinfrastructure techniques and resources" (NSF, 2006, p. 30). Cyberinfrastructure became an important part of scientific activity and in 2001 the NSF established an Advisory Committee for Cyberinfrastructure. Since then the Office of Cyberinfrastructure coordinates the efforts where cyberinfrastructure is involved in tackling 'grand challenges'.

> "All of these developments are part of a revolutionary new approach to scientific discovery in which advanced computational facilities (e.g.,

> data systems, computing hardware, high speed networks) and instruments (e.g., telescopes, sensor networks, sequencers) are coupled to the development of quantifiable models, algorithms, software and other tools and services to provide unique insights into complex problems in science and engineering" (NSF Office of Cyberinfrastructure).

The Fourth Paradigm: Data-intensive scientific discovery

Some scholars call this era the fourth paradigm or the data-intensive scientific discovery (Hey, Tansley, & Tole, 2009). The previous paradigms were experimental, theoretical, and computational –each being the core of the scientific discovery. In the final paradigm, digital data is the core. "All of the science literature is online, all of the science data is online, and they interoperate with each other" (Hey, Tansley, & Tole, 2009) is what the fourth-paradigm envisions. The advancements in information and communication technologies led researchers to use more computational simulation and modeling techniques and remote data collection which resulted in increases in the amount of data collected, used, re-used, and preserved (NSF, 2007). When data is deposited digitally, it can be shared, integrated into bigger data sets, re-analyzed, and preserved much easily compared to analog data. Therefore, the results can be verified by other researchers and also replication studies can be conducted to train future generations of researchers; interdisciplinary research is fostered by integrating different datasets; data integrity is achieved through preservation; data collection costs are reduced (ESF, 2007; Sieber, 1991; ICPSR, 2009).

Sustainable Digital Data Preservation and Access Network Partners (DataNet)

The NSF responded to the change in the research paradigm. Based on NSF's cyberinfrastructure vision (2006) for enable accurately deposited, well preserved, and easily accessible data by specialists and non-specialists; the NSF's DataNet solicitation (2008) had addressed the need for approaches for data-intensive scientific and engineering research by integrating library and information sciences, cyberinfrastructure, computer sciences, and domain science. Thus, collaborations will:

- "provide reliable digital preservation, access, integration, and analysis capabilities for science and/or engineering data over a decades-long timeline;
- continuously anticipate and adapt to changes in technologies and in user needs and expectations;
- engage at the frontiers of computer and information science and cyberinfrastructure with research and development to drive the leading edge forward; and
- serve as component elements of an interoperable data preservation and access network" (NSF, 2008, p. 2).

In addition, the NSF has recently added a new component to the grant proposals they receive. In Fall 2010, the NSF announced that every proposal that is submitted to the

NSF should have a data management plan (NSF Press Release, 2010), which indicates the importance given by the NSF to data issues.

The first two DataNet projects that have received funding are Data Conservancy and DataOne.

Data Conservancy

Data Conservancy is an effort to ensure preservation and curation of scientific and engineering data led by Johns Hopkins University. The subawardees include Cornell University, Duraspace, Woods Hole Marine Biological Laboratory, National Center for Atmospheric Research, National Snow and Ice Data Center, Portico, Tessella, University of California Los Angeles, and University of Illinois. "Through a well-defined management policy, DC will provide the foundation for a diverse, international team to iteratively develop, deploy, and evaluate infrastructure in a manner that combines rapid implementation with research, all with continual progress toward sustainability" (NSF DataNet, 2010). Data practices and curation for astronomy, biodiversity, earth sciences, and social sciences will be studied by scholars in this project.

The Data Observation Network for Earth (DataONE)

The key players in DataONE are the University of New Mexico, the partnership between Oak Ridge National Laboratories, and the National Center for Ecological Analysis and Synthesis. They established DataONE to tackle two problems. The first one

is the environmental problems the world has been facing especially in the last century –the increase in human population and its impact on land-based ecosystems, oceans, and ice sheets, the increase in the surface warming, deforestation, pollution, and the ozone hole are just to name a few of this complex transdisciplinary problem. These problems are so intertwined with each other, it actually is one big complex adaptive system that has become a complex adaptive problem. Yet these problems are studied by different disciplines and even though the problem is one, until recently –two decades at most– they had belonged to different domains of scholarly interest. As a result, there has not been an integrated body of literature on the topic –again until recently.

This brings us to the second problem, which is directly related to not having an integrated body of literature: the lack of integrated data. This problem is understandable, as the need to combine the efforts of different scientists and different disciplines has been realized recently. In addition, there are some data challenges such as data, scattered data sources, data deluge, poor data practice, and data longevity.

Data loss occurs when a natural disaster such as fire or flood damages the facility where the data is stored. Another example of data loss happens when the format of data becomes obsolete. The technology chances so fast that older versions of datasets become inaccessible. Finally, the owner of the data gets retired or deceased and her/his data becomes inaccessible. Scattered data sources is another problem that needs to be dealt with. Unless the data are integrated, bigger datasets cannot be created. In addition, repetitive data collection can occur, which results in increasing costs. Data deluge is the

problem of standards in creating metadata. Different data sources describe their data in different formats that cannot be converted to other formats easily. Again data integration becomes problematic. Data longevity is related to the media that the data is stored. Every media (disk, tape, CD, DVD, etc.) has a life span. They need to be transferred to a newer media when their life span is over, which requires personnel and equipment. It is not common to have researchers to use their limited resources to try to preserve their old data instead of conducting new research that would make them answer new questions, bring them fame, and more resources. Furthermore, scientists are not aware of the data issues, they do not have the resources or the skills to deal with the data issues; therefore, they have poor data practices.

To sum up, DataONE, through dealing with the data problems in environmental sciences, supports the environmental efforts. DataONE aims (1) to provide coordinated access to the current databases (such as Ecological Society for America, National Biological Information Infrastructure, Long Term Ecological Research Network and others) using the available cyberinfrastructure; (2) to create a new global cyberinfrastructure that contains both biological and environmental data coming from different resources (research networks, environmental observatories, individual scientists, and citizen scientists); and (3) to change the science culture and institutions through the new cyberinfrastructure practices by providing educations and trainings, engaging citizens in science, and building global communities of practice.

DataONE is highly collaborative both in terms of institutions and disciplinary interests involved. The collaboration has two levels of participation: coordinating nodes (the initial ones are The University of New Mexico, The partnership between University of Tennessee and Oak Ridge National Laboratories, and the National Center for Ecological Analysis and Synthesis) and member nodes (the rest). Member nodes are responsible for the storage of data, whereas coordinating nodes provide some data storage and importantly provide critical network-wide services such as a registration service, global metadata index spanning, information security, replication services, and discovery services.

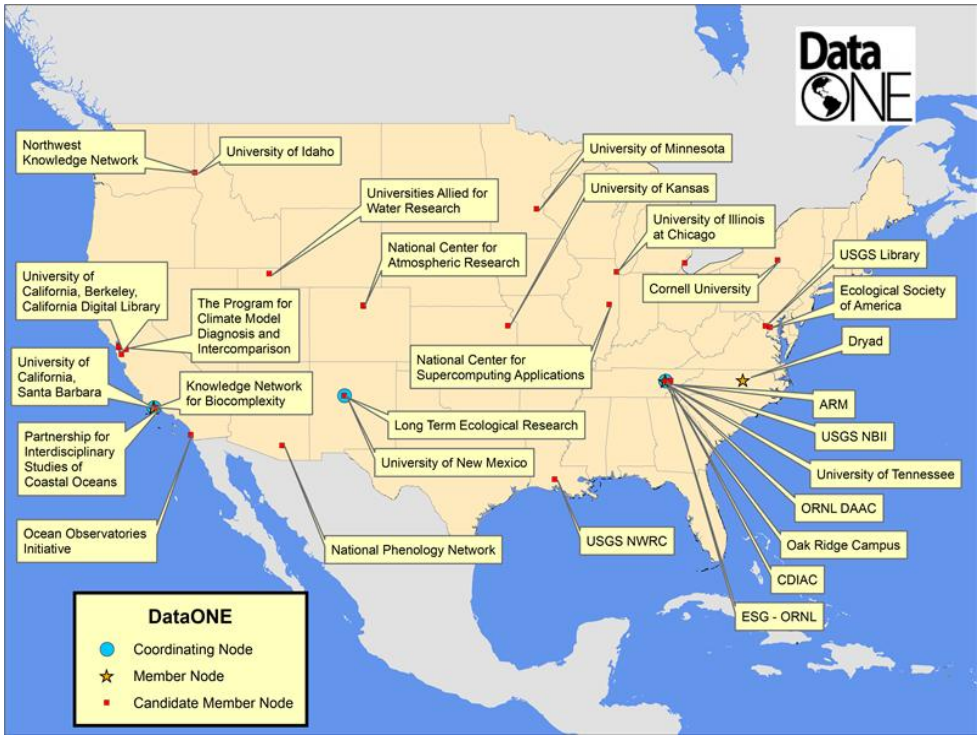


Figure 3 – Coordinating nodes, member nodes, and candidate member nodes in the U.S. as of February 2011

Different types of institutions (universities, research centers, synthesizing centers, libraries, etc.) have joined resources to process the data coming from different disciplines and locations so that data can become accessible to the interested parties (scientists, land-managers, policy makers, students, educators, and the public) and also be stored for future use. The types of institutions that are interested in DataONE activities are:

1. Academic institutions from the U.S. (including three EPSCoR [The Experimental Program to Stimulate Competitive Research] states—

Tennessee, Kansas, and New Mexico) and the United Kingdom (i.e., Edinburgh, Manchester, Southampton);

2. Research networks (e.g., Long Term Ecological Research Network, Consortium of Universities for the Advancement of Hydrologic Science Inc. [CUAHSI], Taiwan Ecological Research Network, South African Environmental Research Network [SAEON]);
3. Environmental observatories (e.g., The National Ecological Observatory Network [NEON], USA-National Phenology Network, Ocean Observatory Initiative, South African Environmental Observatory Network);
4. NSF- and government-funded synthesis (i.e., the National Center for Ecological Analysis and Synthesis [NCEAS], the National Evolutionary Synthesis Center [NESCent], Atlas of Living Australia) and supercomputer centers/networks (Oak Ridge National Laboratories [ORNL], National Center for Supercomputing Applications [NCSA], and TeraGrid);
5. Governmental organizations (e.g., U.S. Geological Survey [USGS], the National Aeronautics and Space Administration [NASA], Environmental Protection Agency [EPA]);
6. Academic libraries (e.g., University of California Digital Library, University of Tennessee, and University of Illinois-Chicago libraries, which are active in the digital library community and are members of the Coalition for Networked

Information, the Digital Library Federation, and the Association of Research Libraries);

7. International organizations (e.g., Global Biodiversity Information Facility, Inter American Biodiversity Information Network, Biodiversity Information Standards);
8. Numerous large data and metadata archives (e.g., USGS-National Biological Information Infrastructure, ORNL Distributed Active Archive Center for Biogeochemical Dynamics, World Data Center for Biodiversity and Ecology, Knowledge Network for Biocomplexity);
9. Professional societies (e.g., Ecological Society of America, Natural Science Collections Alliance);
10. NGOs (e.g., The Keystone Center); and
11. The commercial sector (e.g., Amazon, Battelle Ventures, IBM, Intel) (DataONE Proposal, 2009).

As for the disciplinary interests, by definition there are at least three disciplines involved in the project: computer science, library and information science, and earth/environmental sciences, yet at least two of these groups are highly diversified. Earth sciences consist of geologists, geophysicists, oceanographers, soil scientists, hydrologists, climatologists, ecologists, and also biologists. Library and information sciences have at least two different focuses on preservation and information systems design and access. It would not be a surprise to see that more disciplinary interests are included.

DataONE's structure can be examined in two ways: organizational structure and process structure. Organizationally speaking, DataONE has two big bodies that do the job. First is the cyberinfrastructure team, which consists of six working groups (WG). Each WG works on a different component of the cyberinfrastructure that DataONE is going to operate on. Second is the community engagement and outreach team, which consists of five WGs. WGs deal with the social side of data preservation and sharing issues, the education needs of DataONE users, and the sustainability of the project. DataONE is managed by a leadership team. In addition, an External Advisory Board provides guidance. The organization chart is provided below (Figure 2):

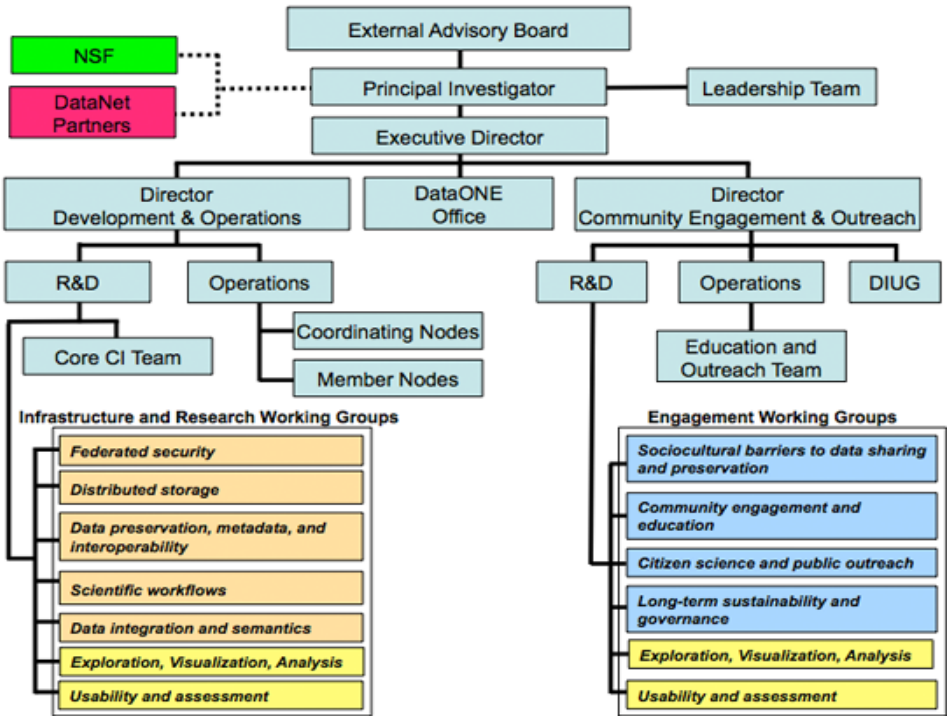


Figure 4 – Organization chart for DataONE as of February 2011

Processwise, as DataONE's main objective is to provide a cyberinfrastructure to deal with scientific data issues, its activities are shaped around data lifecycle which was developed[10] by DataONE members (Figure 5). The different stages of data, requires the involvement of different stakeholders and different activities. For instance, in the collect stage the researchers, the field workers, or the remote sensors collect data. Data assurance, on the other hand, can only be performed by scientists. Data needs to be audited, cleaned, and organized. The describe stage could be the job of a scientist, a data curator, or a librarian. Here data needs to be tagged. To deposit and preserve, equipment and technology are needed. Financing these requires policy-makers to be involved as well. A researcher who uses models and simulations is included in the later stages of the data lifecycle. The data needs to be discovered and integrated to other datasets before it is analyzed. A very brief description of the stages of the data lifecycle above provides various tasks and stakeholders. DataONE provides the necessary cyberinfrastructure to stakeholders so that they can perform data related tasks. In addition, DataONE informs, convinces, and provides training to stakeholders so that they take action to deal with data issues.

---

[10] There had been different data lifecycles that were used in DataONE at the earlier stages. The final version was submitted to the NSF on February 2011 and became the official DataONE data lifecycle.

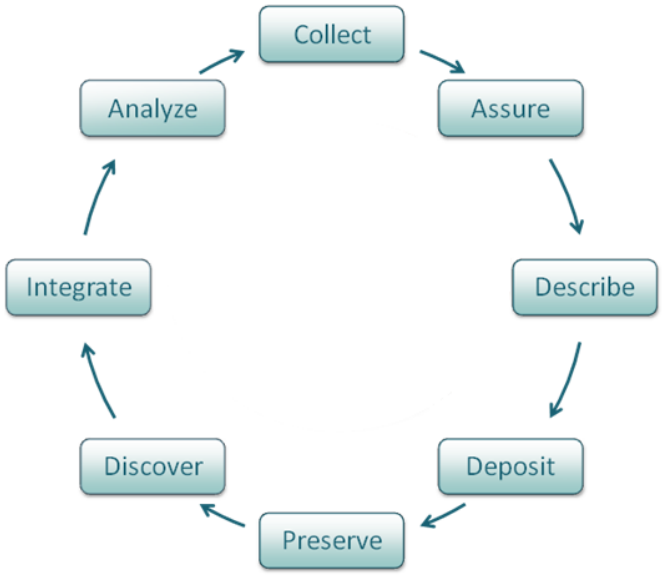


Figure 5 – Data lifecycle as adapted by DataONE as of February 2011

In summary, DataONE, is a multidisciplinary, multi-institutional, multinational virtual scientific collaboration that addresses that data problems in earth sciences. It deals with both technical (cyberinfrastructure) and social (community engagement) sides of data issues through data lifecycle.

Chapter 4

Methods

The main research question guiding this study is: "How can emergence of DataONE –a multidisciplinary, multinational, and multi-institutional scientific collaboration– be explored from a complex adaptive systems perspective?" To answer this question the methodology used in this study is case study. The data is generated through multiple methods including interviews, observations, and surveys. Therefore, in this chapter, first, the case study research method is introduced and reasons for selecting DataONE are discussed. Second, the process of data generation through semi-structured interviews, naturalistic observations, and online surveys is explained. A copy of the interview guide and survey questions are provided in Appendix A and B. Finally, data integration and evaluative criteria are presented.

Case study

To capture the complex nature of the subject, the case study method is employed as this method provides flexibility and rich data. The case study is a "research strategy which focuses on understanding the dynamics present within single settings" (Eisenhardt, 1989, p. 534). It involves an in-depth, longitudinal examination of a single instance, event, or episode (Yin, 1984). Instead of having a generalizable truth, the researcher achieves a deeply focused understanding of how and why that instance, event, or episode happened as it did. From such understanding new areas to focus on might emerge and

lead research to new directions. It is very suitable not only to develop hypotheses but also to develop theories; thus, it is used in grounded theory research. Case studies could be used to test hypotheses in the real world too. It has been used in many scientific disciplines, especially social science, psychology, anthropology, business and ecology.

A key strength of case study is answering the ‘why’ question. For instance, a bibliometric analysis might show the increase in co-authored publications but a couple of case studies might answer ‘why’ scholars are collaborating and reveal the dynamics of co-authorship.

Another key strength of the case study method involves using multiple sources and techniques in the data gathering process. Case studies can be based on any mix of qualitative and quantitative evidence. The researcher determines in advance what evidence to gather and what analysis techniques to use with the data to answer the research questions (Yin, 1984; Eisenhardt, 1989). Planning and design is very important in case studies, otherwise the vast and rich data generated and collected become an obstacle in understanding the phenomenon being investigated. The researcher should make sure that the data generated and collected is relevant, coherent, and concise. Case studies do not have standard procedures for design and reporting methods like in quantitative studies; it is up to the researcher. Therefore, the researcher becomes an important instrument in case study research by his/her approach to the topic and interpretation of the data.

The dynamics of emergence need to be unearthed and this requires deep understanding of the phenomenon. The case study method provides such opportunity because the raison d'être of the case study is deep understanding of a single phenomena through delicate and detailed data collection process (Yin, 1984). In addition, the employment of multiple methods helps the researcher to understand different dimensions of the phenomenon and overcome the limitations of single method.

Rationale for qualitative inquiry

Although a case study employs both qualitative and quantitative methods to collect data, it is generally viewed as a qualitative method and quantitative findings are used to support qualitative findings. This view is favored in this study as well. Here basic features of qualitative research and their reflections in the proposed study are compared to reveal the fit. First and foremost, the aim of the qualitative study is to understand and explain –mostly the mental constructs of the group studied. The aim of this dissertation is to understand how scientific collaborations (DataONE) emerge and explain the role of communication and information behaviors in the process. This process is happening in the mental constructs of the members of the collaborations and it could only be studied through qualitative inquiry. Second, a qualitative study is not interested in manipulation and control, neither is this study. The findings will hopefully increase understanding and contribute to the literature. Third, in a qualitative study data is local, specific, and time bound –which indeed what a case study is. Four, context means everything in qualitative studies and almost everything in case studies. Five, data is generated through the inquirer

and the inquired, which gives an important role to the researcher. Above, the importance of the researcher as an instrument in data collecting and analyzing has already been mentioned. Finally, theories are inductive in qualitative studies. In case studies, cases are selected on dimensions of a theory –here complexity theory.

Sampling

Information-oriented sampling[11] is used to select the case. Information-oriented sampling is selecting a case that has the potential to provide the richest data (Flyvbjerg, 2006, p. 229-30). The reason behind such logic is that an average case, which is selected through random sampling, most probably does not provide the richest data but the average data; however, extreme or atypical cases are filled with interactions among agents (compared to an average) and could provide better insights. The deeper causes of why things happen might be revealed through such cases. Such revelations might lead hypotheses that could be tested in future research and generalizable findings could be achieved. The downside is generalizable findings might not be possible, for the same reasons obviously. Moreover, even though case studies cover a narrow area, they provide more realistic responses to everyday problems than a purely statistical survey. However,

[11] Although the term 'sampling' brings 'random selection' to mind, here it is used as purposeful selection.

it has to be mentioned that some scholars, such as Flyvbjerg (2006), disagree with this argument and claims that generalizable findings might be possible.

Rationale for selecting DataONE

The selected case is DataONE (Observation Network for Earth). There are three reasons that DataONE is appropriate for study. First, DataONE was being formed at the time of the study. The National Science Foundation funded the project in August 2009. It is a great opportunity for a researcher to witness the emergence of a collaboration. The data is generated through interviews, observations, and surveys (details are explained below). Data is retrospective. As time goes by, people add and subtract emotions and thoughts, develop new positions towards the phenomenon. In previous studies the researchers were involved after the collaboration was underway. Collins (2004) on Laser Interferometer Gravitational Wave Observatory covered 100 years. The International Virtual Observatory Alliance started in 2002 but the research started five years after (Kertcher, 2009). Vertesi's study (2009) on Mars Rover and Saturn Cassini collaborations is a recent study but the collaborations went back decades. However, in the DataONE case, the fresh memories of the interviewees and other participants can provide more intact data.

Second, the interdisciplinary structure of the collaboration provides opportunities to observe emergence (here 'emergence' is in its complexity science meaning). Sawyer (2005) explains in what kind of systems emergence occurs:

> "Complexity theorists have discovered that emergence is more likely to be found in systems in which (1) many components interact in densely connected networks, (2) global system functions cannot be localized to any one subset of components but rather are distributed throughout the entire system, (3) the overall system cannot be decomposed into subsystems and these into smaller sub-systems in any meaningful fashion, (4) and the components interact using a complex and sophisticated language" (p.4-5).

The DataONE collaboration, as a complex emergent system, fits to all of the four criteria compiled by Sawyer.

(1) The interdisciplinarity of the phenomenon –environmental problems and data needs– requires frequent interactions among the agents who are dedicated to tackle it. This collaboration is not the kind of collaboration where the members do their own research and study, and meet to share their findings. Quite the opposite, members are creating this cyberinfrastructure all together with continuous communication and information flow.

(2 & 3) The objectives of DataONE cannot be localized to one group as the tasks are diversified yet interconnected to other. For instance, creating a platform for data sharing and preservation does not make sense if scientists do not have an interest in using it. On the other hand, if you have scientists interested in this, it would not be enough as the necessary tools and platform is missing. Thus, a look at the organization chart (provided earlier) and the objectives of DataONE is quite promising.

(4) In the data lifecycle (provided earlier), the need for interoperability, standards, and integration requires a complex cyberinfrastructure, which is the backbone of the collaboration and as a matter of fact that could be considered as the grammar book of a complex language that would help researchers from different disciplines to be able to communicate with each other regarding the environmental phenomena. Given the nice fit in four criteria, DataONE seems to be very promising as an excellent organization to observe complex emergent behavior as a system.

Finally, the sample is accessible. The researcher is a graduate student in one of the coordinating nodes (explained later) and personally knows two of the members in the leadership team. Through their reference, the researcher acquired access to the rest of the members. In addition, the researcher lives in the town where one of the coordinating nodes and several key personnel are located, which made accessing them convenient. To sum up, due to the emerging stage of the collaboration, the interdisciplinary structure of the collaboration, and the accessibility of the participants; DataONE has been selected as the case for the study.

Data collection

It was mentioned that conducting a case study employs multiple methods in order to obtain as much data as possible. Therefore, the research questions are explored by employing both qualitative and quantitative methods. Through semi-structured interviews, naturalistic observations, and surveys with co-principal investigators and co-

investigators, who have more knowledge about the functioning of the collaboration, data was generated. The data for this study was collected between February 2010 and March 2011.

1. Semi-structured interviews

The semi-structured interview method is applied for this component as interviews "can take us into the mental world of individual, to glimpse the categories and logic by which he or she sees the world" (McCracken, 1998, p.9). This "sharply focused, rapid, highly intensive" (p.7) method is the appropriate data generation method as it allows the researcher to understand the mental framework of the participants through free conversation.

By the time that the study started, according to the Appendix A4 of the grant document, there were thirty-five key members in the collaboration –four co-principal investigators (Co-PIs) and thirty-one co-investigators (Co-Is). The researcher would have liked to have as many interviews as possible; however, due to time and budget constraints interviews were conducted with the members of the Leadership Team. The leadership team consists of 17 people, four of whom are the Executive Team (PI, Executive Director, Director of Development and Operations, and Director of Community Engagement & Outreach). This team encompassing the Co-PIs and representatives from key institutions and focal areas, "confers weekly with DataONE key personnel to provide advice and guidance with respect to strategic organizational directions (including routine risk assessment), project implementation, collaborative opportunities and community

engagement, personnel, and other matters that are central to project success" (DataONE, 2009); therefore, due to their knowledge, expertise, and opportunity to see the big picture, they provided rich data and thick descriptions about the dynamics of collaboration.

Before initiating the interviews, the researcher conducted two pilot interviews to test the interview guide: one with the co-lead of SocioCultural Working Group and one with the project postdoctoral associate. Although both of the interviewees were not in the leadership team, they both have extensive knowledge of the project –as one is the co-lead of working group and the other being full time employee of the project working for two working group co-leads. After the transcription and analysis of the interviews, the interview guide was fine tuned and the interviews started.

Due to time constraints of the members of the leadership team only 13 of the 17 people were able to participate in the interviews. This reflects 76% of the leadership team. The researcher was able to conduct interviews with everyone in the executive team and also with the first five originators/founding fathers of the project. Redundancy was reached around the 10th interview, so the researcher was able to pursue some emerging themes in the remaining interviews.

Semi-structured interviews focused on the dynamics of the emergence of a collaboration by asking 'how' and 'why' questions. They were conducted as informal conversations, which were guided by a discussion guide with several open-ended

questions. The first few questions, which are called ‘grand tour questions’ (McCracken, 1988, p. 34) such as demographics, education, and affiliation, were designed to make the respondents feel more familiar with the interviewer and more comfortable in discussion. The subsequent questions asked the respondents to express their thoughts and feelings toward DataONE.

The researcher wanted to explore complex adaptive behavior; therefore, the questions were designed to observe some of the basic features/themes of complex adaptive systems, which are emergence, complexity & interaction, and adaptation. In ‘emergence’ related questions, the researcher aimed to observe ‘emergent behavior’ (as explained in literature review above), bottom-up formation, and self organization. In ‘complexity & interaction’ related questions, the non-linear relationships and interactions among agents, and the counter-acting forces in the system (such as different institutional or agential goals) were the focus. In ‘adaptation’ related questions, the evolution of the collaboration over time due to the changes in the internal dynamics (such as addition or subtraction of a member) and the environment (a change in law, funding, public perception etc.) which are communicated through various feedback loops. The combination of themes helped to reveal the, complex behavior of the system. The interview guide is provided in the appendix.

Encouragement and relevance are crucial in interviews; for this purpose, the subsequent question was emerged from the last reply of the interviewee whenever possible, which resulted in asking questions in different order. This was also appropriate

to qualitative inquiry because it sees inquirer and inquired together and “findings are literally the creation of the process of the interaction between the two” (Guba, 1990, p.27). As for relevance, it is possible to be pulled out of the phenomenon of interest to another topic by the interviewee for various reasons. Such cases happened, the interview guide above served as framework to help the researcher to stay on track during the interview. However, the researcher took the advantage of flexible qualitative research design –that is to be on alert to realize serendipitous/emerging categories and ready to pursue them if needed.

The interviews lasted between 30 to 50 minutes. Six of them were conducted face to face at the interviewee’s office (4) or a coffee shop (2). The rest of them were Skype (7) or phone (2) interviews due to geographical and other constraints. The discussion guide was sent to the interviewees beforehand to save time. Interviews were audio recorded and verbatim transcribed for analyzing the data and quotes. After the transcription, the texts were sent back to the interviewees for member check and additional editing if desired. This process was crucial for two reasons: First, to avoid any mistakes during the transcription and be able to reflect interviewees’ thoughts correctly. Second, the interviewees can be identified easily due to the small number of people in the leadership team. With the editing opportunity, they could feel more comfortable about expressing their thoughts and feelings related to DataONE.

The method of analytic induction was applied to find common patterns by reviewing the transcripts line by line for themes or categories emerging from the initial

cases, then modifying and refining it on the basis of subsequent cases. The researcher was interested in observing a coherent relationship of the themes in the actions of the agents that are reflected in the actions of the collaboration. These themes were non-linearity, counteracting forces, positive and negative feedback loops, prediction impossibility, action irreversibility, co-evolution, self-organization, emergence, dissipative structures, bifurcation, self attractors, dynamic equilibrium, and sensitivity to initial conditions.

2. Naturalistic observations

Ethnographic methods are also frequently used in case study research designs. Naturalistic observation, observing subjects in their natural environment, seems to be a good fit as this method is used when little is known about the phenomenon being investigated or questions involving the natural flow of behavior (Grazione & Raulin, 2000). Emergence of DataONE had both criteria. In naturalistic observation, the observer does not intervene at all. For all intents and purposes, the researcher is unobtrusive and works hard not to interrupt the natural dynamics of the situation being investigated. Naturalistic observation provides rich descriptions about the nature of the social world where there is little or no manipulation of the environment; therefore they would provide valuable insights to the researcher in analyzing the data generated through semi-structured interviews.

It has to be mentioned that this method has two important limitations. First, the findings are not generalizable. Second, even though the procedures of naturalistic

observation are clearly specified, there might be some changes as the study continues; thus, the procedures might not be followed exactly. As a result, such studies, like other qualitative approaches, are more flexible (Marecek & Fine, 1997) but harder to replicate. This is not a bad thing, just a trade-off between flexibility and replicability. As little is known about the phenomenon being investigated, it is very expected to have a research design that does not fit the needs of the phenomenon 100%. It is the researcher's skill and flexibility of the method that adjust the fit between the phenomenon and the research design.

The researcher had the chance to attend two All-Hands-On meetings and one Community Engagement & Outreach Team meeting. Around a hundred people attended the former whereas the latter had around 35 people. All meetings lasted for three full days. The researcher took notes and avoided professional contact in order not to intervene the group dynamics. By attending the meetings, the researcher explored the functioning of DataONE and its agents all together in its natural setting, the formal and informal communications behaviors, the evolution of DataONE, and most importantly had a better understanding of the collaboration.

Furthermore, the researcher gained access to the internal DataONE website and had the opportunity to examine the artifacts created by DataONE members for various purposes which helped the researcher to interpret his findings.

3. Survey

As for the quantitative component, an online survey with 24 questions was prepared and posted on a server and the link was distributed to the members of the collaboration. This component is descriptive only. In addition to some demographic questions, quantitative values related to the frequency of communication, types of communication channels and information sources were sought. The survey instrument is provided in the Appendix.

The link was distributed to 100 email addresses. 51 responses were received, for a response rate of 51%. The reason for such a high response rate on an online survey might be that the group is small and the participants know the researcher. The survey stayed live for two months on surveymonkey servers (www.surveymonkey.com). Three weeks after the first invitation email, a reminder was sent to the potential participants. The email list was obtained from the DataONE website.

Data Integration, Evaluative Criteria, and Analysis

Analyzing results for a case study tends to be more opinion based than statistical methods are. The data was collated into a manageable form and it was constructed in a narrative way around the basic concepts of complex adaptive systems theory. Concise and interesting findings are supported with numerical data (if possible).

All methods have limitations. By using multiple methods to generate and collect data, the researcher overcame some of the limitations. For instance, naturalistic observation

provides mostly descriptive data, whereas semi-structured interviews provide explanatory data. Yet, findings coming from both of them were not weak in terms of representativeness, and thus were not generalizable. Using multiple methods is one of the four triangulation methods that Denzin (1978) proposed: data (use of variety of data sources), investigator (use of several researchers), theory (use of multiple perspectives to interpret data), and methodological (involves using more than one method to gather data, such as interviews, observations, questionnaires, and documents). Through methodological triangulation, the researcher looked for similarities and regularities in the results which increased the validity[12] of the results. In addition, data triangulation is also used. The data came from different sources: from the participants and also the internal website of DataONE.

The main concepts for evaluation for such studies are authenticity, credibility, and trustworthiness (Corbin & Strauss, 2008). Authenticity is reached through member check. The transcripts of the interviews were sent back to the interviewees for revisions. Everything that was used in the analysis was confirmed by the participants in order to ensure that the analyses were based on what they meant. Credibility and trustworthiness are embedded in the researcher's competence. The researcher has done similar studies

[12] Validity in broad meaning, not statistical validity.

before that were presented and published in various venues (Aydinoglu, 2010a; Aydinoglu, 2010b; Tenopir et al. 2011).

The themes generated through semi-structured interviews were compared to the findings of the naturalistic observations in order to see if they support each other. For instance, the behaviors of the meeting participants were explained by the claims of the interviewees. The findings of the online survey supported the themes that were generated through interviews. There were also differences in results. Both for the analysis of data generated through the semi-structured interviews and naturalistic observations context were given special importance. Thick descriptions –behavior of the participant and its context (Geertz, 1973)– and emic language –language of the participant in his/her own language and culture (Headland, Pike, & Harris, 1990)– were reflected in reporting.

As for the semi-structured interviews, the analysis of data started once the interview started. However, the interviewer suspended judgment, eliminated, or at least gained clarity about, preconceptions (Patton, 2002, p. 407), manufactured distance (McCracken, 1998, p. 22) in the pilot interviews; otherwise the data could have been both generated and analyzed with biases. Since every word said and action taken by the researcher during the interview has an effect on the response (such as an encouragement on particular topic by the researcher, might make the participant focus on that topic only during the interview); the researcher minimized these effects and encouraged the participant to express his or her mental construct. These were ensured by the subsequent

questions and the floating responses which were emerged from the immediate analysis of data.

Data must exhibit “symptoms of truth”, which are exact, economic, mutually consistent, externally consistent, unified, powerful and fertile (McCracken, 1998, p. 50). Unless these conditions are present, it means that the study does not have the appropriate standards. These standards establish the credibility needed and the researcher believes that they are present as quotes from the participants are provided in the manuscript as much as possible.

The analysis and discussion is presented as McCracken suggests (1998, p. 52-8). Since this study is an exploratory one, an ‘open-topic write-up’ approach was employed which “allow(s) rich and abundant data to speak to the reader” (quoting as much as possible) and “provide(s) a clear and vivid sense of the ethnographic particulars while also showing the general formal properties and theoretical significance of these data” (make the necessary connections with previous studies, if possible) (p. 58). Special attention was given to use the passages of respondents’ words and descriptors because they provided a basis for accepting, rejecting, or modifying the conclusions to the reader. Moreover, they were needed in assessing the validity of the study.

Complexity theory and the emergence concept were used for data analysis and interpretation. Thinking of scientific collaborations as complex adaptive systems is a new way of improving our understanding of communication and information behaviors of

scientific collaborations. Complexity theory provides a novel look at scientific collaborations because of its ability to explain and make sense of unique, unrelated or one-time events (Thietart & Forgues, 1995) by especially focusing on the non-linear relationships and interactions among units. Furthermore, the emergence concept, following Sawyer's interpretation (2005), elaborates the importance of complex communications among agents. DataONE's complex multidisciplinary nature fits very well for such an analysis. Wagner had a similar understanding (2008). She also considered scientific collaborations as complex adaptive systems; however, she was interested in the policy applications of such understanding and focused on incentives, stimulations, and inducements.

Chapter 5

Results

In Chapter 2, the basic features of complex adaptive systems theory are summarized. That compilation is used to prepare a framework for complex adaptive systems. Through that framework (Figure 10 - 10-concepts) it might be possible to assess whether a collaboration is a complex adaptive system or not. An organization/collaboration that operates according to complex adaptive systems theory are different from one that operates according to a linear model. First, they are smart, they learn things by themselves. Second, the ability to learn makes them adaptive to changing environments. Third, adaptation gives them resilience to external and internal threats, which is another advantage they have over traditional (linear) systems. Fourth, since they are self-organizing and also dissipative, they are cost effective; they emerge when needed and disappear when not needed. Finally, they are quite innovative –often in an unpredictable way; which drives forward development. Therefore, the assessment of a scientific collaboration is crucial in deciding allocating limited resources to the one that has the maximum potential to be successful.

This chapter follows the outline of the complexity framework (see Table 3 below) that is developed for scientific collaborations using the literature summarized in Chapter 2. First the components of DataONE are introduced –individuals and organizations. Second, the diversity of these components is explained –disciplinary diversity, institutional diversity, geographical diversity to name a few. Third, the interdependency and

communication in DataONE is described. This part is merged with feedback concepts as two-way communication involves feedback process. Fourth, this is the story of emergence of DataONE as a whole, yet it is in fact the emerging substructures that create DataONE. Fifth, the internal structure and external environment are used to illustrate the edge of chaos and adaptation concepts. Finally, the early impact of DataONE on the scientific community is discussed.

Table 3 - Complexity Framework

Concept	**Short definition**	**Analogy**
Components	Agents in the system	Ingredients
Diversity	Variation of agents in the system	
Interaction & interdependency	The nature of the relationship among agents	Cooking
Feedback	Assessment of the relationships among agents	Tasting
Unpredictability	System's behavior arising from nonlinear relationships among agents	Unskilled cook
Edge of chaos	The environment that a complex system could exist	Heating, stirring
Emergence	Self-organization, the outcome of the change	Ready-to-serve
Adaptation	Learning and the new equilibrium for the system	Changing ingredients
Historicity	A cryptic determinism	Unskilled cook
Co-evolution	Contagious/spreading adaptation or repositioning according to the other systems	Others meals

It has to be mentioned that the results presented here are still related to DataONE's emergence phase. The project started to receive funding and a lot of work has been done;

however, at the time of the study the system was not public yet. According to the presentation done for the NSF second year review (February 2011), DataONE will be active by the end of 2011. The data for this study was collected between February 2010 and March 2011. The project has received funding for five years and a second five years of funding is plausible according to the interviewees. This study was conducted in year two. Given these facts the collaboration should still be considered in the emergence phase. All the results reflected in this study belong to this phase.

1. Large number of components and counteracting forces:

If DataONE is going to be considered as a complex adaptive system, it should have a number of components and it has. Moreover, this number is increasing because DataONE grew rapidly in its first two years. When the researcher started the study in February 2010, the number of individuals directly included in DataONE activities was around 40. After 18 months the working groups alone have over 100 individuals, to which the survey was distributed in February 2011. More than 200 people expressed their interest and gave their emails to be contacted for related DataONE activities.

In terms of institutions involved, there are eleven types of institutions, which means at least eleven different types of institutional goals exist. The maps below (Figure 3 and 4) show the institutions that are already involved in DataONE activities as member or coordinating nodes and the ones that were interested in being a member node as of 2014. At the time of this study was written, DataONE has three coordinating nodes –University

of New Mexico, Oak Ridge National Laboratories & University of Tennessee partnership, and University of California Santa Barbara National Center for Ecological Analysis and Synthesis (NCEAS)– and three member nodes –the Knowledge Network for Biocomplexity (UCSB), the ORNL Distributed Active Archive Center (DAAC), Dryad (at Duke's NESCent). By the end of 2011 this number is expected to increase to six, by 2012 to 10, by 2013 to 20, and by 2014 to 40. As of 2011 there are 29 institutions in the US and 19 more worldwide that mentioned that they are already involved or interested in DataONE. (Details can be seen in Figure 3 & 4)[13].

[13] DataONE website, retrieved on February 2011, from https://docs.dataone.org

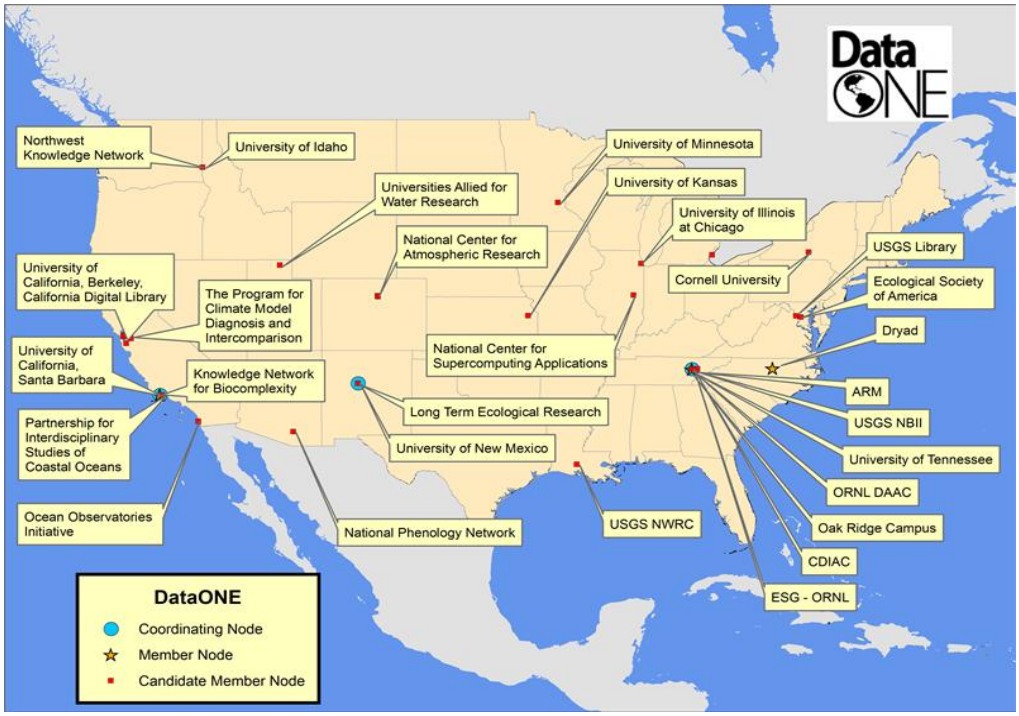


Figure 6 – Involved or interested institutions in the U.S. as of February 2011.

Figure 7 – Involved or interested institutions world wide as of February 2011

It was mentioned before that there is not a ‘certain’ number in the literature to argue that ‘this amount is sufficient to have complex system.’ Different systems have different number of agents. For a human social system, a scientific collaboration such as DataONE, the researcher believes that the number of agents on different levels (individual & institutional) is sufficient regarding DataONE to be considered as a complex system.

2. Variation and Diversity

The real reason to have variation and diversity in the system such that it can be considered as complex system is that the system needs counteracting forces in it. A complex system exists at the edge of chaos or between order and disorder (chaos). If all agents are the same it would become a highly ordered linear system. The data collected revealed diversity at different levels: stakeholder perspective, disciplinary perspective, career age perspective, motivations to join DataONE, types of institutions involved, and geographical diversity. The management team, which was the name for the leadership team at the beginning, also recruited new members while considering diversity.

The existence of counteracting forces in the system has been realized by the DataONE team early in the project –even during the grant proposal writing time. The team realized that problems related to data practices involve different stakeholders because the data lifecycle revealed that there are many perspectives and concerns regarding data. In order to be able to create the technology-enabled science capacity, in which data access, sharing, and preservation is crucial, the functions of the data lifecycle

has to be understood. I am going to use a model developed by the SocioCultural Working Group which later has become the official data lifecycle used by whole DataONE and also approved by the NSF.

The importance of data lifecycle model is that all of DataONE services and products are created from this model; hence, it constitutes the base for DataONE activities. The phases in the data lifecycle involve different types of agents. For instance, data is collected by scientists (or remote sensors) but they might or might not be involved in the rest of the phases until they have been analyzed –'analyze' phase. Librarians, data curators or data managers might 'describe', 'deposit', and 'preserve' the data. However, to do that, they need the necessary tools, equipment, and training which are supposed to be provided by another party –policy makers. There have to be platforms developed for this system which is the job of computer scientists and also information scientists. However, it is not enough to have a cyberinfrastructure if the scientists do not see the value of doing this. These activities are time and money consuming; thus, scientists need to be motivated to take care of their data. Thus, they have to be made aware of the benefits. Yet, they are generally conservative because there are other issues such as copyrights of data, acknowledgement, etc. Furthermore, they lack the skills (which can be taught by librarians) and the resources (which can be provided by librarians or other sources).

In conclusion, Figure 5 demonstrates the variation of actors and their activities in the data life cycle (DataONE, 2011). Representatives from each phase in the figure below are included in DataONE.

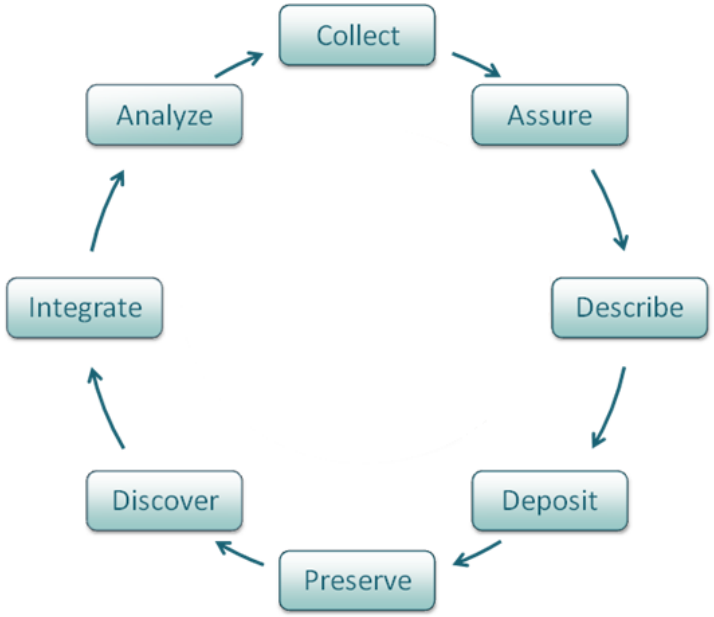


Figure 8 – Data lifecycle as adapted by DataONE as of February 2011

One participant describes this deliberate process as such:

> "We have spent a lot of time on you know, sort of what are the primary groups we are trying to serve, right? What are our primary audiences and stakeholder community? To sort of serve those effectively in some form or fashion, whether it is in the leadership team or working groups or management, whatever, we know we need those disciplines properly represented, right? So, I guess, from my perspective, you know, we know we need to work within the library community. We know we need to work within the computer science community. We know, of

> course, our primary stakeholder is some kind of earth scientist or biologist or ecologist, whatever it is. So, by that defining of our stakeholder group or primary audience we are trying to serve, that sort of identified the various disciplines that we need to make sure were involved."

DataONE is a multidisciplinary collaboration. At first the project focused on cyberinfrastructure only. However, on one of the very early NSF consultations, the NSF requested library science involvement to connect the goals of the project with the scientific community. One of the interviewees describes the process as such:

> "So, you know, one of the basic premises of DataONE is that libraries can impact the community in terms of whether it is training or being that first line that researchers go when they start their project to educate them or even to deposit data. That sort of has been the reason why we have involved people who lead, say, university libraries or USGS, for instance."

Another interviewee tells how two strong library partnerships were established.

> "So we had early discussions…this has been over two years ago, but early on into the project it was clear that NSF expected a significant involvement of what we could loosely call library science community in the DataNet partners. At the time when I got involved in this we really did not have a strong library science partner in the organization, in the proposal team. … So X and I proposed to bring in Y and Z into the discussion and at the same time there was a parallel that was

brought in, T and the California Digital Library folks so that actually brought us two strong library partners."

Of the 51 respondents to the online survey, almost one third of them (16) responded that their subject discipline is Library and Information Sciences. Computer science (7) and ecology (7) follow with 15% each. As it can be seen in Figure 6 below the collaboration is quite multidisciplinary.

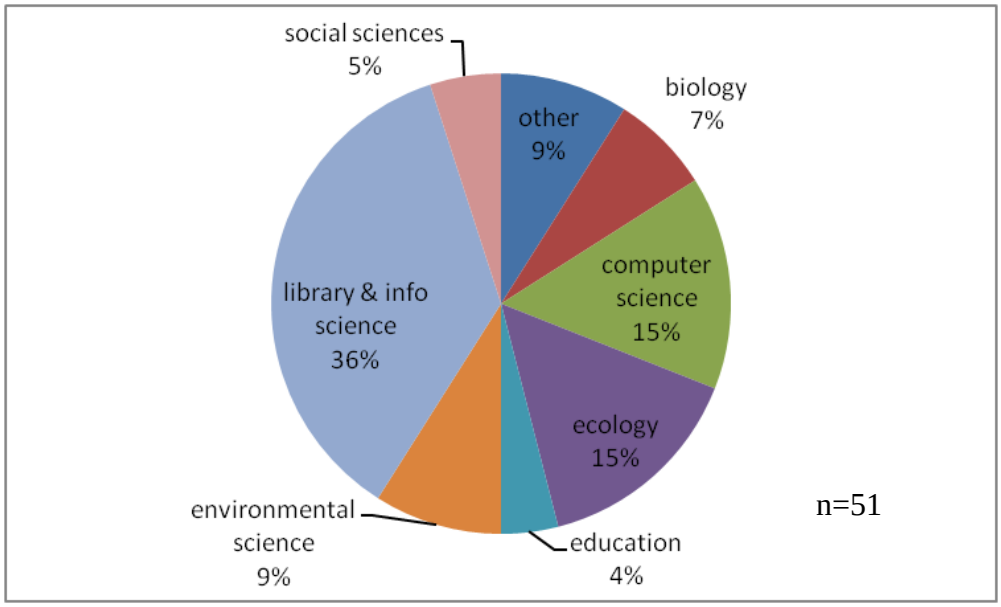


Figure 9 – Subject disciplines in DataONE according to the responses to the survey

This multidisciplinary structure is also reflected in the leadership team. There are chemists, ecologists, biologists, library and information scientists, and computer scientists. They work in the academia or for government.

During the interviews, the participants mentioned that they spent some time on the multidisciplinary nature of the collaboration and one of them defined three different aspects

of this nature: multidisciplinarity in earth sciences, such as oceanography, geology, geochemistry, etc.; multidisciplinarity from cyberinfrastructure perspective such as integrating visualization tools; and using library and information sciences as a community engagement tool.

> "I mean, the goals of the project are very much to serve the science community through technology development. So naturally, without really active engagement of the science community and the social science community, we cannot effectively reach the goals of the project. So, I think the multidisciplinary nature of the project is critical and the value it brings I providing a mechanism for us to actually meet the goals."

Institutional diversity is prominent in DataONE as well. Different types of institutions have different agendas. They are different stakeholders –only in an institutional level –not an individual level. Thus, they also create a variation in the DataONE system. There are eleven different types of institutions in DataONE:

12. Academic institutions from the U.S. (including three EPSCoR [The Experimental Program to Stimulate Competitive Research] states—Tennessee, Kansas, and New Mexico) and the United Kingdom (i.e., Edinburgh, Manchester, Southampton);
13. Research networks (e.g., Long Term Ecological Research Network, Consortium of Universities for the Advancement of Hydrologic Science Inc.

[CUAHSI], Taiwan Ecological Research Network, South African Environmental Research Network [SAEON]);

14. Environmental observatories (e.g., The National Ecological Observatory Network [NEON], USA-National Phenology Network, Ocean Observatory Initiative, South African Environmental Observatory Network);
15. NSF- and government-funded synthesis (i.e., the National Center for Ecological Analysis and Synthesis [NCEAS], the National Evolutionary Synthesis Center [NESCent], Atlas of Living Australia) and supercomputer centers/networks (Oak Ridge National Laboratories [ORNL], National Center for Supercomputing Applications [NCSA], and TeraGrid);
16. Governmental organizations (e.g., U.S. Geological Survey [USGS], the National Aeronautics and Space Administration [NASA], Environmental Protection Agency [EPA]);
17. Academic libraries (e.g., University of California Digital Library, University of Tennessee, and University of Illinois-Chicago libraries, which are active in the digital library community and are members of the Coalition for Networked Information, the Digital Library Federation, and the Association of Research Libraries);
18. International organizations (e.g., Global Biodiversity Information Facility, Inter American Biodiversity Information Network, Biodiversity Information Standards);

19. Numerous large data and metadata archives (e.g., USGS-National Biological Information Infrastructure, ORNL Distributed Active Archive Center for Biogeochemical Dynamics, World Data Center for Biodiversity and Ecology, Knowledge Network for Biocomplexity);
20. Professional societies (e.g., Ecological Society of America, Natural Science Collections Alliance);
21. NGOs (e.g., The Keystone Center); and
22. The commercial sector (e.g., Amazon, Battelle Ventures, IBM, Intel) (DataONE Proposal, 2009).

Another variation is the career age of participants in DataONE. Although the leadership team consists of seasoned scholars, the overall collaboration is open to researchers from any career age from newly minted PhDs to senior researchers. As it can be seen in Figure 7 below it is quite diversified. Naturally, the goals of a senior researcher could be very different from a junior one.

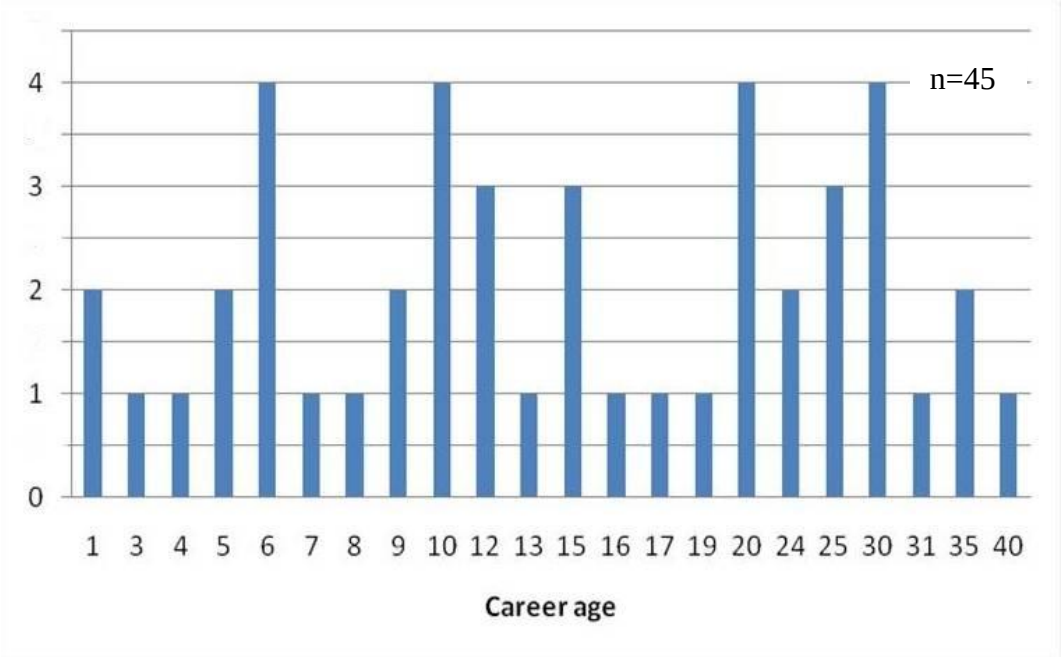


Figure 10 – Career ages of DataONE members according to the survey

Another reason for counteracting forces in the collaboration is the motivation to join DataONE. Individuals in the leadership team mentioned different reasons about why they joined DataONE. The most stated reason was professional/research fit, which is the motivation directly related to the participant's career such as a research interest or a job task. For instance one participant expressed that s/he needs to conduct research to get her/his tenure:

> "Well, I pretty much welcome any opportunity to work with the folks at T, it has been a very good match for me in terms of, you know, part of my promotion and a tenure process here."

Another one said that s/he has a research career build on scholarly communication and this project provides opportunities to work with the scholar s/he studies:

> "… and I realized, 'you know I am doing all this work, my whole career is looking at scientific communication and scientists and publishing, I really ought to be involved in a project that is working directly with scientists'."

The potential to be productive is also important. Some participants found the problems dealt by DataONE intellectually stimulating, which could result in scholarly production.

> "I also thought it could be very professionally productive to do this in terms of the infrastructure being developed as well as the publications and other products that could come out of the endeavor as well."

A non-academic participant sees the parallelism between his daily tasks and DataONE's goals:

> "I am motivated by this desire to build the tools and infrastructure necessary to support reproducible science, especially focused on the environmental sciences but science in general."

The second reason to join the collaboration was the institutional fit. The participants explained their motivation through the organizations they work for. The fit between their institutions short- or long-term goals and DataONE, the desire to broaden their reach, and make use of others' resources were considered under this title. For instance, NCEAS, as a

synthesis center combines methods and perspectives from different disciplines to address major problems in ecology. The director sees the fit and opportunity:

> "Well the need for it is apparent in the work that we do at (the center). As an independent researcher myself, I know that there is a need for sustainable, usable, accessible infrastructure for data and in my role at the center, one of the things I do is facilitate research that uses existing data so the need for it is quite obvious and this is the right group of collaborators to do it."

In some cases, the participants used 'we' instead of 'I' even though they were asked for their personal motivation. Joining forces with DataONE helps them to achieve their organizational objectives.

> "There is a lot of real relevance to the processes that DataONE is planning to build for a cyberinfrastructure that would be really useful for the kinds of research that we do here. … We have won several fairly significant National Science Foundation awards for (our facility) based on those kinds of things (informatics & data interoperability). And we are really excited about supporting DataONE so it can provide a platform for us to do our work."

Another participant who works for government pointed out the importance of networking and effective use of resources through sharing tools:

> "I thought it was the wave of the future. We have our ... data center but it is really isolated. We realize there are a lot of other activities going on out there and we need ways to link our holdings with other holdings

> and we need to take advantage of other practices, what folks are doing, citations, and tools and services. We just cannot do it all in isolation. So, we have some skills that we would like to share with others and we want to see what others are doing and see if we can incorporate those practices without having to reinvent the wheel."

Some of the diversity mentioned above were actually the result of the recruitment process. The members of the management team expressed how their concerns on diversity shaped their recruitment decisions. Although, the collaboration is open to anyone who is interested now, early on people were invited according to their background, research interests, gender, and institution.

> "And in addition, we wanted to have as much diversity in the mix as possible –both institutional diversity as well as gender and other types of diversity as well. And as part of that, we did not want to overload it with too many people from any one institution so even though there may have been multiple people from the same institution that we could have invited, in a lot of cases we did not so we could expand the institutional diversity as part of the mix."

The diversity in the project ensures equity among different perspectives according to a participant. When a particular perspective is dominant, the minorities do not get enough attention or even feel neglected.

> "I have been on other projects where there is such a diversity and one or two members from a different field and in those cases it was much more difficult because they were much more of a minority. And so,

> there is a tendency to, you know, not address that particular discipline so much."

The final type of diversity is the geographical diversity. The maps in the previous section (Figure 3 &4) show the institutions from different parts of the world that are going to be involved in DataONE activities which will increase the diversity.

In conclusion, in order to achieve project goals DataONE has employed people with diversified and rich backgrounds who have different motivations, different organizational objectives in mind. Yet, for a system to demonstrate complex adaptive behavior, these agents should interact with each other frequently and their existence or tasks should be interdependent to each other.

3. Connectivity, interdependence, and interaction

The components of a complex system should have an impact on each other like a ripple effect. The ripple effect is explained through the interdependency of different working groups and the interaction is through the communication behaviors. The barriers and problems regarding communication are discussed in detail in Chapter 6.

DataONE is a virtual organization. Thus a lot of communication/interaction happens among its members and they happen online. Frequent communication/interaction is also important for daily tasks to continue because one unit's job is dependent on the others. In this section, first the units (working group structure) are introduced; second, the results related to the communication/interaction among these units and individuals are

provided; finally, the problems occurred regarding communication and how they were dealt with are explained.

In order to ensure that members are communicating with each other, the management team invited people who have experience in virtual organizations, who are known be good communicators, and who are able to work in teams. As one participant put it, apart from the diversity criteria, this was the fundamental principle.

> "The criteria was that we wanted to have people who were good communicators who would listen and would really not want to do their own thing so that was a criteria. People who were difficult to work with or whatever, we tried to avoid that. So we built a team of people based on that. It was a real fundamental principle to start."

Another participant also emphasized the importance of compromising, which is an important concept when dealing with conflicts.

> "We definitely wanted to make sure the people we were bringing on board had a good reputation for, again, being able to work in a group. So, their abilities to communicate, their willingness to compromise, and their effectiveness at working virtually were all critical components in the decision making."

In sum, the agents have the necessary skills and experience to communicate but before moving forward, the units (working group structure) have to be explained in order to describe the interdependency and also the reason for frequent communication/interaction.

Interdependency & Working Groups

Working group structure is quite common when traditional funding mechanisms do not let researchers come together and conduct their research on especially interdisciplinary phenomenon. It fits to the goal of DataONE as the problem being studied is multidisciplinary and the participants are volunteering their time (except for a small fraction of employees and travel grants). A key player in DataONE, NCEAS, has extreme experience and research in similar structures which was mentioned by both of the interviewees. Basically there are two themes for the Working Groups (WGs) in DataONE: cyberinfrastructure and community engagement & education/outreach. As it can be seen from the organization chart below, there are five WGs on the cyberinfrastructure side and four on community engagement/outreach. Two of the WGs are cross-cutting.

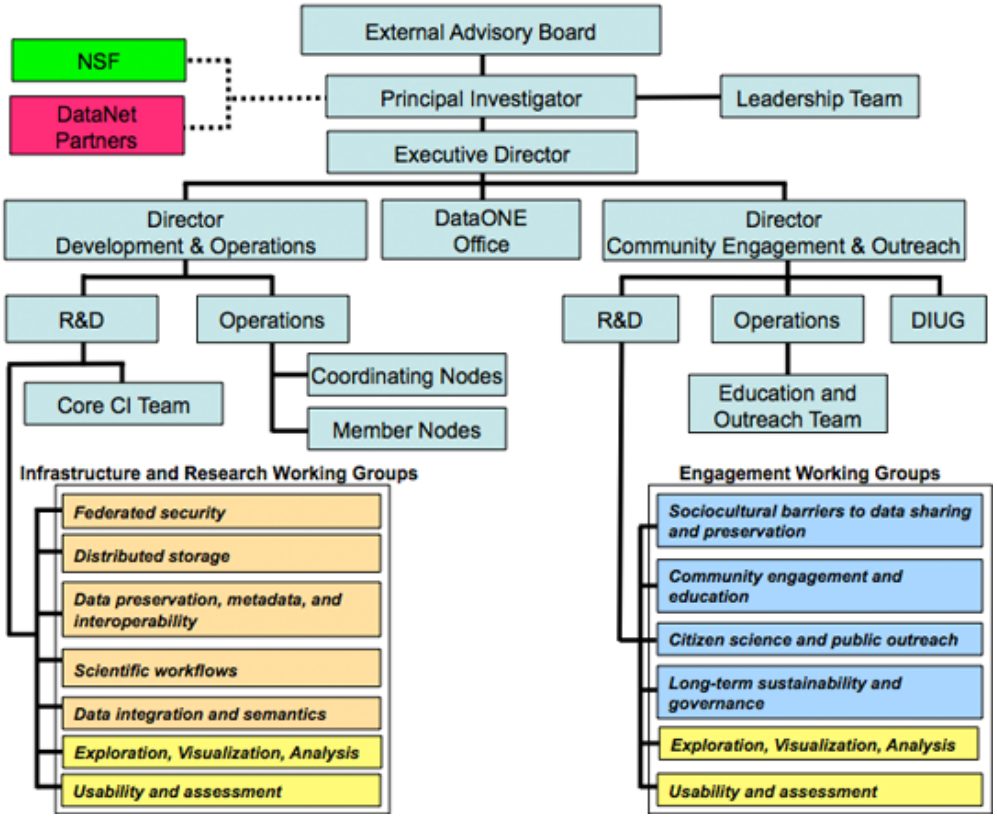


Figure 11 – Organization chart for DataONE adapted as of 2010

The development of the cyberinfrastructure is not a merely technical task. The users have to be considered when developing the systems. Therefore, the tasks of the two big teams –cyberinfrastructure (CI) and community engagement (CE)– are interdependent. Below, both approaches are provided to demonstrate the interdependency.

> "So, for example, on the CE side of things, we talk about user scenarios and user scenarios are made up of a lot of different activities that the user is engaged in. For example, they would go to the computer and log in to the system. They would conduct a search of the coordinating node and download data from the member nodes. They would integrate it.

> We talk about all of those things being activities within a scenario and the whole scenario is perhaps searching data, gaining data, analyzing the data, writing the data and publishing it. It is just a very broad scenario and then we have these integrated activities. On the CI side of things, they talk about those individual activities as being case studies and that is because they need to break it down to; the user sits and logs in, okay, what types of cyberinfrastructure support do we need to have for that capability?"

When the data collection was started, some of the WGs were active, some just established and some of them not active. At the beginning the attention was on the cyberinfrastructure component of the project as it is the main product. Without it, there was not nothing to promote to scientific community, no feedback from the scientific community, nothing to educate the scientific community on, etc. However, the cyberinfrastructure was going to be created for the scientific community and it was supposed to be built on their needs (some of which they are aware of and some of which they are not). Thus an Ad Hoc Group was created to start assessing the stakeholders' current conditions to have a baseline. One participant explains the Ad Hoc Group as such:

> "The usability is not happening yet but the assessment had to get started right away so we could create a baseline to measure future activities against. And the way we are building the working groups couldn't happen fast enough to get the baseline out and so a smaller group of people were assembled that started working on developing and deploying the baseline"

The first assessment was done on scientists' scientific data practices and attitudes towards data sharing. The results indicate that "Barriers to effective data sharing and preservation are deeply rooted in the practices and culture of the research process as well as the researchers themselves" (Tenopir et al., 2011).

The WGs are simple collective units in DataONE. Each one has goals that are interdependent to another. For instance, the usability & assessment WG conducts surveys to measure a baseline for different stakeholders (scientists, libraries, educators, etc) and later will measure the same variables to see whether DataONE had an impact on them. Another task that they do is conducting usability tests to provide feedback to developers from users. In a nutshell, the activities of the usability & assessment WG are connected to other WGs. The same holds for other WGs. Without the cyberinfrastructure there is nothing for usability & assessment WG to provide feedback for or to measure the impact of. Another example is from the preservation, metadata, and operability WG. For this WG to achieve its goals, community engagement & education WG provides education and trainings to both librarians and scientists who would like to use the system; sociocultural issues WG investigates for instance the organizational support (or lack of) towards data preservation and providing tools for metadata preparation and so on. Each of the WGs are connected to other WGs. The table below summarizes the goals of the WGs and a close examination of them demonstrates the interdependency among each other.

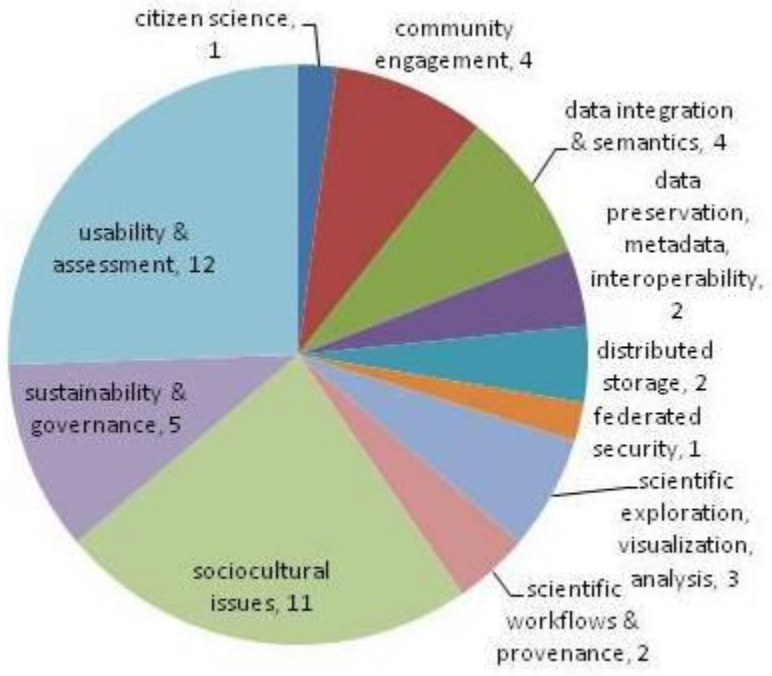
Table 4 – Working Groups in DataONE

Working group	**Goals of the working group**
Federated security	i) establish federated identity management scheme and authorization/access-control for provisioning resources within a distributed DataNetONE infrastructure that supports a large user-base.
Distributed storage	i) define and select production-wide area file system(s); ii) define and select production data (file, block, storage object) movement services for transfer of data between nodes and for transfer to and from users; iii) define and select production data-related services including tools for file replication management, replication location, staging, and planning as well as the specification of needs for continuous validation, data warming, and consistency checking.
preservation, metadata, and interoperability	i) identify, evaluate, select, and implement the standards, tools, procedures, and internal policies needed to support data curation and preservation and metadata management; ii) exercise the standards, procedures, and tools deployed at the initial system implementation; iii) develop a plan for a comprehensive internal summative evaluation to determine the effectiveness of tools, procedures, and systems.
Scientific workflows	i) evaluate and co-develop workflow archival formats; ii) develop data and workflow provenance interoperability framework; iii) generalize existing, emerging workflow repositories; iv) gather/develop workflow design patterns for commonly used systems.
Data integration and semantics	i) design schema object repository architecture; ii) specify schema classification and interoperability assessment services; iii) research and prototype source registration, mapping, integration services.
Usability and assessment	i) interact with DIUG Community and initiate research to assess current practices and future needs using user-centered design approach

	(e.g., initially survey users interacting with existing archives and metadata management systems in use at participating institutions for rapid input on usability issues); ii) recommend enhancements to tools, products, and services; iii) oversee assessment plan that assures deliverables and schedules are met, and that broad community involvement occurs throughout the project lifecycle.
Sociocultural issues	i) identify and examine the sociological and cultural issues that inhibit effective data sharing and long-term preservation; ii) evaluate and recommend strategies that overcome sociocultural barriers and create incentives for data preservation; iii) explore and make recommendations regarding the roles for libraries in training data authors, supporting data curation, and acting as a facilitator of digital preservation practices.
Community engagement and education	i) determine effective mechanisms for community input on tools for data providers and consumers and for the dissemination of products appropriate to scientific and non-scientific audiences; ii) establish a training program for both science and citizen science initiatives; and iii) establish metrics that will be used to determine the adoption success and utility of DataNetONE products.
Citizen science and public outreach	i) determine requirements for management of citizen science data and visualization, exploration, and analysis of data by disparate users (from citizens to scientists); ii) create a comprehensive data management strategy for highly disparate citizen-based observational networks; iii) build tools to allow project managers, researchers, educators, or networks to develop a customizable web-based data gathering system.
Long-term sustainability and	i) investigate different organizational models, including a stand-alone non-profit 501(c)(3) organization; ii) investigate different funding

governance	models to ensure long term sustainability; iii) establish the governance of DataNetONE and a representative stand-alone organization (i.e., DIUG) to ensure that stakeholders provide direction.
Exploration, visualization, analysis	i) to develop examples that generate scientific publications and multiple data visualizations and explorations that highlight the value of the DataONE process and exhibit the enormous potential of the synthesis of large and disparate data resources.

According to the survey, usability & assessment WG (12) and sociocultural issues WG (12) have the most members. Although, there are not any restrictions or rules, the membership on the community engagement & education side is stricter compared to the cyberinfrastructure team. For instance, in meetings the members of the former WGs stay together and work on their own WG's agendas whereas cyberinfrastructure WGs often uses the workshop structure to deal with software development issues in which members from different WGs come together around a specific problem. This might be due to a higher level of interdependency among cyberinfrastructure tasks. Indeed, one individual in the cyberinfrastructure team, who is also in the leadership team, objected to the "please identify your primary working group" question in the survey on these grounds. He considers himself and some others in the "core cyberinfrastructure team" and strongly disagrees with the notion of having a primary or secondary working group. Because of the casual formations around cyberinfrastructure issues and move in between cyberinfrastructure teams, the membership in the cyberinfrastructure teams are considered less strict. The responses regarding 'primary working group' are provided in Figure 9.



n=47

Figure 12 – The number of individuals in working groups according to the survey

Some WGs are closer to some WGs than the others yet all of them are interdependent with each other. The members –especially the ones in the leadership team– are aware of this fact. There is frequent interaction and feedback among them as expressed by one participant below:

> "...we could build all the technology in the world and if it is not adopted, it will have no impact on the fields of science we are interested in. These things are all critical and they are all motivating for me to interact with these other groups. Like the sociology group, for example, they have done these baseline assessment surveys to assess the attitudes towards data sharing and data management and other

> things within the scientific community. I have been real interested in the results of those because they tell us something about where we should be heading with respect to development activities on technology. So, it is really useful to interact with that group a lot."

The online survey also found similar results. Eight percent of the respondents mentioned that they do communicate with anyone that is in other WGs; nearly 20% of them communicate weekly, and 30% of them bimonthly or monthly (n=43). Email (26) is the most used tool in communicating with people in other WGs. Videoconferencing (10) and plone website (9) is also frequently mentioned.

Communication behaviors

The interactions among members are analyzed through communication behaviors. The interviews with the leadership team revealed some valuable information about the types and frequency of communication. Email is again the most frequent tool to communicate among different groups. Videoconferencing is also quite common, using the software Maratech and Skype. For instance, the leadership meets once a week through the videoconferencing software, Maratech, for around an hour; WG meet at least once a month; and the core cyberinfrastructure development team meets every morning for half an hour to 45 minutes. In case of a deadline, such as the External Advisory Board meeting or the NSF review, the frequency and the duration of these meetings increases.

Face to face meetings are also important; even though, the frequency of them is very rare compared to virtual meetings. The whole cyberinfrastructure team and

community engagement & education/outreach team meets twice or three times a year for a three day meeting. These meetings are intense. DataONE All-Hands meetings are held once a year in which everyone attends. The schedule of these meetings is similar. First, new members are brought up to speed in a plenary session with “an introductory synthetic talk about DataONE”. As DataONE is still in the emergence phase, at every meeting there are some new faces that are not familiar with the project. Second, the all-hands meetings include a summary of past activities. Third, WGs at the meeting meet by themselves. This part is the most intense and longest part. Agenda is prepared beforehand by the WG leads and the leadership yet priorities could change or new items could be added to it. Generally, both new activities are planned and even new products emerge in these three intense days. However, the dynamics of each WG are different. For instance, whereas members of the usability & assessment WG ask for more autonomy in the topics they want to cover, members of the sociocultural WG demand to be given tasks from their WG leads. The last day (or last half day), again in a plenary session, issues that concern DataONE as a whole are discussed. These meetings are quite important and effective in establishing a collegial environment and organizational identity, creating a network, and producing scholarly work and software.

Besides email, videoconferencing, and face-to-face meetings, subgroups use different tools for communication purposes. For instance, IRC (internet relay chat) is quite common among the cyberinfrastructure team and they are quite fond of it.

> "…we have an IRC channel that we run. And so we use IRC and the main developers from the project are all logged into IRC whenever they are working. And so that is one of the ways we get, it is sort of like being in the same room in the sense that you can say, "hey X, did you know this," and he can answer. If he is not there, it is no big deal. If he is there, it is kind of like being able to yell across the room. Except the room is the difference between Colorado and Alaska."

> "…with the software development team, we use IRC on a daily basis, but I do not think some of the other groups use that so much."

Another favorite of the cyberinfrastructure team is Subversion, which is software that helps to keep the track of different versions of program that the software developers working on. The community engagement & education/outreach team relies heavily on the plone website, which is a website created as a repository for DataONE related documents, presentations, images, etc. Etherpad is another popular software among DataONE participants which is both used in virtual meeting and face-to-face meetings. Etherpad is word processing software that can be accessed and edited by multiple users simultaneously. Heavy using of etherpad in every face-to-face meeting was observed. Etherpad is also used on weekly virtual leadership team meetings to follow the agenda and to keep the minutes. Another tool used for communication is wikis, yet it is limited to sociocultural WG and usability & assessment WG only at the time of the study. They have been established as not only to communicate among WG members but also to utilize the knowledge of interested parties all over the world.

In conclusion, members of DataONE have to interact/communicate frequently with each other as their tasks related to the project are interdependent on each other. Since DataONE is a virtual organization, most of the communication is facilitated through the computer. Email, videoconferencing, and shared space applications are the most frequent tools that are used by DataONE; however, face-to-face meetings are also important, effective, and productive though they are rare. Due to large the diversified member structure of the collaboration, coordination and communication problems occur. They are expected and dealt with delicately. The details of the communication challenges are discussed in detail in Chapter 6.

4. Feedback

Two-way communication and the encouragement of participation of every member by the leadership team ensured that the feedback processes are working properly in DataONE. The problems that occurred regarding communication and how they are solved (which has been possible through healthy feedback) are discussed in Chapter 6.

5. Unpredictability

DataONE is still in the emergence phase; although, it has been two years since the funding started. It is too early to observe unpredictability in the system. However, as it can be seen in section 8 – adaptation & learning, there have been some unexpected changes and DataONE has responded to them successfully.

6. The edge of chaos

In order for emergence or self-organization to occur, the system should exist in a special environment called the “edge of chaos” or “far from equilibrium”. Self-organization or emergence happened and DataONE came to being considering the funding environment the bigger system. The data-intensive research era (the advances in data collection, storage, sharing, and analysis technologies; the acknowledgement of data problems by the scientific community; and the NSF DataNet Solicitation) prepared the right conditions to a new structure emerge, in this case DataONE. In a smaller level of analysis, the DataONE management fostered the right conditions to new structures, relationships, and products emerge inside DataONE such as WGs, academic publications, cyberinfrastructure, new collaborations, etc. The details of this special environment are discussed in the subsequent sections when adaptation and the management style of DataONE are explained.

7. Self-organization, emergence, and strange attractors

When a complex system reaches a critical point, additional energy or matter or information will cause an emergence which could be a new rule, relationship, structure, feature, etc. The system has something different now, a new component or feature or player. This is a new equilibrium. What the emergence forms around is called ‘strange attractor’.

In the DataONE case, the scientific community has reached a critical point due to the advancements in data related technologies and practices such as simulation and

modeling with high speed computers, automated data acquisition, new databases that are connected to each other. The new equilibrium, the emergence is a new way of doing science: the fourth paradigm or data-intensive research era as some scholars call it. The NSF's DataNet solicitation was the right charge to the system (as an attractor) so that new collaborations around this data-intensive research idea through funding could be formed. The background for this process is summarized in Chapter 4; thus, it is not going to be repeated here. In a nutshell, the first attractor in this study (or which led this study by letting DataONE emerge) is the DataNet solicitation. The two first emergent structures or self-organized structures are DataONE and Data Conservancy (It can be thought of the virtual collaborations or the cyberinfrastructure of the projects).

The second attractor is the severity of the problem. One of DataONE's long term goals is to support the efforts to tackle environmental problems through robust, accessible, and secure data. In the interviews, some of the participants mentioned the seriousness of the environmental problems –specifically climate change– the earth is facing and they felt that it is their responsibility to take action. This topic is especially brought up by members who for government agencies in the context of different agencies doing the same work in different times without being aware of the others' efforts, and thus, wasting resources. However, they believed that the severity of the problem cannot afford us to waste neither time nor resources. DataONE is aiming to create a single platform so that everyone could be aware of what is happening.

> "I have strong opinions about cross agency efforts minimizing duplication of efforts. I get very frustrated when I see something where USGS pays for the same thing that the NSF pays for the same thing that the Department of Energy doing that is identical to what NASA does. Just looking across some of those –even within NSF –I see or different groups do the fundamentally same thing and it is not in their perceived best interests to collaborate. And I find that waste of resources that in the context of the things like climate change and ecology can't afford."

Joining forces, creating synergy, not repetitive works but complementary works are what must be done regarding interagency efforts in the fight with the climate change problem. These themes have been repeated by the participants often.

> "The idea that we can join forces and learn about different things, like DataCite and VisTrails Scientific Workflows, that are really going to help us do our job better in the long run, that is what NASA is looking at. More ability to access other data products, learn about how other organizations operate and we might benefit from that. It is all good. It is all good."

> "So, some of it is receiving benefits from DataONE but some of it is also, you know, sort of hoping that the lessons we learn within our networks, and primarily maybe the bad things we did or the things we would do differently, DataONE would do differently and take advantage of."

In summary, because the scientific community as a complex system has reached a critical point (data-intensive research paradigm), emergent structures formed (DataONE)

through the funding from the NSF's DataNet solicitation around a severe problem, climate change.

Due to fractal expression of complex behavior, emergent properties could be observed at different levels. There are some in DataONE as well. The first one is the new relationships among institutions, subject disciplines, and individuals. Every interviewee mentioned these. Sometimes it was a relationship between a university and a government agency; sometimes it was between two distinct subject disciplines such as ecology and library & information science; and sometimes professional or personal relationships between individuals who did not know each other before DataONE. Here are some testimonials:

- "We (ORNL DAAC) have opened up our relationship with Cornell (Lab of Ornithology)."
- "And then another, of course, is the work with Cornell Lab of Ornithology. We had a planning meeting for the DataONE proposal. We were sitting together, talking about bird monitoring and analyzing the observation. When X described the analysis he was doing, I said, 'you need my data.' And so that was a connection that never would have happened otherwise. He just went bonkers with our remote sensing data and downloaded millions of our data records. So, that was another link that arose out of the DataONE connections. And to be real honest with you, I don't think NASA ever thought about using their remote sensing data for this sort of purpose so it is really pretty novel and exciting."

- "... so there is relationship with people on Data Conservancy as well."
- "We will be going out in a couple of weeks to the scientific computing, school at the University of Utah to work with some people on data visualization that all stems from the EVA working group."
- "I'm actually linked in and I've probably expanded by about an extra 80 people since joining DataONE."
- "I have also been working more closely with California Digital Curation Center."
- "That is a link (partnership with California Digital Curation Center) that we just never really had without DataONE."
- "The interactions with, for example, the ORNL and that team, are somewhat new to us."
- "So, this is kind of a new community for me to interact with-this library community."

These relationships, naturally, resulted in many products and outcomes. First, the cyberinfrastructure of DataONE is obviously the most important of all. The DataONE website is open to the public, yet at this point (May 2011) its data features are not active for public use. Second, a variety of scholarly products (such as papers, articles, posters, book chapters, and presentations) are the outcome of this fruitful collaboration. Third, one is the grant proposals leveraging the DataONE collaborators are starting to receive funding. Finally, due to DataONE's collegial and cozy environment, a network of scholars has been

established, which is the precursor of new outcomes. DataONE has proven to be quite productive and yet it is still in the emergence phase.

8. Adaptation to environment (context) / pattern recognition / learning

Adaptation happens only if the system is capable of it. Therefore, in this section first the management, the PI, data lifecycle, and the working group structure is explained to demonstrate that DataONE is organic and capable of adaptation and learning. In the second part, the changes since the inception of the project are reported under 'other changes' title to illustrate the adaptations that DataONE experienced.

a. The management

The leadership team manages DataONE; although, there are other bodies that have an influence on the leadership team. These include:

- the executive team (manages day to day task, responsible for external and internal communication),
- the external advisory board (provides strategic direction, input, and guidance),
- the working groups (different expertise groups that work towards the objectives of DataONE), and
- the NSF (the funding agency).

Everyone in the leadership team is lead or co-lead of a working group, which ensures sound communication and representation of the working groups.

The management style is neither very hierarchical nor loose; it is just in between which is the best environment for a complex adaptive system to emerge. It could be remembered from previous chapters that this state is called the edge of chaos or far from equilibrium. Structures that are overly hierarchical hinder creativity, foster status quo, and, in the long run, systems get rigid. Thus, they lose their adaptivity. The opposite of hierarchical systems, loose systems, on the other hand, are far from being efficient and productive. They do not even have the characteristics of a system and do not exist as a meaningful structure for a long time. By being in between, DataONE demonstrates the potential to be an adaptive system. The NSF is the funding agency, the goals on the grant proposal are the commitment, they are the framework/structure. The goals, the rules, the deadlines, etc. have a huge impact on local agents (members and working groups). However, the collaboration is actually a bottom-up formation or self-organization that emerged through interactions among local agents –researchers in the DataONE case. Individuals come together to form the collaboration and also the working groups, defined their goals by themselves, set the deadlines accordingly. Hence, their individual actions have an impact on the overall system (DataONE).

In a nutshell, the funding structure is a cycle. The NSF call adds energy to the system, individual researchers come together around a cyberinfrastructure vision, and the collaboration emerges. The individuals and the collaboration influence each other, which is possible through an in-between management style. The interview participants are aware of the management style and quite welcome it.

> "In some ways, it is self-organizing and I think we have some real tight deadlines and we are really focused on those. ... a lot of ideas are generated from the bottom up."

> "...it does not seem in the culture of DataONE to be a hierarchy of roles. ... We do not have a very hierarchical system. Although, if you look at our org charts, we have distribution of activities and personnel. In reality, the style of communication is very inclusive."

> "I think it is always a balance between doing too much and trying to control too much and being too loose. ... I don't know. It is good. It is a good mix of formal project management skills and also good interpersonal skills."

Here is the figure for complex adaptive systems introduced in Chapter 2 after it is applied to DataONE (see Figure 10). NSF DataNet Solicitation and the data-intensive research era are the changes in the external environment. The interactions among individual researchers such as the previous interactions in the SEEK project and the Creation of a Virtual Data Center for the Biodiversity, Ecological and Environmental Sciences Project (Interop Grant) led to the emergence of DataONE. The PI and the leadership team encouraged interaction, self-management, and creativity whereas discouraged disharmony and individuality.

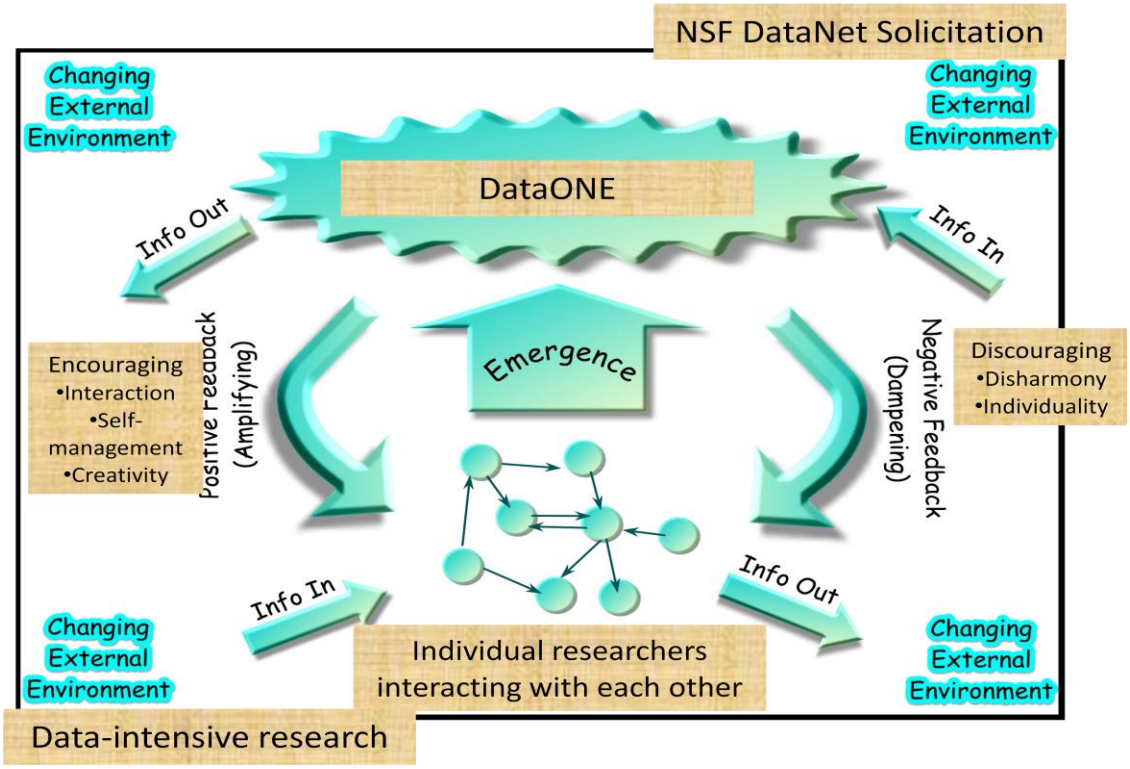


Figure 13 – DataONE adapted to complex adaptive systems figure

b. The Principal Investigator (PI)

The PI of the project fosters this productive environment through his experience and knowledge in similar projects. He is an ideal leader for complex adaptive systems. He is aware of his authority and other's capabilities. He does not issue orders like a despot but he is very observant and in control. He supports dialogue, team work, and collective work. He has created a team of experts and lets them do their job. He reminds them of certain deadlines and objectives, and provides feedback from an executive point of view when researchers get lost in impractical, ineffective discussions. His leadership skills are admired by many. People try to learn from him, imitate him.

> "I am very impressed with the way he [the PI] leads the project. A part of what I do is to watch and study what he [the PI] does and how that is effective for him and try to use that as a lesson in developing my own leadership skills and abilities to work with people."

> "[The PI] came to pick us up at the airport and that was important because he is really all about building team and right from the start he started making us ... he had us all coordinated, it already started instantly, started to build that team work thing going on."

> "He is not a dictator, not a yeller, you know, he is very quiet, he gives people a little bit of structure and lets them do their own thing. He believes in the power of multiple minds and he lets us, multiple minds to work, he does pull things together at the end but he lets people do their own thing, he brings together a great team and he lets that team work."

c. Data Lifecycle

In previous sections it is explained that the activities in DataONE are based on the data lifecycle. The eight phases here (collect, assure, describe, deposit, preserve, discover, integrate, and analyze) have become static yet the operationalization of these phases happen to be different in different tasks or among different working groups. The flexibility provided on the interpretation of each phase allows DataONE the opportunity to develop the appropriate response to the problem, whereas the static structure harmonizes the activities conducted by different parties. Therefore, by existing between order and disorder (the edge of chaos), DataONE is capable of adapting.

d. Working group structure

The management and the leadership style led to the interdependent working group structure that has been explained in previous sections. What has been left out are their flexibility, nimbleness, and ability to adapt to a changing environment. Working groups consist of experts in their fields. When they recognize an important problem, they know whether it needs to be prioritized or not, and allocate their resources accordingly. This approach allows them to deal with multi-faceted cyberinfrastructure issues. When problems are short-term, they shift to the workshop approach (federated security); some are long-term, they employ stricter membership policy (sociocultural issues); some exist in both cyberinfrastructure and community engagement side (exploration, visualization, analysis); some provide current situation with stakeholders to other working groups (usability & assessment); etc. With such freedom and flexibility, they become very resilient and effective. They adapt, evolve, and react to the changes in the internal and external environment; in other cases they get proactive and change the environment. They dissolve and emerge, there is nonstop action in them.

> "...but I think there is much more flexibility than was perhaps originally envisioned in terms of having the opportunity for short-term working groups or workshops, having these super groups with working groups collaborating together on projects and also having sub-groups. So, that concept, I do not think was there initially, but has evolved through the working styles of our individuals within the working groups."

> "Because we had multiple disciplines and multiple domains represented and we have changing needs. I mean, some projects that a working group might tackle, can be very short term, other can extend through the life of the project. So, you know, there is not really a solution that fits every type of project need."

> "We did change a couple of our working groups so they became more like a series of individual workshops. Meaning that, for example, security was one that, we did not feel we could identify a group of 8 to 10 people that could address all the various security issues for the long-term. We had one initial workshop to work out some of the federated identity and authentication type issues and recognized that some of the featured topics would require a slightly different group of people so we made that flexible in terms of being able to add in a whole new mix of people to address subsequent topics."

> "That (moving between working group & workshop structure) has probably been one of our biggest changes. ... also being more flexible with how working groups are structured so they are not necessarily all consistent membership but they can, you know, be flexible and evolve over time."

Working group structure is quite useful when there are limitations on funding. Members of DataONE are mostly volunteers and, except the four full time employees. People do get travel support for face-to-face meetings and a small honorarium, but for most that is all. If they are employed by the U.S. federal government, they receive nothing. They need an incentive and that incentive is to work with like-minded experts on the problems

they are interested in. Working group structure let individuals to pursue their research agenda and produce scholarly work as long as the agenda is parallel with DataONE objectives. During the interviews, it was mentioned that presenting a problem is one of the main incentives to make people work for DataONE.

> "Primarily, we present them with a challenge. So some groups respond nicely to being given some sort of academic or intellectual challenge or some problem that they need to solve. Other groups respond more to, say, 'I would like your expert opinion on how this work is progressing' so that works really well with things like user interface evaluations for example. When we have usability experts on a project we basically say, 'here is our template, what do you think of it?' With other groups if we approach them and say, 'We cannot solve this problem but we know you are experts in the field so we would like you to go away and think about this and engage whatever resources you can to come up with a solution to the problem that works within the context of DataONE'."

The concept of secondary or more working group membership is also another indicator of nimble structure. With that opportunity, members could work on topics that are interesting for them or join forces on ad hoc problems without feeling remorse or guilt. The survey showed that members do have secondary working groups (n=46). Here is the breakdown of the secondary working group memberships.

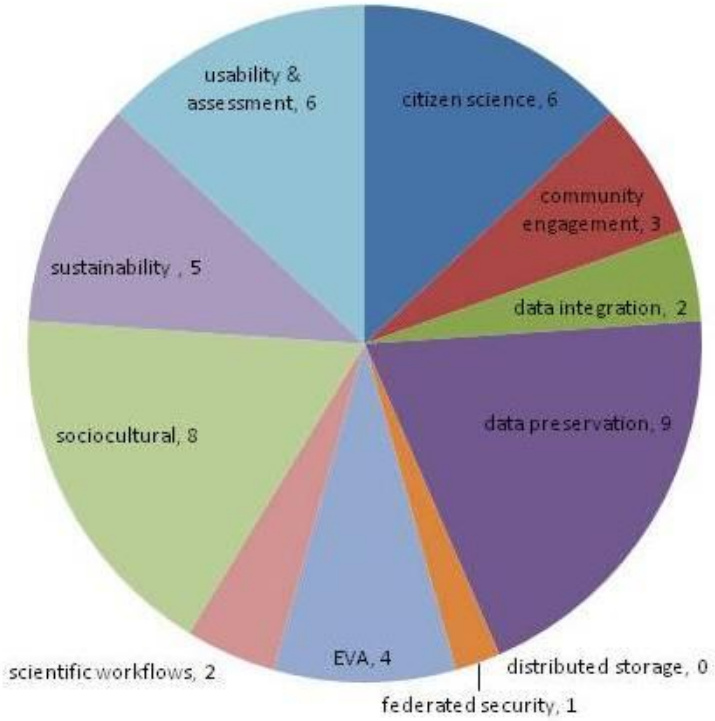


Figure 14 – Secondary working group membership according to the survey

This approach is feasible in given conditions of DataONE. In one case, there is not a working group for one problem and in the end it led to the establishment of one. The exploration, visualization, and analysis working group was not even planned. However, when the opportunity met the need/problem, DataONE took action and created the exploration, visualization, and analysis working group. When some members met X from the Cornell Lab of Ornithology, who prepares simulations, exploiting the data and simulation that is available seemed to be a very good idea. They proved to be right. The simulations that use data from DataONE increased the promotional power of DataONE and it was decided to have a permanent working group focusing on the topic.

> "...because we needed something that could be shown, we needed something that people could visualize, data isn't sexy but what we can do with data, we needed to get that sexy part, we needed to get the part that looks good."

> "There have been a few side trips that were not planned. For example, our one fabulous example, the EVA working group that did the bird work, that was not really planned that way. That one was sort of a great idea that got rolling –a wonderful group of people, they generated so much enthusiasm and actually more money to keep on going. So it is our poster child."

> "That was the exploration and visualization analysis working group so that we could more heavily engage domain scientists and hoping to set direction for DataONE."

The interaction among working groups is explained in the previous sections; thus, it is not repeated here; however, here, the establishment of an Ad Hoc Group has to be mentioned. A need, a reaction to the environment resulted in a temporary working group formation. When the conditions were fixed, the Ad Hoc Group disbanded itself and its tasks were transferred to the working group (usability and assessment) which was originally responsible for them. Ad Hoc Group, for instance, conducted the baseline assessment study for scientists to examine the data sharing and preservation behaviors of scientists (Tenopir et al., 2011). To sum it up, when the need arises, working groups join forces, create a new working group, or come up with a new approach to deal with it because they have the necessary skills and environment/opportunity given to them.

e. Other Changes

Working groups are not the only change that has happened in DataONE. On an individual level, the participants expressed that their role and time commitment have changed unexpectedly. In some cases, the individual became a working group lead or become responsible for new tasks; in some cases the time committed increased or decreased. For instance, the reply to the questions "has your role in DataONE changed?" is as follows:

> "Yeah, in the original proposal I was not lead of a working group. I was sort of a PI and helping out in a bunch of different areas. This was a cool thing to be able to do, to jump in. I was helping out in many, many different areas and then focus [on EVA], so that was cool."

Another member describes how s/he got into the leadership team.

> "… I personally got more involved than what I originally imagined or intended … So my role in it has become bigger than I expected it, I never thought that I would have time to commit to be on a leadership team for example.

Another important change is hiring an executive director. Due to the change in the role of PI (he had to become the PI of another project as well), a need for an executive director has emerged so that day to day overseeing of the project could be done. The process is described as such by a participant:

> "We didn't have any second director written in and that has changed because as the organization morphed and as we got more structured …

> we realized that is something that we needed to add and we changed, we had called him assistant director, we changed that into directors for the two groups, infrastructure and community engagement and over time changed what the responsibilities for those would be a little bit…"

There have changes in the environment as well. The management who is responsible for the NSF changed. There had been some uncertainties and delays during the transition period which were solved rapidly. The real change happened in the funding environment. The financial crisis in 2008 hit research and development funding all over the world including the U.S. Originally there were supposed to be five DataNet projects; however, after the crisis it was decided to have two (DataONE and Data Conservancy). Fewer DataNets means, there are fewer opportunities to interact with other DataNets. The importance of cross-disciplinary interaction should be obvious by now. DataONE and Data Conservancy are deprived from such interaction and limited to each other. The impact on DataONE is the issue of sustainability. Although, it is too early to comment on that, the leadership team has some worries and has started to think about it.

> "Probably, from a financial point of view, DataONE is supposed to be self-sustaining into the future and one strategy for making that happen is with donations and so forth from different groups. That source of funding is certainly dramatically impacted by the financial situation of the company and the economic situation of the country. So, that's caused some change but I would not say it was a major change."

In addition, the DataONE Business Plan has been created "to build the capacity to preserve DataONE content and services and to increase their value to the user community over time" (NSF DataONE Progress Report, 2011).

The change in the organizational structure can be seen in the organizational charts as well. In the previous organizational chart (Figure) the leadership team and the PI is not included. DataONE Office is not envisioned, it was established at the UNM later in 2010. Furthermore, as it was reported earlier, the Exploration, Visualization, and Analysis working group does not exist. These changes demonstrate how organic DataONE is and adapts to the changes and needs so that it can perform better.

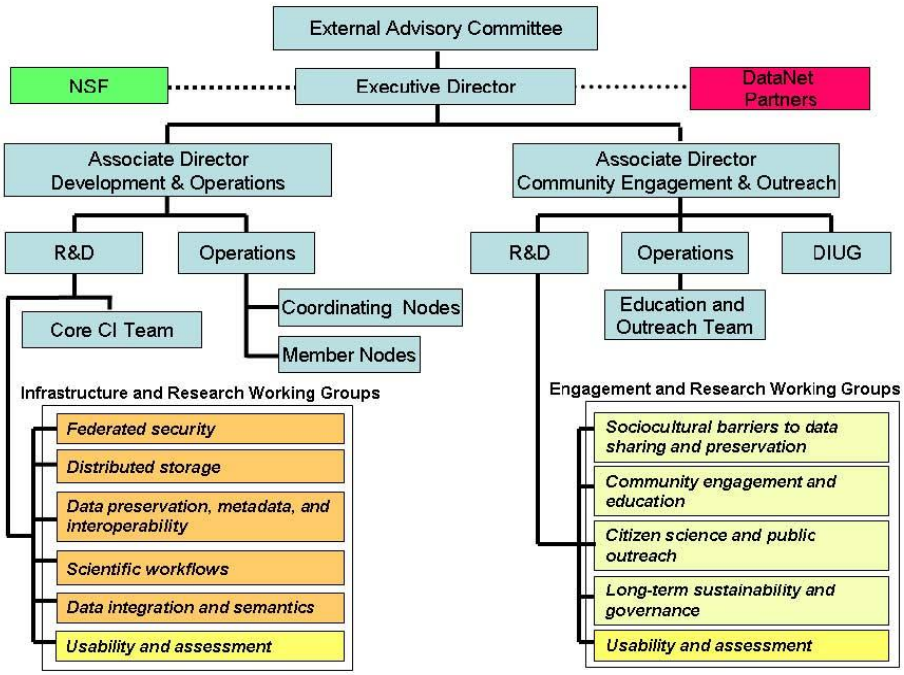


Figure 15 – Organization chart that was submitted in the grant proposal as of 2009

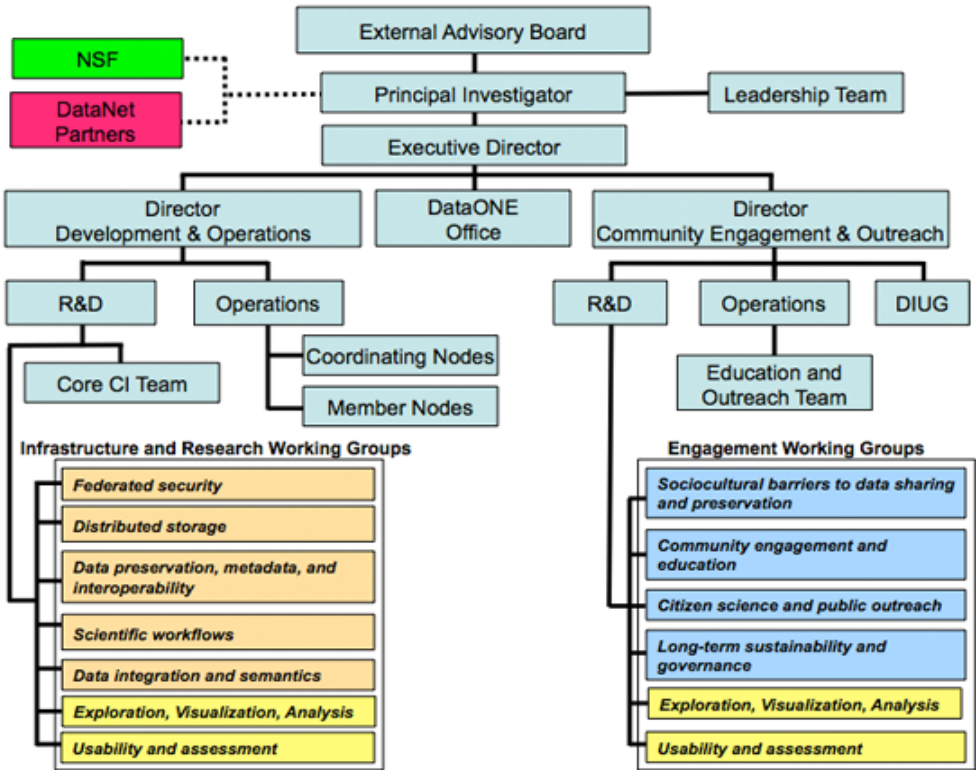


Figure 16 – Organization chart for DataONE as of February 2011

In summary, it is too early to observe big shifts and changes in DataONE's relatively short life so far; the project is still in its emergence phase. However, there have been some changes and DataONE reacted to them quite well and adapted to the new conditions. The structure that is formed is quite flexible to accommodate such modifications.

9. Historicity and path-dependence

Historicity means the consequences of past actions result in current events. Although there is not enough evidence to make such a conclusion, some of the interviewees

mentioned previous cyberinfrastructure projects that in a way led to DataONE. Two projects step forward: SEEK (the Science Environment for Ecological Knowledge) and the Interop Grant (Creation of a Virtual Data Center for the Biodiversity, Ecological and Environmental Sciences). One of the participants expressed that SEEK also interested in data integration for ecological sciences; however, for the time being its goals were 'ambitious' and could not be accomplished yet there were some promising results, provided some useful tools and lessons, and a following project might succeed.

> "So, before this project, I was involved in another projected called SEEK, which is the Science Environment for Ecological Knowledge, and DataONE kind of flowed on naturally from the work that was done on that project. In fact, there are several participants from the project, which are core people in DataONE as well. … Basically, the SEEK project was looking at essentially the problem of data integration to biological sciences and ecological sciences. And, it had fairly ambitious goals, and there was also a lot of research work involved in that project. … So ambitious that we did not reach many of them. I will put it that way, after five years of the project (we did not reach). So what we did figure out was how to do, sort of, this low level integration between data repositories and what was really required to make that happen. And so, that experience really helped drive the development of the DataONE proposal. So, there was a lot of background that came out of that project that sort of flowed directly into the overall architectural design and even the sort of day to day activities of that communication and so forth of the project participants."

The Interop Grant (Creation of a Virtual Data Center for the Biodiversity, Ecological and Environmental Sciences) was the second project. It was prepared and implemented by some of the core members of DataONE. During the preparation, the basic features of DataONE had been realized. Later, the NSF's DataNet solicitation had been announced. It was a good match. That proposal was modified heavily yet it was undeniably the basis for DataONE. The PI summarizes that work:

> "Well it came about as a result of thinking through another proposal for NSF called interoperability. And, this one was also related to building interoperability solutions that would enable more readily transparent data sharing and data use for the environmental sciences. And, in thinking about that project, it became clear to me that we needed a full-blown data center, federated data center, like DataONE, … to help make it available to as broad as possible community environmental data tools."

Another participant remembers the process in more detail:

> "…NSF did a solicitation called interoperability and so we all joined forces and wrote a virtual data center proposal and it was funded, although at a low level. It included M, of whom I knew, it included N, who I did not know at that point. It included the guys at Nescent, P and others. So, we just started building on the ideas from this group that met in Santa Barbara in 2005. Then, we got funded from NSF for the Virtual Data Center proposal. Shortly after that activity was funded, they sent a call for DataNets."

Path-dependence is limiting the options over time that the system has and being obliged to one option. The fragmented structure for repository purposes that resulted from the conflict between the cyberinfrastructure team and community engagement & education team can be considered as one. In the beginning all options were possible to store data. However, over time, due to the expectations, habits, and culture, the options were eliminated except two: Subversion and the plone site. Ironically, a collaboration that is formed to solve long-term data issues has a data issue of its own. The detailed discussion of the problem is presented in Chapter 6 but the problem is, briefly, the discrepancy/conflict between the cyberinfrastructure team and community engagement team on how to communicate and exchange documents. The result is a fragmented structure in which the cyberinfrastructure team uses Subversion and the community engagement & education team uses the plone website. Some consider this an important challenge that the collaboration experiences.

> "That is one of the challenges the organization faces. On one hand, I think it would be valuable for everyone to share the same type of interface or system for sharing material. But, I know that CI do enjoy using subversion and CE do enjoy using plone. So, the potential for one group or other to use the other system might be a hurdle to overcome. So, it is whether that hurdle encourages people, on either side, encouraging people to use something that is not their first response or first nature –may be more of a challenge than just having two systems operational. I think that is something we need to think about as more and more material is produced in moving forward. Because, one of the

> challenges of having two systems, or repositories of documents we share, is duplication and also non-conformity between the two. So there might be some things that CI has put into plone to share with CE and they have been updated on Subversion but have not been updated on plone for example. So, that is something we really need to be mindful of and to find some sort of resolution for if we are going to maintain two different systems."

However, not everyone agrees that the fragmented structure was a serious problem but not anymore because of the better linkages between the two systems.

> "Yeah. I won't say it is as fragmented as it was. You know, there is the plone DataONE website and pretty much now all of the documents are being managed through it. There is still Subversion stuff that is getting used, but primarily for a lot of the code and stuff or architecture-type documents. There is better linkages between those repositories. But at one point, there really was not good linkage. The reason I think it caused some issues was people sort of manually had to deposit documents in both places, which, of course, they are not going to do. They do not have time, you know, and it was repetitive. That has sort of been resolved to the most extent to be honest with you."

10. Coevolution

Since DataONE is still in its emergence phase, it is too early to observe co-evolutions in other systems. However, there have already been some changes. One of them is the data management plan requirement for project proposals to the NSF. As of fall 2010, all of the project proposals that are submitted to DataONE must have a data management

plan(NSF, 2010). This study is not in a position to make an argument regarding causality or precedence between the new regulation and DataONE. It simply acknowledges the fact that they both exist. In the long run, more interaction between the scientific community and DataONE on this data management plan dimension is expected.

Another co-evolution potential results from the interaction between DataONE and Data Conservancy, the other DataNet project which is receiving $20 million in five years like DataONE. Although Data Conservancy is first targeting astronomy data, there has been constant communication between the two projects reported by the participants of the study. Members from each project attended the meetings of the other; as a result "the Data Conservancy has agreed in principle to act as a DataONE Member Node, and DataONE as archival store for Data Conservancy" (NSF DataONE Progress Report, 2011). Moreover, these two projects are the role models for smaller scale projects that receive funding from the same solicitation. Thus, it is quite likely that they follow the work happening in and publications from DataONE and adjust themselves accordingly. They are already or will be repositioning themselves. The system is co-evolving.

The potential area for co-evolution is the scientific community at large, assuming that DataONE will be successful. DataONE's third goal is to engage the scientific community and change the scientific culture to a culture of data sharing. This ambitious goal signals a variety of changes in a variety of fields. The formation of the exploration, visualization, and analysis working group could be interpreted as an impact on the scientific community. The researchers at the Cornell Lab of Ornithology realized the

potential of accessing huge databases through DataONE and decided to be involved in the project. Some projects that address issues of long term data management, reuse, discovery, and integration have been identified for future collaborations: Filtered Push, the Scientific Observations Network and the Semantic Tools for Ecological Data Management (SONet/SemTools), TeraGrid, Federation of Earth Science Information Partners (ESIP) are to name a few. However, it is too early to discuss the impacts as DataONE is still in its emergence phase.

In summary, DataONE as a scientific collaboration proves to be an organization that operates according to complex adaptive systems theory. It has the necessary elements, the relationship among these elements, and a structure and environment that nourished nonlinear relationships. As a result, DataONE is an emergent structure. It is able to learn and adapt. Finally it shows promise to have an impact on other systems.

Chapter 6

Additional Results Regarding Library & Information Science and Communication Studies

The results in this chapter are of interest to scholars who do research in information science and communication studies. The data collected for this study reaches beyond the framework that is developed to assess the complexity and adaptivity of a scientific collaboration. They are reported here.

1. Library and Information Science

It has been mentioned that the library and information science component was added to the project after feedback provided by the NSF during the grant proposal writing. The University of California Digital Library, the University of Tennessee Library and School of Information Science, and the University of Illinois-Chicago Library have been heavily involved in DataONE since early on. The role of library and information science scholars includes engaging the community by providing training on data issues, converting libraries to digital repositories, work on digital object identifiers, developing assessments, etc. It was an interdisciplinary connection that had not been thought of before. In fact, some of the participants were not aware of the services that library and information scientists could offer. Participants summarized it as such:

> "...but early on into the project it was clear that NSF expected a significant involvement of what we could loosely call library science community in the DataNet partners. At the time when I got involved in this we really did not have a strong library science partner in the organization, in the proposal team."

> "The other thing is the idea of the library community. I never really realized what they are up to, to be honest with you. I never knew."

The contribution that library and information science professionals provide falls on both sides of DataONE activities. On the cyberinfrastructure side, the libraries operate as member nodes for storing and providing access to data. For some, being responsible for data should be the future mission for libraries.

> "I think, one of the things I think is so exciting about it is the opportunity to work with people in a library background. This is the first time that I have done that. What I see for DataONE is such a wonderful opportunity for libraries in the future. I think we are moving away from books and libraries are going to need a new mission. I think being responsible for data is an excellent mission and they really have a fantastic background for this."

In addition, they provide both knowledge and also network for digital object identifiers[14] in this matter. The extent of library community's contribution on DOI had not been known before among the non-library members and the partnership with California Digital Library proved to be quite important.

> "And another one is with the California Digital Library and they are working on digital object identifiers. We sort of jumped into this four or five years ago. We decided to add DOI's and we are going to move forward with this without really knowing, what the rest of the community was doing. I should say … that there is a Data Cite group that recently formed dealing with digital identifiers for data sets. So that was something that was totally new to us and I think that those folks appreciate what we are doing and where we are and we appreciate how they are leading the field forward and we want to go with them. That is a link that we just never really had without DataONE."

As for community engagement, they are providing training to different audiences. The training has not yet started for all of the stakeholders, DataONE has not yet become

---

[14] Digital Object Identifiers (doi): A kind of tag that helps to identify an electronic object (a physical, digital, or abstract entity) in a digital environment. "The DOI system provides identifiers which are persistent, unique, resolvable, and interoperable and so useful for management of content on digital networks in automated and controlled ways" (Paskin, 2010, p.1586).

public as of May 2011. However, there have been some early activities regarding students: a summer internship program and coordinating science links[2] students with IMLS funded data. In the summer internship program, students work with mentors from DataONE on issues that are related to DataONE's goals such as data management, environmental data in the classroom, data lifecycle, data science, programming, and developing animations. Science data students are engaging in some DataONE activities as part of their science information program.

> "At UT [the University of Tennessee] we are looking at bringing in some students, who will learn science data, we call them science data students. Hopefully what they will do is be able to help us prepare data for the archives –like documentation, quality checks and things like that. Having the opportunity of going and teaching some of our ideas and practices is a collaboration that really strong."

Another activity for library and information science students and professionals is the Environmental Information Management Institute which is going to take place in summer 2011 at the University of New Mexico (UNM) sponsored by both DataONE and UNM (DataONE, 2011). The courses will provide the conceptual and practical hands-on training that allows the participants "to effectively design, manage, analyze, visualize, and preserve data and information."

The role of libraries and librarian will grow when the community engagement and education activities take off. Since the product –DataONE cyberinfrastructure– is not fully ready, there is nothing to promote. Moreover, the baseline assessment for libraries and

librarians has not been concluded. These assessments will tell the current conditions and afterwards it will be possible to create and implement right strategy to mobilize the resources so that the involvement of library and information science component can be fully reflected in DataONE.

> "I still think the role of the library is somewhat untapped. I think right now some of the library participation has been through the expertise of maybe the technical information people, you know, in terms of doing assessments or metadata or some of the training. But, not so much in terms of working specifically in a library to figure out how they manage their data and how they can use the services of DataONE. I have not seen that directly in the project and that has to occur for that to be successful. … I think when … we get assessments from those groups, that will push, "we need to do this, we need to do that," or "here is the current practices" within those libraries, if you will. So, I think that will help drive even more participation involvement and things like that of libraries and librarians to DataONE. So, it is just sort of a phasing thing to some extent."

The participants are also asked what kind of information they need regarding DataONE matters (n=51). Scientific information ranked first (24), followed by technical information (16). It seems that in order to conduct daily tasks legal (1) and financial (0) information are not needed very much. The participants were also asked which channels they use to seek related information (see Figure 12). Email is the leading one on both scientific (33) and technical (30) information. The second place to look for information regarding DataONE related matters is the plone website. For technical information 19

people and scientific information 17 people expressed that they visit the website. Virtual media such as Skype, IRC, Maratech is popular and wikis are referenced quite often.

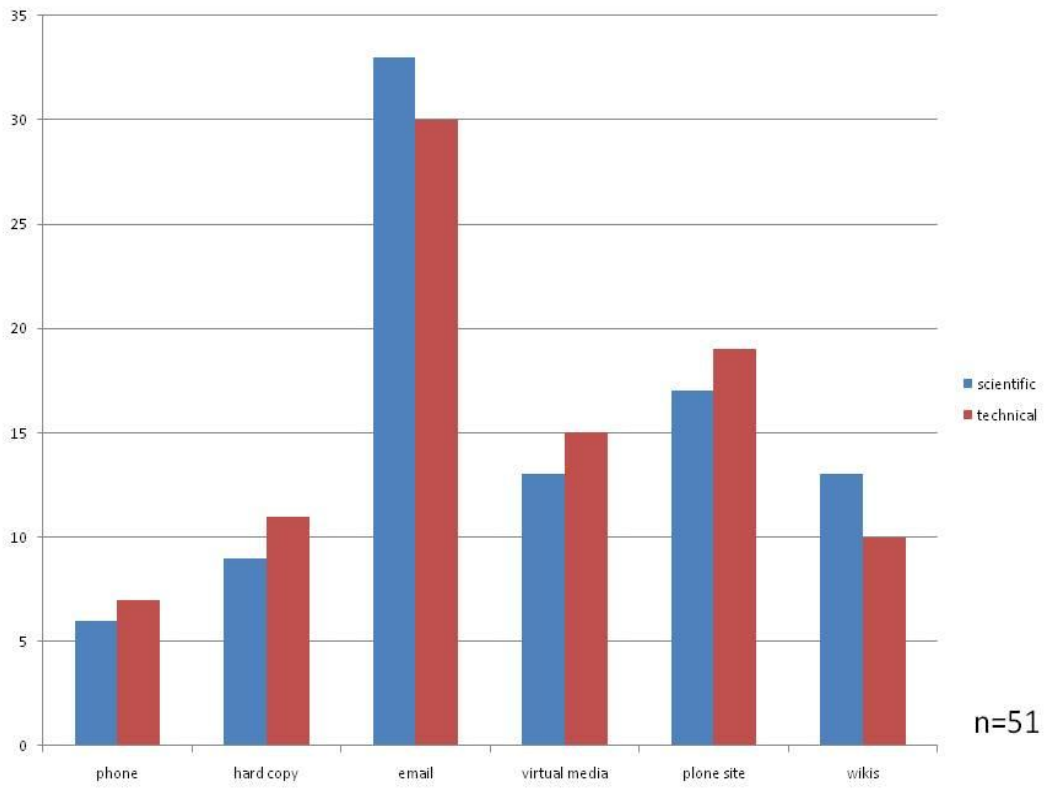


Figure 17 – Information channels used to seek information regarding DataONE matters

2. Communication Studies

Communication is crucial for the success of DataONE (and of course for any other scientific collaboration). The online survey revealed some information about the frequency

of communication among the working group members (see Figure 13). Nearly a quarter of participants (24%, n=42) communicate with their own working groups members weekly or more frequently and one fifth of the participants (19%, n=43) with other working group members. As for communicating with their own working group leader, 40% (n=40) of the respondents expressed that they communicate weekly or more frequently, nearly one third (30%, n=40) expressed monthly or more frequently, and more than a quarter third (28%, n=40) expressed less than monthly. These results are consistent with the interview results that indicate the frequent interaction among individuals and working groups.

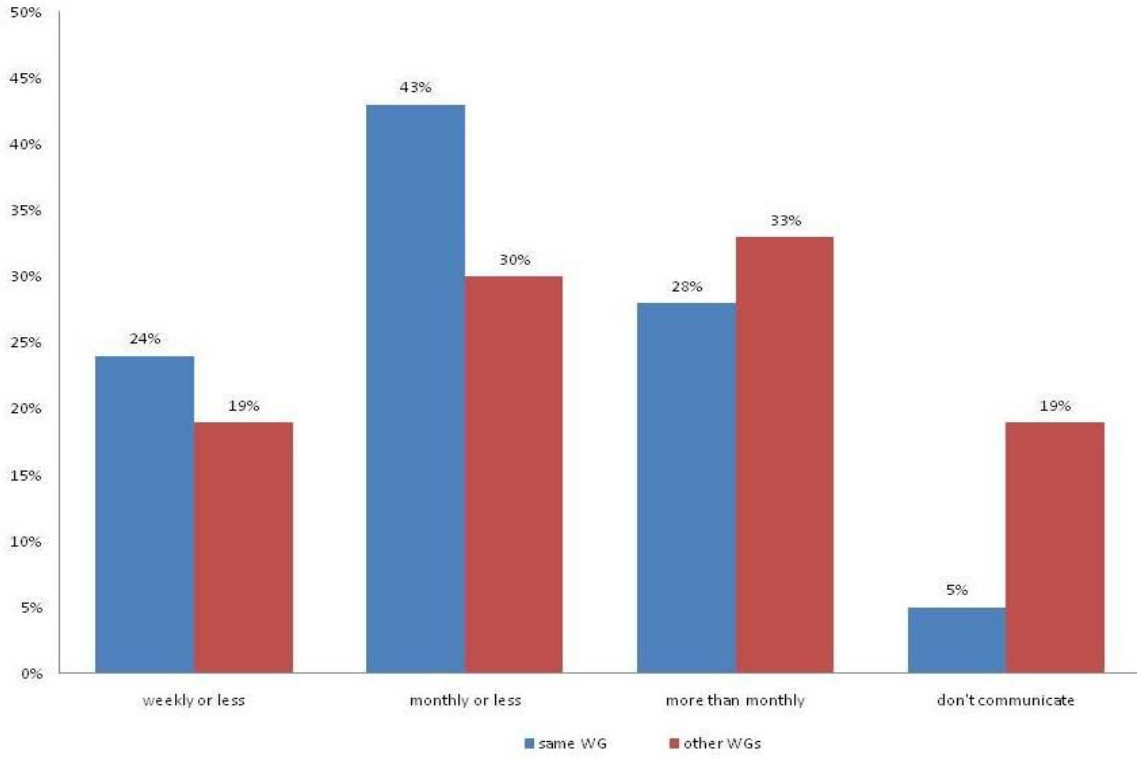


Figure 18 – Communicating with own working group members and other working group members

However, communication has its challenges mainly because of the diversity and the structure (being a virtual organization) of DataONE. Communication in a virtual organization is mostly mediated with the help of computers because the members live in different places. The effectiveness relies on the software capabilities used to communicate. In addition, they live in different time zones, which also cause problems. These problems have some solutions to a degree; however, problems that are caused by diversity are harder to handle. People from different disciplines have different terminology and different workflow. Coordinating and harmonizing their activities have become a challenge for DataONE. In this section, based on the interviews with the leadership team, the problems and responses to them are reported in four subsections: geographical location, software capabilities, the divide between cyberinfrastructure team and the rest, and intimidation.

1. Geography

DataONE consists of carefully selected individuals who have experience and expertise in many different fields which is discussed in the sections above. There is not one single location, institution, or city that could have all of them as residents. Some institutions have better computer scientists, some people have better relations with the scientific community, some are more experienced in dealing with funding agencies, etc. Simply, there is not one place that could supply such a high quality human resource. Thus, DataONE is a virtual organization, and members (including the external advisory board) are all over the world –even in the emergence phase. In such a widespread group having

some communication and coordinating issues is only natural. These issues are a tradeoff between having the right people and having fewer issues.

The first problem is the most obvious one. People are located in different time zones. At this point, almost all of the working group members are in North America[15] which means a four-hour difference in time at most; however, the external advisory board has members in the UK and in Australia. Having virtual meetings with everyone at the same time is not easy, yet there is a cumbersome solution: having two meeting with two different time zone clusters. The executive director describes the process as such:

> "The one area that we have a little difficulty is that our external advisory committee has people in Pacific, the pacific time zone in the US, central time zone, eastern time zone, somebody is in the UK and someone in Australia. What we have had to do is, we do face-to-face meetings so everybody can come but we also have to do two phone conferences just to hit that many time zones."

2. Software capabilities

DataONE is a virtual organization; thus, almost all communication activities are facilitated through computers such as emails, videoconferencing, chat, document sharing, etc. The capabilities of such software are crucial to the well-being of the project as the

---

[15] There are a couple of people in the UK.

importance of communication/interaction in complex systems was explained in previous sections. Even though most of the participants are satisfied with the abilities of the software used, several software limitations were mentioned in the interviews with the leadership team.

The first one is related to videoconferencing tools. The team tried some commercial software and then decided to use the open source system: Maratech[16]. Even though technology has advanced and web-conference has become widespread, and there are many freeware solutions available, none of the software tried and used were the perfect solution. One participant believed that what is provided to consumers is far from what can be provided because the technology is available and defined it as 'pathetic'.

> "The video conferencing solutions that are available today are still, in many ways, kind of pathetic with respect to what is possible on the internet versus what is actually delivered."

> "We have the technology today to do it if somebody would actually put together a good software package for it. Like, I have been disappointed in the quality of the software packages for, you know, remote meetings. They are just not that great."

---

[16] Maratech is a web-conferencing tool developed by a Swedish company. It was later bought by Google in 2007 and its services were disbanded. A version is still being hosted at NCEAS servers.

The problem with videoconferencing tools is that there is not a package that serves all. For instance, they do not come with a document sharing component –which has to be brought in additionally to supplement these tools. DataONE has around 100 active members who produce a variety of documents and software codes that needs to be shared, edited, and rewritten over and over again. In order not to lose track of different versions an effective document sharing and collaborative working space is needed yet the available software does not meet the standards of everyone; "they [the software] are not that rich." For instance, etherpad is widely used among DataONE members; however, the cyberinfrastructure team is not fully satisfied with it because, although they can collaboratively work on a code from different locations, they are "out of (their software) developing environment." Googledocs is found to be "updating slowly" and "the whiteboard tools and stuff that come with the ICT projects are really clunky to use."

Given the increase in the number of virtual organizations, tools that facilitate collaborative working, especially in the areas of communication and information flow, have gained importance. The interviews revealed that there is room for improvement to increase efficiency and productivity of virtual teams.

> "So, there is a lot of detail in the technical areas that could be tied into these collaboration tools that would really improve productivity of the distributed development team."

3. Cyberinfrastructure team vs. Community engagement team

This problem was not an easy one to deal with unfortunately. In every interview with the leadership team, the conflict between the cyberinfrastructure team and community engagement team was brought up. The problem was in order to keep track of different versions of documents the cyberinfrastructure team introduced Subversion and the plone website, which was found cumbersome and impractical by the community engagement team. Subversion is popular among software developers; it keeps different versions of software codes automatically. The plone website is more like a shared space or internal website to post documents and presentations. A big advantage of the plone website is to share the files that cannot be sent through email because of their size. However, people on the community engagement team were not familiar with either of these methods and they tended to rely on email. They find the methods promoted by cyberinfrastructure 'too complicated', 'arcane' and irrelevant to their job. An interviewee expresses the situation as such:

> "There is a huge split between the community engagement and technical working groups. Technical working groups want everything on the official sites, the ticket system, the plone site, etc. And the community engagement likes to do everything via e-mail. Huge, huge split!"

A member of the cyberinfrastructure team, who identifies himself as a bridge between two sides came to one of the community engagement & education/outreach

meetings to introduce these systems and show how to use it; however, it was not well received.

> "We had a little bit of a rebellion there … everybody said 'forget it, we are not going to do it' –we said it nicely."

> "What they (CI) did not take into account is you have to be a real geek to enjoy that and to be able to actually do it."

The cyberinfrastructure team was not happy as well. Senior researchers had been using Subversion for so many years, it had proved to be a very practical tool and successful.

> "Unfortunately, I do not think a lot of people on the CE side have done that. They are use to working in much smaller teams, teams of less than 4 or 5 people, where they are largely in control of the project. They are not really dependant on the work of 10's or 100's of additional people. So, they are not used to the idea that they need to report on what they are doing, not just on the broad strokes of what they are doing, but on all the details of what they are doing. … So, you know, it is a cultural difference between the two sides of the project. I think the engineering, CI side, is much more amenable to that because I think it is hard to manage any reasonably sized software endeavor without it. So, they are just used to it. The other side has not seen the light."

It can be remembered that one of the criterion for recruitment was 'willing to compromise.' In this conflict, both parties had to compromise to find a solution and leave their comfort zone a little bit. The compromise was a fragmented structure. The community

engagement team is learning and using plone website because it is much more reliable and organized repository system than email whereas cyberinfrastructure team is using Subversion since the features of it are quite indispensible to their work.

> "I know that CI do enjoy using subversion and CE do enjoy using plone."

After long discussions, both parties understood the needs of each other and they developed empathy for the other. A cyberinfrastructure team member explains why Subversion is not liked by community engagement team and a community engagement team member expresses the value of it for cyberinfrastructure team.

> "…others felt that (Subversion) was too burdensome. I think that is totally understandable." (CI team member)

> "…you have got to have versioning system, you have got to make sure that everybody can get to version 1, version 1.1, version 1.2,… and it is very important to have it in one place all" (CE team member)

It is expected to have some conflicts with such a diversified group –people with different disciplinary and professional backgrounds with different agendas. However, picking up the right people –who are good communicators, willing to compromise, have a good reputation in teamwork– to work in DataONE and the interdependent nature of tasks to accomplish for DataONE paid off. A solution was inevitable and both parties met halfway and overcame the problem.

Another important element in resolving such problems is the involvement of bridging members –people who are able to ensure sound communication between two groups. The interviews revealed two such members (one even identified himself as a bridge).

> "I am very cognizant of cultural differences that arise between disciplines. … I also see myself as one of the bridges between the cyberinfrastructure and community engagement side of things."

It is important to have a common understanding in a system in order to have a collective behavior. The variation among members also results in a different jargon, which is not surprising because different disciplines focus on different concepts or name them differently. DataONE is developing 'personas' to define the needs of various users. The outcome will be not only stereotypes of users and services but also a reference for vocabulary that is shared by everyone. Furthermore, this is a process not an end product. Throughout the process open communication will help to explore and discuss the differences among different parties and in the long run might result in a common understanding.

> "So, it might not be a common vocabulary, but when you have something in writing that everyone can refer to that depicts a particular individual, a particular type of user, also, extending that to a particular process. … ultimately, we might not be able to have a completely shared vocabulary, but we can have the same concept behind the language we are using, provided we are aware of the language in the

> other areas and other domains, then we can move forward. I think that the way this is done is: A) documenting the process and developing materials that can be read and shared across all groups, but B) having a lot of opportunity for discussion and clarification. It is during this process of presenting materials to one other that we are able to ask our questions and get that understanding. So, I think that dialog is essential, and that time, for the question/answer is essential for this common understanding."

In this quote, the importance of dialogue, to be able to ask questions without reservations, and communication is emphasized in detail.

4. Intimidation

The problem of intimidation was experienced only among people who come from library and information science discipline and easily resolved. The problem is caused by the etiquette of 'hard science' or 'computer science.' These LIS people felt intimated at their very first meetings by just being in the presence of these people, although these people had expressed no intimidating behavior. It was just the etiquette.

> "I will just be truthful and say that I thought the group of people that had been gathered together to work on this was overall such high quality people, you know, they have all achieved so much , that it was mildly intimidating, to be involved with that crowd at the very beginning."

> "It was a little bit intimidating at first. Especially everyone that is involved, sometimes I get the feeling that everybody who is involved,

> except for me, understand a lot more than I do, especially again sort of on the science side of things and on digital preservation side of things."

> "Well, I was definitely very quiet the first day because I was just taking everything in, learning everybody's names, figuring out who did what because some of the people came totally from the hard sciences, other people came from the mix of hard sciences and computer science. I was the only librarian in the room, Information Science person, so part of it was there in terms of, gosh; it is a really tough question. It was scary to be the only person…"

However, these people are really good communicators and the PI creates a very welcoming environment and encourages participation and dialogue. The computer or hard science people are good explaining what they do, what the problems are, and what the possible actions are into non-technical people, in this case to library and information science people.

> "… the people that are involved, that I have interacted with, are from such different backgrounds and yet each one of them has an ability to, you know, speak to someone outside of their discipline in such a way that you understand what they are talking about and yet you don't feel like, you know, you don't know anything about their discipline. They have very good ways of communicating and teaching what they know best."

> "Yeah, I think I have a hard time keeping up with the jargon and CI. But, the people on the CI side of the project are very sensitive to that and they truly want to communicate, and can communicate, so it is very

> easy with these particular individuals to say, “ I don’t understand,” and they can explain it (laughing) in words that I understand.”

The PI acknowledges the fact that DataONE is a multidisciplinary project and still growing which means in every new face-to-face meeting there are new faces with different backgrounds who have to be brought up to speed. Thus, accommodating their questions by creating a participatory environment is a priority. In addition, working group leaders have gone through facilitation training. The PI and the former members together invite everybody to join in discussions. The PI sees this as an important component of DataONE’s communication strategy.

> “… we needed to lay out on the table and make sure we were clear about and also make people feel comfortable asking questions when they don’t understand where someone is coming from. So, that is sort of all been, we try to make that ingrained in our approach for communication anyway.”

The researcher also experienced such an attitude during his observations many times. More than once, he was invited to join in discussions and express his opinions, even though he mentioned that he is not a participant but an observer. An interviewee who witnessed one of the incidents remembers it during the interview and refers to it.

> “They made un-scary it very quick, it was like ‘we really value everybody’s opinion’. The way the meeting was run, everybody had a chance to say something in terms of the way…it is hard to explain…but

> it was like, you were not allowed just to sit there and say nothing, I think you experienced that."

In conclusion, DataONE requires frequent communication among its members; however, there are challenges due to being a virtual organization and diversified member structure. These challenges were expected. In order to overcome them smoothly, the members were selected according to their communication skills and experience in working such environments (the details of selection process is discussed in previous chapter). It paid off. Highly qualified and experienced DataONE teams developed solutions that made DataONE so far one of the most successful projects[17].

3. Bridging Role

One theme that emerged from the interviews in dealing with cultural differences is that a couple of members are identified as a bridge between the 'computer technical folks' (cyberinfrastructure team) and community engagement team. These roles are not given to these members but they see the need and given their skills, experience, and desire, they take on the bridging role. These people have a mixed educational and professional background that would help them in this intermediary role.

---

[17] The success of DataONE so far is discussed in Chapter 7.

> "I am working in two disciplines, or even three disciplines in which I have zero academic training. I am very cognizant of cultural differences that arise between disciplines…. . I think it is important to bridge all of these kinds of cultural issues among academic disciplines."

> "I am sort of an IT person but a manager too so I to some extent crossed both camps. You know, I will participate in some of the technical working groups but then some of the management like sustainability and governance group too. So, that was interesting to see that dynamic of which tools, which groups are more comfortable with and stuff."

During the conflict between the cyberinfrastructure team and community engagement team on what to use for communication and repository, one of them volunteered to demonstrate the software. Although, it did not work and the community engagement team decided not use the Subversion, the involvement of someone who can speak for both sides is a crucial advantage in not only dealing with conflicts but taking care of daily tasks in the project.

> "For a successful project and in a successful organization one needs a combination of people who are deep technical experts within their given area and one also needs somebody who can speak the language of a broad range of experts. That has historically been one of the roles that I have had in projects because that is something that appeals to my personality and matches up with some of my skills. So I can talk to V and P about the issues in development of the communications plans and the communication strategy, understand some of the issues in

> educational approaches in engagement, and then turn around and talk to the developers about details of communication protocols and so forth. And I don't understand any one of them to the level that those particular experts do, but I understand enough of what they do that I can translate the language from one to another. "

Every member is communicating with others –some more, some less. Working group leads are generally more that the members as they being the hubs. The bridging people are also hubs and they serve as translators.

Chapter 7

Conclusion & Discussion

In this chapter, first the summary of the study is presented. Second, a brief assessment of DataONE is reported through the reflections from the NSF 2nd year review. The third section describes the contributions of this study, focusing on developing a complexity framework and applying it to virtual scientific collaborations. The fourth section is a discussion about multidisciplinarity and the potential of communication studies and information science in scientific collaborations. Future study ideas are discussed at the end of subsections in italic whenever needed.

1. Summary

The current study has explored the emergence of DataONE; a multidisciplinary, multi-institutional, multinational scientific virtual collaboration that aims to provide access and storage for earth sciences data. Briefly, findings of this study reveal that, DataONE behaves like a complex adaptive system: various individuals and institutions interacting, adapting, and coevolving to achieve their own and common goals; during the process new structures, relationships, and products emerge.

The literature on scientific collaborations is rich; however, systemic studies that could produce generalizable or comparable findings and studies that treat scientific collaborations as CAS are lacking. Furthermore their relevance to the focus of this study is limited for two reasons. The first reason is the recent developments in information and

communication technologies that have changed the rules of the game. The ease of information sharing and communication has given birth to a new type of collaboration: virtual collaborations. The number of studies examining virtual collaborations is increasing, but more needs to be done. The second reason is a new paradigm that is used in many disciplines to explain the relationships among components and systems: complex adaptive systems theory. Even though, complex adaptive systems theory is used in organizational studies, its applications to scientific collaborations or virtual collaborations have been very limited.

Complex adaptive systems perspective is beneficial to understand both the scientific collaborations itself and the environment in which the scientific collaborations exist. Complex adaptive systems theory is proven to be useful in explaining multi-agent systems, nonlinear relationships among agents and among systems, self-organization, adaptation and learning, and finally unique, unrelated or one-time events –the areas where traditional systems perspective fails (Aydinoglu, 2010b). Collaborative science, by nature, requires multiple researchers and nonlinear relationships among them. For instance, the relationship among a famous researcher, an average researcher, a graduate student, a technician, an engineer, etc is nonlinear. In addition, assume these agents belong to different institutions and different disciplines and the relationships get more complex. However, they learn from each other, adapt new skill sets and perspectives, establish new relationships, rules, and structures (emergence). Their collective impact is bigger than the sum of their individual impacts. Finally, they deal with unique or one-time events because

the funding agencies have limited resources. It is uncommon for one funding agency to support two projects that have the same goal. There is one large hadron collider, one Hubble telescope, one human genome project etc. This uniqueness is one of the reasons why we do not have a body of literature that systemically investigates scientific collaborations. Each one is so different than the other. However, complexity theory, basically a systems theory, is able to assess and compare different systems/scientific collaborations -which is the biggest contribution of this study.

This study fills the gap by applying complex adaptive systems perspective to a virtual scientific collaboration. In order to observe complex adaptive behavior, this study develops a tool, the complexity framework. The development of the framework has been a necessity because there has not been a unified complex adaptive systems theory. To build this framework, the researcher used the common features, concepts, and propositions of complexity theory in the seminal articles from the field of organizational studies and identified ten concepts (see Figure 19).

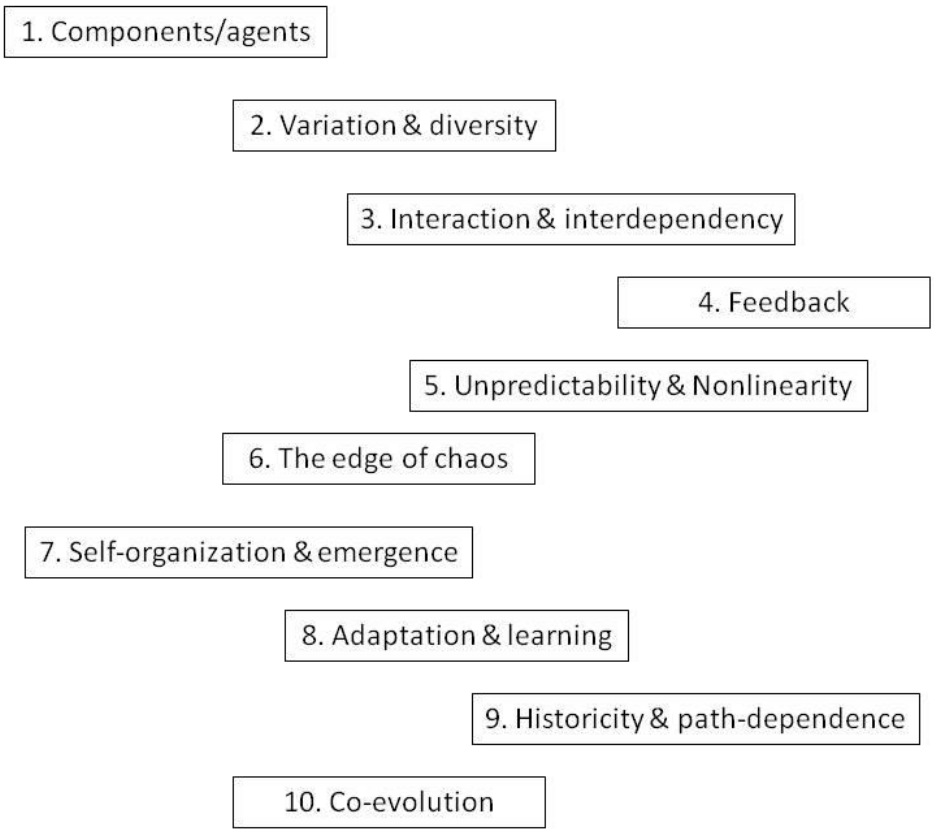


Figure 19 – Complex adaptive systems framework

The data generated through interviews, observations, and surveys are compared to the framework to see whether DataONE operates as a complex adaptive system, a system "that have a large numbers of components, often called agents, that interact and adapt or learn" (Holland, 2006, p. 1). The answer is 'yes'. Let me explain the process using the complexity framework above. *Italics* are used to refer the related concept in the complexity framework.

1. DataONE has a number of *components* (individuals and institutions).

2. They are *diversified* in terms of disciplinary background, career age, type of the institution, and motivation to join in. Hard scientists, computer scientists, and social scientists collaborated to create DataONE. They come from different institutions with different institutional agendas. Some of them are motivated by their career goals, some of them by personal goals.

3. However, despite the huge diversification that is present in the collaboration, their existence and tasks are *interdependent and interconnected* to each other. DataONE employs working group structure (and sometimes workshop structure). One working group's objective is the output of another and also input for another. A change in one, affects the rest. Moreover, they *interact* frequently via different media (face-to-face, email, video call, shared space, chat, etc.). The members are indeed selected based on their reputation of being good communicators which has proven to be important not only on taking care of daily tasks but also dealing with conflicts among people arising from different disciplines.

4. Interactions among members ensure *feedback* processes which also become significant in dealing with conflicts.

5. Although the goals are established in DataONE and the deadlines are met on time, the processes are *unpredictable*. A new working group emerges, some working groups change their names and priorities. In this way, DataONE is organic and reactive.

6 & 7. The environment that DataONE operates creates the right conditions for *emergence* –in this case DataONE itself. Inside DataONE, new structures, relationships among individuals and institutions, and products emerge. The management and the PI provide a balance between order and disorder to foster emergence. They encourage dialogue, interaction, innovation, and creativity. It is a suitable strategy for a bottom up formation. DataONE is mostly a volunteer organization, with a balanced management style in which the members are able to pursue their own goals by setting their own agenda that is in harmony with DataONE's goals. They all feel very motivated to do it because they are free. On the other hand, everyone feels responsible for the commitments (and deadlines) to the funding agency. Both are served by the balance between loose and tight management, which is translated as '*being at the edge of chaos*' in complex systems theory. Other *emergent* features are a new working group (exploration, visualization, and analysis working group), partnerships between institutions and scholars, and various outcomes (such as papers, posters, grant proposals, software codes, book chapters, etc.).

8. DataONE also demonstrates that it is able to *learn*, *adapt*, and *survive*. There have been changes in the internal structure and external environment that endangered the success of DataONE, yet they have been parried successfully. The recession, as a serious change in the external environment, for instance, has brought up some concerns about the sustainability of the project. The long term sustainability and governance working group has created the DataONE Business Plan to deal with financial insecurities; two of the five year of funding has been received; and a second five seems quite possible. However, the

expected donations and new grants might experience a setback. Besides the self-organization of exploration, visualization, and analysis working group, other changes occurred in the internal structure –mostly on the time committed by and the responsibilities of the individuals.

9. DataONE is a natural result of its predecessors: SEEK (the Science Environment for Ecological Knowledge) and the interoperability grant.

10. Other changes outside the DataONE (*coevolution*) have happened as well: the NSF's mandatory data management plan, interaction with Data Conservancy, and impact on scientific community. However, DataONE's real impact will occur in the future as it is still in the emergence phase[18].

In a nutshell, DataONE operates as a complex system which makes this virtual scientific collaboration resilient, adaptive, and successful.

---

[18] As of June 2011 DataONE is in the 2nd year of its 5 year funding and the cyberinfrastructure has not become public yet. Therefore, it is considered in the emergence phase. In order to be considered in the mature phase, the cyberinfrastructure has to become public and operational for some time.

2. The NSF's 2nd year review

In February 2011 DataONE went through the 2nd year review by the NSF and received positive feedback. The NSF and DataONE negotiated the goals of DataONE. The deliberate review of the NSF revealed that DataONE has reached the goals that were expected. The cyberinfrastructure team has been developing cyberinfrastructure for three major components (coordinating nodes, member nodes, and investigator toolkit) and deployment of prototypes for each of them has been done. In addition, the community engagement team "has made progress in its four major activities: (1) providing responsive governance and management; (2) engaging the broad community in DataONE and building an extensive data resource; (3) creating an informatics literate populace; and (4) ensuring financial support and sustainability" (NSF DataONE Progress Report, 2011, p.6).

DataONE is progressing according to the plan; although, there have been a number of changes internally and externally that has been explained in the previous chapters. This is the strength of a resilient and adaptive system. This study, so far, has discussed the factors that made it possible through complex adaptive system theory perspective. As a complex adaptive system, DataONE has proved to be successful so far by the NSF standards and also with the threats dealt with which are explained in Chapter 5, Section 8 and summarized above under adaptation and learning title. Being a complex adaptive system is definitely a strength for the collaboration. However, they have to be treated accordingly. For instance, treating them as they are hierarchical structures would cause disasters as it did in the Columbia space shuttle disaster when the administration did not

take into account the concerns of the engineers (Aydinoglu, 2010b, 27). Scientific endeavors are full of surprises, they are unpredictable, and the power of a collaboration is actually lies in the strength and harmony of the collective minds it employed. In order to respond and adapt to the surprises, changes, and threats in the scientific and non-scientific environment[19] these minds should be set free but not let chaos reign. The PI(s), the management, and the funding agencies should be aware of the strengths and weaknesses of complex adaptive systems.

3. Contributions of this study

Contribution 1. A complexity framework for virtual scientific collaborations has been developed.

The first contribution is the development of complex adaptive systems framework for human organizations. Complex adaptive systems exist in many realms from micro to macro and from inorganic to organic worlds. The variety and diversity of these systems prevented us having a unified complex adaptive systems theory. Therefore, even though there is a consensus on some of the basic concepts in different applications and some frameworks developed in some disciplines (such as education, computer science,

---

[19] The importance of the non-scientific environment is going to be explained when I discuss the super-conducting super collider fiasco in the next section.

management), this is the first framework developed for scientific collaborations. However, it needs to be tested more; one case is not enough.

As a future research idea, the assessment of other scientific collaborations with complexity framework is needed to see whether the framework works and if it works, it needs refining. The refining should also include the scientific collaborations at later stages (mature & dissolving) because this one only addresses the emergence phase. The problems experienced in different stages are likely to require different responses. For instance, in mature collaboration routine tasks might hinder creativity and innovation. Also, such as the data lifecycle developed by DataONE team, the development of a scientific collaboration lifecycle might be very useful to assess and investigate scientific collaborations. In addition, a quantitative complexity framework might be developed in some areas such as identifying communication patterns among members or tracking who collaborates with whom in the next research partnership.

Contribution 2. The complexity framework to virtual scientific collaborations has been applied to real case.

The employment of complex adaptive systems theory in scientific collaborations has been very limited. Wagner (2008) for example focused on science policies in developing countries, Vasileiadou (2009) focused on messages among research teams. Aydinoglu (2010b) proposed that scientific collaborations could be studies as complex adaptive systems. This study; on the other hand, is the first step of developing and applying a framework to assess scientific collaborations using complexity theory. The framework is

going to be useful in three areas: having comparable results, lessons for the management/PI; and lessons for the funding agencies. The complexity framework has been applied to DataONE. It has proved to be useful in explaining the success, strength and resilience of DataONE.

Contribution 3. It would be possible to have comparable results that increase our understanding of scientific collaborations.

It has been mentioned that there is a lack of comparable studies regarding scientific collaborations. In complex adaptive systems theory the common denominator is the system yet different units of analysis (individual, team, collaboration, system, environment) is possible with it through fractal/self-similarity feature of complex systems. It is; therefore, a powerful tool to compare different scientific collaborations in different funding environments to each other. With this tool it might be possible to convert lessons learned in one to another. Complexity theory does not aim to manipulate or forecast; however, it envisions short-term prediction through the use of fractals or self-similar structures. Unfortunately, it has not been possible to collect data about this feature due to DataONE being in the emergence phase. More time is needed. Also a bigger collaboration, as it has more agents, and thus more interactions, provides more data to observe self-similarity at different levels.

Complex systems are able to learn. Success in one collaboration could be mimicked in the other or failure in one could be avoided in the other. Best practices could be shared. If a framework and related literature could be developed, transfer of experience and

knowledge in between collaborations could be possible. After all, the scientific community is a bigger complex adaptive system that is capable of learning from its subsystems. This study might be of help in facilitating learning and establishing a formal experience transfer method. Complex adaptive systems theory could act as a master key in our approach to scientific collaborations.

A best practices toolkit would be beneficial to scientific community. While developing and refining the complexity framework to assess scientific collaborations, hopefully enough cases will be accumulated to develop a best practices toolkit. Spreading out the lessons from one to another is already envisioned in complex adaptive systems theory under 'learning' concept as each collaboration is a system and the scientific community is a bigger system and complexity theory is capable of explaining the relationships among systems as well.

Contribution 4. There are some lessons for the management/Principal Investigator (PI).

This study's finding is similar to the literature on complex adaptive systems regarding that complex adaptive systems are resilient, capable of learning and adaptation, and innovative. In order to have a functioning complex system, certain conditions have to be met, which is the job the PI(s) or the management of the collaboration. The management/PI(s) should create an environment that is in between order and chaos. Micromanagement is not a good idea, for instance. Tight management is another bad idea (Simons, 1995). They both smother creativity, innovation, and emergence. Of course,

scientific endeavor is more about discipline, tenacity, tedious and routine work in the lab/field, yet creativity and innovation could lead to groundbreaking results. At the end, that is what science does: come up with a new, fresh perspective to explain how things happen/behave. On the other hand, chaos should be avoided too. Especially in big science where many nations, organizations, individuals involved in a number of tasks, things have the tendency to get out of control easily and quickly. Again it is the management/PI(s) job to make the collaboration meet its goals on its deadlines. In a nutshell, the management/PI(s) has to maintain a balance in between loose and tight or sweet and tough. Moreover, in order to make use of collective minds, the PI(s) should encourage communication and interaction.

The emerging rituals and culture through safety protocols at the U.S. Los Alamos National Laboratory (Sims, 2005) is reflected in the preservation procedures (Subversion and the plone website) in DataONE. In a distributed organization like DataONE, the safety protocols become the protocol to keep different versions of software code, which made Subversion the right tool. However, community engagement team operates in a different realm; therefore, that is not their concern and they do belong to a different realm. They did not adapt that ritual and use the plone website for preservation. The organizational culture is affected by the setting, venue, and procedures. However, in a virtual/distributed organization only procedures remain. It is harder for the PI and the management to create an organizational culture by using these means.

As a future research idea, an approach to leadership studies from a CAS perspective might provide useful insights. There is some literature on leadership and complexity theory; however, as complexity theory is in favor of bottom-up formation, leadership studies are considered to be anti-thesis of it. I believe it is a negotiation between both. The leaders (formal & informal/natural) play a key role in scientific collaborations. More studies are needed to reveal the dynamics of how can leaders facilitate interaction and communication, foster creativity and innovation, and motivate other scholars to contribute. Furthermore, creating an organizational culture remains a challenge due to the proximity of members. Communication becomes even more important.

Contribution 5. There are some lessons for funding agencies.

Funding agencies are bureaucratic institutions and like all bureaucratic institutions they are slow to respond to changes in their environment. The complex adaptive systems assessment tool could be of help to them to assess both the proposals and the environment. There is not much empirical evidence that scientific collaborations that operate according to complex adaptive systems theory are more successful than the traditional ones; however, we know from organizations studies that organizations have a better chance of survival if they are able to learn and adapt (Pascale, Millemann, & Gioja, 2000). If we extend that knowledge on organizations to scientific collaborations and assess them from complexity theory perspective, we might have a better perspective on how to allocate our limited resources.

Moreover, complex adaptive systems theory could be used to assess the environment in general. For instance, the wrong assessment of the environment led the U.S. Congress to approve the super-conducting super collider (SSC) project, the biggest particle accelerator of its time (1990s) planned to be built. The project kicked off in late 80s and after $2.6 billion was spent, the project was terminated (Goodwin, 1993). There are several reasons for this experience, all of which are rooted in different areas making it difficult to identify. An assessment with complexity theory perspective might have provided the decision makers with a clearer, more comprehensive picture of the situation. The decision makers could not integrate the data they have because data belonged to different domains and there are not many people to make sense of such disparate data. The reasons are:

- The physics community was not in favor of the project (Lemonick, 1988) – of course except the particle physicists; other disciplines (biomedical and space research) were not in favor of it as well due to the competition for funds.
- The funding paradigm had shifted from 'national preeminence to international partnership' (Goodwin, 1994, p.88) because the Soviet bloc had collapsed, there were no communist threat (later when the U.S. government was out of money and ask other countries to help, they did not due to the national preeminence rhetoric employed by the U.S.).
- The Congress and public were against the project as the U.S. economy had been experiencing the highest budget deficit so far (Anderson, 1993).

- The project was handed to an incompetent management (Greenberg, 2001), the cost had risen from $4, to $6, to $8, and finally to 12 billion.

As it can be seen each reason represents a complex system that falls to the interest of a different domain (see Table 4); however, integrating them into a bigger system, analyzing all of them together is possible through complex adaptive systems theory because in such an analysis the unit of analysis becomes system. For instance, the policy makers could have taken the record U.S. deficit and thus public opposition to the SSC into account before approving the project. Or were the policy makers able to reassess the foreign political arena, they could position the SSC as an international project rather than a project to show off the U.S. dominance, which later could have helped to receive funding from other countries when it became obvious that the U.S. could not do it alone. Or they could have realized that the scientific community –including the physicists– were not ready for such a project that exhausts all the funding as they do not consider it as a priority. If the system do not reach critical point, emerge does not happen. Everything was connected yet because each one fell into a different disciplinary domain, a comprehensive analysis could not be done. Each reason is a system that has an impact on the others. $2.6 billion could have been saved if we had a tool to make sense of the relationships among different domains/systems. The complexity framework might be that tool. Although, long-term prediction is not possible in CAS theory, it provides some insights that could be beneficial. In addition, with the accumulation of data and examination of many cases, some generalizations might be possible in the future.

Table 5 – The reasons that led the termination of SSC

Reasons to fail	**Domain**
The U.S. deficit	Finance
Public reaction to the money spent for SSC	Politics, public relations
The reaction of the physicists community	Science policy
The reaction of the scientific community	Science Policy
Unipolar world, Soviet threat no more	International relations, political sciences
Incompetent management	Management science, organizational studies

As can be seen from the example above, not only is the scientific community a complex system, but also the funding environment in which scientific community resides is one. The snapshot of the scientific community through a complex adaptive systems theory, identification of agents (from funding agencies to the decision makers) and the environment (political, economical, technological, social environment) and the relationships among them could be quite useful. For instance, the scientific community as a complex adaptive system reached a critical point and as a result, DataONE emerged. The data-intensive research paradigm, the NSF's vision, and the rise of virtual collaborations are the excess energy/matter/information that led the system to reach the critical point and the combination of certain agents made DataONE's emergence possible.

4. Discussion on Multidisciplinarity, Communication Studies, and Information & Library Sciences

For a scientific collaboration to be considered as a complex adaptive system, the easiest way to assess diversity and variation is to check the disciplines involved in the collaboration. If a collaboration is operating according to complex adaptive systems theory, it is likely that that collaboration is multidisciplinary. Therefore, the complexity theory perspective does provide insights on multidisciplinary collaborations. There are studies (and common sense agrees with this argument) that the more disciplines involved in a collaboration, the harder to communicate, cooperate, and collaborate. Indeed, the findings of this study are parallel with this argument that such collaborations are prone to challenges, tensions and conflicts. In DataONE, the division between the cyberinfrastructure and community engagement teams on document preservation and sharing; the differences on the terminology; the differences on research questions are to name a few of these challenges.

However, the multidisciplinary structure also provides opportunities. The members of DataONE believe that it is one of the main reasons of DataONE's success: "The diversity of organizations and participants involved in the project is one of our greatest strengths" (NSF DataONE Progress Report, 2011, p. 17). Multidisciplinarity brings new perspectives, new tools, and new methods that are not thought of before to deal with the problems. Like super glue, different communities are connected to each other through diversified backgrounds and expertise, which is indeed important because they are different

facets of the same problem whether we think it is as like that or not. Climate change, spreading of populations or infectious diseases are examples of such multi-faceted problems that cannot be tackled by one perspective/discipline. Furthermore, multidisciplinarity (which indeed is simply diversity) is the source of learning, innovation, resilience, and adaptation. Members from different disciplines learn from each other, make use of each other's knowledge and experience, and become cost effective. Multiple perspectives might result in the emergence of novel solutions, relationships, and structures. In the end, the collaborations survive until they fulfill their goals, become more effective and successful.

Exploring multidisciplinarity through the concept of variation and diversity in complex adaptive systems theory might provide new insights. Multidisciplinary projects are by definition diversified because of the different disciplines involved in them. Complex adaptive systems theory explains how order emerges from such diversity and variation; hence, it can be applied to multidisciplinarity to provide a new perspective.

Two disciplines become prominent and different in multidisciplinary collaborations that operate according to the complex adaptive systems theory: communication studies and information sciences. It can be remembered from the framework that interaction among agents is a crucial element. The interaction depends on communication. Agents should be able to communicate, understand each others' terminology, and provide feedback to each other. Science communication is generally referred as 'public understanding of science' and tends to focus on communicating scientific results to the public and/or policy-makers.

Communication among scientists is studied through scientometrics and its limitations are summarized in early chapters.

I believe, the lack of studies on communication among scientists, is a big challenge in multidisciplinary scientific collaborations. Theories from organizational communication might be of help to a certain degree; however, that literature focuses on mostly profit-based organizations. The environment these organizations live in differs greatly from the environment that the scientific collaborations live in. Moreover, the end goal is not profit in scientific collaborations. Profit-based organizations have infinite life span (they live as long as they profit which is the reason they come to being) whereas scientific collaborations, have a limited life span –until the funding is over and/or they achieve their goal (Aydinoglu, 2010). It is quite likely that different motivations require different approaches.

The goal is knowledge for scientific collaborations. Considering what is at stake; considering the problems we suffer because we do not have the knowledge (climate change, infectious diseases, energy demand…); considering how much money we spent to obtain that knowledge (the amount of money allocated for R&D in the U.S. was$350 billion last year) (UNDP, 2008); and considering the impact of that knowledge, if we have, through science and technology on our civilization (higher living standards, creating jobs, security)– this community is too big and important to be ignored. Our society simply does not have enough time or resources for scientists to realize and overcome the problems they have in communication –especially in multidisciplinary environments.

As for information and library sciences, this study revealed that scientists in this project were not aware of the capabilities, skills, and potential contributions of information and library scientists. The NSF thought that it would be good to have some involvement from the information and library scientists at the very early stages of proposal writing. The hidden potential was realized immediately and information and library science scholars were involved in the project really early on. Before they come on board, there was no one who had graduate education in information sciences in the DataONE collaboration. People had professional experience on data issues; however, it is not enough. We are now in a new paradigm of scientific research: data-intensive research (Hey, Tansley, & Tole, 2009). Data (and all related issues) are at the core of this kind of research. We are not even able to store the data we collect (Gantz, 2008, p.4). Analyzing, preserving, accessing, sharing, and reusing will come after that. We need people who are trained in information sciences. We also need scientists who are trained in information sciences so that they can have the notion of metadata and data lifecycle or least know that this is a serious problem and also who to call for support. Scientific collaborations should have data management plan and now the NSF requires one for proposals they receive (NSF, 2010). This is a promising start yet more needs to be done.

Multidisciplinary and data-intensive research requires not only information intensive collaborations but also communication intensive ones. The importance of these disciplines needs to be explained. These disciplines differ from others because they are more like the support personnel for the collaboration. Apart from the content of the

research (earth sciences, particle physics, climate change, etc.), every collaboration needs good communicators and data people (because communication and data is about facilitating the processes, they are independent from the content). Communication is crucial for the well-being of a collaboration, and the frequency of communication has increased due to the developments in ICTs. As for data, it has become the core of research (Hey, Tansley, & Tole, 2009). In addition, data reuse, which is only possible if preservation and access is provided, is a very cost effective way of doing science (Hey, Tansley, & Tole, 2009; Tenopir et al., 2011). Therefore, every collaboration needs professionals trained in these fields to help them facilitate the research process. The theoretical contributions of these fields are important; however, it is obvious that they have practical use in scientific collaborations so that these collaborations can perform more efficiently and effectively.

Discussing and exploring the role of communication studies and information sciences in scientific collaborations become more important than ever. Because of being at the core of the activities, people who are responsible for communication and information management will probably develop a gate-keeper role in the collaborations. This role needs to be studied. In addition, assessing the data management needs of a scientific collaboration, for instance, has already been a necessity for applying for an NSF grant. Integration and preservation of different research projects are going to be an issue in the near future; thus, it is time to start thinking about them. These will definitely create a demand for people who have the skills and experience. The library and information science programs might think revising their curriculums to supply the demand. On the same token,

the increase in the number of multidisciplinary research projects will require communicators specialized in multidisciplinarity. More studies are needed on the current condition of professionals in both fields, projections of what is going to be demanded from them in the future, how to respond to that demand, and the assessment of the current communication and information needs of virtual scientific collaborations.

One final note about interaction lies at the intersection of communication studies and information sciences, which is especially critical for virtual scientific collaborations. The interaction in a virtual collaboration heavily relies on computer mediated communication. This is an area where communication studies and information science could work together and actually they do. However, the interviews revealed that the commercially available software on the market that is available to DataONE team is not sufficient to take care of midsize virtual collaboration. There are others that are better. It is a challenge –especially considering the different workflows different disciplines have. However, this is neither a small nor insignificant community to be ignored. The open source community might be mobilized to develop better software and tools. The identification of problems and needs regarding communication tools for virtual scientific collaborations might be useful.

In summary, the main contributions of this study are introducing a new perspective (complex adaptive systems theory) to explain the emergence of virtual scientific collaborations, developing a framework to assess them, and applying it to a real life case (DataONE). DataONE's success (so far), and adaptivity and resilience to external and

internal threats can be explained through the framework. The framework can be applied to other virtual scientific collaborations to assess their emergence. In addition, it might provide insights about the resilience of them to external and internal threats.

List of References

Adamic, L., & Glance, N. (2005). The political blogosphere and the 2004 U.S. election: Divided they blog. In J.Adibi, M. Grobelnik, D. Mladenic, & P. Pantel (Eds.), *Proceedings of the 3rd International Workshop on Link Discovery* (pp. 36-43). New York, NY: ACM.

Agar, J. (2006). What difference did computers make? *Social Studies of Science, 36,* 869-907.

Allison, P.D. (1980). Inequality and scientific productivity. *Social Studies of Science, 10,* 163-79.

Ancona, D. G., & Caldwell, D. F. (1992). Demography and design: Predictors of new product team performance. *Organization Science*, *3*(3), 321–341.

Anderson, C. (1993). The anatomy of a defeat: Congress cuts funding for Superconducting Super Collider. *Science, 262*, 245.

Anderson, P. (1999). Complexity theory and organization science. *Organization Science, 10,* 216-32.

Arthur, B. (1994). Increasing returns and path dependence in the economy. Michigan: University of Michigan Press.

Arthur, B. (2009). *The nature of technology: What it is and how it evolves.* New York: Free Press.

Axelrod, R. and Cohen, M.D. (1999), *Harnessing Complexity: Organizational Implications of a Scientific Frontier.* New York, NY: Basic Books.

Aubert, B. A., & Kelsey, B. L. (2003). Further understanding of trust and performance in virtual teams. *Small Group Research, 34*(5), 575-618.

Aydinoglu, A. U. (2010a). Emergence of a scientific collaboration: DataONE case study. *4S (Society for the Social Studies of Science) Annual Meeting*, Komaba I Campus, University of Tokyo, Tokyo, Japan.

Aydinoglu, A. U. (2010b). Scientific collaborations as complex adaptive systems. *Emergence: Complexity and Organization, 4*, 15-29.

Baranger, Michel (2000). *Chaos, Complexity, and Entropy: A Physics Talk for Non-Physicists*. Retrieved on December 2010 from http://www.necsi.edu/projects/baranger/cce.pdf

Barth, K.H. (2003). The politics of seismology: Nuclear testing, arms control, and the transformation of a discipline. *Social Studies of Science, 33,* 743-81.

Benbya, H. & McKelvey, B. (2006). Toward a complexity theory of information systems development. *Information Technology and People, 19*, 12-34.

Bentley, R.A., & Maschner, H.D.G. (2007). "Complexity Theory" in *Handbook of Archaeological Theories* (RA Bentley, HDG Maschner, C Chippendale, eds.): 245-270. Maryland: AltaMira Press.

Bodnarczuk, M. & Hoddeson, L. (2008). Megascience in particle physics: The birth of an experiment string at Fermilab. *Historical Studies in the Natural Sciences, 4,* 508-534.

Bogg, J. & Geyer, R. (2007). *Complexity science and society*. New York: Radcliffe Publishing

Boshoff, N. (2009). Neo-colonialism and research collaboration in Central Africa. *Scientometrics, 81,* 413-34.

Braun, T., Glanzel, W., & Schubert, A. (2001). Publication and cooperation patterns of the authors of neuroscience journals. *Scientometrics, 51,* 499-510.

Bruun, H. & Sierla, S. (2008). Distributed problem solving in software development: The case of an automation project. *Social Studies of Science, 38,* 133-158.

Bush, V. (1945). *Science: The endless frontier*. Retrieved on March 15, 2009 from http://www.nsf.gov/od/lpa/nsf50/vbush1945.htm

Cambrosio, A., Keating, P., & Mogoutov, A. (2004). Mapping collaborative work and innovation in biomedicine: A computer-assisted analysis of antibody reagent workshops. *Social Studies of Science, 34,* 325-64.

Choi, H. & Mody, C.C.M. (2009). The long history of molecular electronics: Microelectronics origins of nanotechnology. *Social Studies of Science, 39,* 11-50.

Cilliers, P. (1998). *Complexity and postmodernism: Understanding complex systems*. Kentucky: Routledge.

Clemens, Jr., W.C. (2006). "Understanding the coping with ethnic conflict and development issues in Post-soviet Eurasia" in *Complexity in World Politics: Concepts and Methods of a New Paradigms* (N.E. Harrison ed.) New York: State University of New York Press.

Cloud, J. (2001). Imaging the world in a barrel: CORONA and the clandestine convergence of the earth sciences. *Social Studies of Science, 31,* 231-51.

Collins, H. (2004). *Gravity's shadow: The search for gravitational waves.* Chicago: The University of Chicago Press.

Computer Science and Telecommunications Board (1995). *Evolving the high performance computing and communications initiative to support the nation's information infrastructure*. Washington, D.C.: The National Academies Press.

Cogburn, D., Santuzzi, A., & Vasquez, F.K.E. (2011). *Developing and validating a socio-technical model for geographically distributed collaboration in global virtual teams*. Proceedings of the 44th Hawaii International Conference on System Sciences.

Corbin, J. & Strauss, A. (2008). *Basics of Qualitative Research (3rd ed.)*. California: Sage.

Cramer, F. (1993). *Chaos and order: The complex structure of living systems*. New York: Wiley-VCH.

Cronin, B. (2001).Hyperauthorship: A postmodern perversion or evidence of a structural shift in scholarly communication practices? *Journal of the American Society for Information Science and Technology, 7,* 558-69.

Cronin, B., & Meho, L. (2006). Using the h-index to rank influential information scientists. *Journal of the American Society for Information Science and Technology*, *57*(9), 1275-1278.

Cummings, J.N. & Kiesler, S. (2005). Collaborative research across disciplinary and organizational boundaries. *Social Studies of Science, 35*, 703-22.

DataONE (2009). *Proposal to the NSF DataNet Solicitation*. Retrieved on March 2010 from htpps://dataone.org

DataONE (2009). *About* retrieved on November 18, 2009 from https://datanet.ecoinformatics.org/about

DataONE (2011). *Progress report to NSF for 2nd year review*. Retrieved on April 2011 from https://docs.dataone.org

de Solla Price, D. (1963). *Little Science, Big Science.* New York: Columbia University Press.

de Solla Price, D. (1977). *Science, technology, and society: A cross-disciplinary perspective.* California: Sage Publications.

Dean, K., Naylor, S., Turchetti, S., & Siegert, M. (2008). Data in Antarctic science and politics. *Social Studies of Science, 38.* 571-604.

Denzin, N. (1978). *Sociological Methods (6th Ed.)*. New York: McGraw Hill.

Ding, Y., Foo, S., & Chowdhury, G. (1999). A bibliometric analysis of collaboration in the field of information retrieval. *The International Information & Library Review, 30,* 367-76.

Douglas, H. (2009). *Science, Policy, and the value-free ideal*. Pennsylvania: The University of Pittsburgh Press.

Duque, R.B., Ynalvez, M., Sooryamoorthy, R., Mbatia, P., Dzorgbo, D.B.S., & Shrum, W. (2005). Collaboration paradox: Scientific productivity, the Internet, and problems of research in developing areas. *Social Studies of Science, 35,* 755-85.

Eisenhardt, K.M. (1989). Building theories from case study research. *Academy of Management Review, 4,* 532-50.

European Science Foundation (2007). *Shared responsibilities in sharing research data: Policies and partnerships*. Retrieved on August 2010 from http://www.esf.org/index.php?eID=tx_nawsecuredl&u=0&file=fileadmin/be_user/CEO_U nit/Science_Policy/Sharing_Data/ESF_DOC_SHARINGDATA_V01ppp.pdf&t=1287487773&hash =89d4ad56d3544ba3edb68c1683369ed4

Flyvbjerg, B. (2006). Five misunderstandings about case study research. *Qualitative Inquiry, 12,* 219-45.

Garfield, E. (2009). From the science of science to Scientometrics visualizing the history of science with HistCite software. *Journal of Informetrics, 3*(3), 173-179. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S1751157709000297

Gantz, J.F. (2008). *The diverse and exploding digital universe: An updated forecast of worldwide information growth through 2011*. IDC White Paper, retrieved on December 2010 from http://www.emc.com/collateral/analyst-reports/diverse-exploding-digital-universe.pdf

Geertz, C. (1973). *The Interpretation of Cultures: Selected Essays*. New York: Basic Books.

Gerovitch, S. (2001). 'Mathematical machines' of the Cold War: Soviet computing, American cybernetics and ideological disputes in the early 1950s. *Social Studies of Science, 31,* 253-87.

Glanzel, W. & De Lange, C. (1997). Modelling and measuring multilateral co-authorship in international scientific collaboration. Part II. A comparative study on the extent and change of international scientific collaboration links. *Scientometrics, 3*, 605-26.

Glanzel, W. (2001), National characteristics in international scientific co-authorship, *Scientometrics, 1,* 69-115.

Glanzel, W. (2002). Coauthorship patterns and trends in the sciences (1980-1998): A bibliometric study with implications for database indexing and search strategies. *Library Trends, 50,* 461-73.

Gleick, J. (1987). *Chaos: Making a new science*. New York: Penguin Books.

Goodwin, I. (1993). Congress cancels SSC and allocates high budgets for technology in 1994. *Physics Today, 46*, 77-80.

Goodwin, I. (1994). After agonizing death in the family, particle physics faces grim future. *Physics Today, 47*, 87-91.

Gossart, C. & Ozman, M. (2009). Co-authorship networks in social sciences: The case of Turkey. *Scientometrics, 78,* 323-45.

Graziano, A.M. & Roulin, M.L. (2000). *Research methods: A process of inquiry*. Boston: Allyn and Bacon.

Greenberg, D.S. (2001). *Science, money, and politics: Political triumph and ethical erosion*. Chicago: The University of Chicago Press.

Guba, E.G. (Ed.). (1990*). The paradigm dialog*. California: Sage Publications.

Guston, D.H. (2000). Between *politics and science: Assuring the productivity and integrity of research*. Massachusetts: Cambridge University Press.

Guston, D.H. & Keniston, K. (1994). *The fragile contract: University science and the federal government.* Massachusetts: MIT Press.

Hackett, E.J. (2005). Essential tensions: Identity, control, and risk in research. *Social Studies of Science, 35,* 787-826.

Hara, N., Solomon, P., Kim, S.L., Sonnenwald, D.H. (2003). An emerging view of scientific collaboration: Scientists' perspectives on collaboration and factors that impact collaboration. *Journal of the American Society for Information Science and Technology, 54,* 952-65.

Harper, K. (2003). Research from the boundary layer: Civilian leadership, military funding and the development of numerical weather prediction (1946-55). *Social Studies of Science, 33,* 667-96.

Hart, R. (2007). Collaboration and article quality in the literature of academic librarianship. *Journal of Academic Librarianship, 33,* 190-5.

Haythornthwaite, C. *(*2009*).* Crowds and communities: Light and heavyweight models of peer production. *In* R.E.Sprague *(Ed.), Proceedings of the 42nd Hawaii International Conference on System Sciences* (pp. 1–10*).* Los Alamitos, CA*:* IEEE Computer Society.

He, T. (2009). International scientific collaboration of China with the G7 countries. *Scientometrics, 80,* 571-82.

Headland, T. & Pike, K, & Harris, M. (1990). *Emics and Etics: The Insider/Outsider Debate.* California: Sage Publications.

Hedgecoe, A. & Martin, P. (2003). The drugs don't work: Expectations and the shaping of pharmacogenetics. *Social Studies of Science, 33,* 327-64.

Hey, T., Tansley, S., & Tole, K. (2009). *The fourth paradigm: Data-intensive scientific discovery*. Retrieved on June 2010 from http://research.microsoft.com/en-us/collaboration/fourthparadigm/4th_paradigm_book_complete_lr.pdf

Hine, C. (2006). Databases as scientific instruments and their role in the ordering of scientific work. *Social Studies of Science, 36,* 269-98.

Hoffmann, M.J. (2006). "Beyond regime theory: Complex adaptation and the Ozone depletion regime" in *Complexity in World Politics: Concepts and Methods of a New Paradigms* (N.E. Harrison ed.) New York: State University of New York Press.

Holland, J.H. (1995). *Hidden Order: How Adaptation Builds Complexity*. Reading, MA: Addison-Wesley.

Holland, J. (1998). *Emergence: From chaos to order.* Reading, MA: Addison-Wesley.

Hong, W. (2008). Domination in a scientific field: Capital struggle in a Chinese isotope lab. *Social Studies of Science, 38,* 543-70.

Houchin, K. K., & MacLean, D. D. (2005). Complexity Theory and Strategic Change: an Empirically Informed Critique. *British Journal of Management, 2,* 149-66.

Horning, S.S. (2004). Engineering the performance: Recording engineers, tacit knowledge and the art of controlling sound. *Social Studies of Science, 34,* 703-31.

Inter-University Consortium for Political and Social Research (2009). *Guide to social science data preparation and archiving: Best practice throughout the data life cycle*. Ann Arbor: University of Michigan Press. Retrieved on September 2010 from http://www.icpsr.umich.edu/files/ICPSR/access/dataprep.pdf

International Energy Agency (2009). *World Energy Outlook 2009*. Retrieved November 23, 2009 from http://www.worldenergyoutlook.org/

International Panel on Climate Change (2007). *Climate change 2007: Synthesis report.* Retrieved June 30, 2009 from http://www.ipcc.ch/pdf/assessment-report/ar4/syr/ar4_syr.pdf

Janicik, G. A., & Bartel, C. A. (2003). Talking about time: Effects of temporal planning and time awareness norms on group coordination and performance. *Group Dynamics: Theory, Research, and Practice, 7*(2), 122-134.

Jeffrey, P. (2003). Smoothing the waters: Observations on the process of cross-disciplinary research collaboration. *Studies of Social Science, 33,* 539-62.

Kacen, L. (1999). Anxiety levels, group characteristics, and members' behaviors in the termination stage of support groups for patients recovering from heart attacks. *Research on Social Work Practice, 9*(6), 656-672.

Katz, J.S. & Martin, B.R. (1997). What is a research collaboration? *Research Policy, 1,* 1-18.

Kauffman, S.A. (1993). *The origins of order: Self-organization and selection in evolution*. New York: Oxford University Press.

Kauffman, S.A. (1995). *At Home in the Universe: The Search for the Laws of Self-Organization and Complexity*. New York, NY: Oxford University Press.

Kertcher, Z. (2009). *Epistemic gaps and bridges in interdisciplinary collaboration*. Paper presented at the Society for Social Studies of Science Annual Meeting in Washington, D.C.

Kitcher, P. (2001). *Science, truth, and democracy*. New York: Oxford University Press.

Large Hadron Collider. (2009). Retrieved on November 28, 2009 from http://lhc-machine-outreach.web.cern.ch/lhc-machine-outreach/

Laser interferometer gravitational-wave observatory (LIGO) retrieved on February 10, 2009 from http://www.ligo.org/about.php

LaFollette, M.C. (1996). *Stealing into Print: Fraud, Plagiarism, and Misconduct in Scientific Publishing.* California: University of California Press.

Lee, S. & Bozeman B. (2005). The impact of research collaboration on scientific productivity. *Social Studies of Science, 35*, 673-702.

Lemonick, M.D. (1988). A controversial prize for Texas: The superconducting super collider has a home at last. *Time*, 132, 79.

Lewin, A.Y. (1999). Application of complexity theory to organization science. *Organization Science, 3,* 215.

Lomnitz, L.A., Rees, M.W., & Cameo, L. (1987). Publication and referencing patterns in Mexican Research Institute. *Social Studies of Science, 17,* 115-33.Lopresti, M. (2008). WikiProteins brings scientific collaboration to the WikiSphere. *EContent, 31,* 16-7.

Marecek, J. & Fine, M. (1997). Working between worlds: Qualitative methods and social psychology. *Journal of Social Issues, 4,* 631-44.

McCracken, G. (1998). *The long interview*. California: Sage Publications.

McKelvey, B. (2001). Energizing order-creating networks of distributed intelligence. *International Journal of Innovation Management, 5,* 181-212.

Maienschein, J. (1993). Why collaborate? *Journal of the history of biology, 26*, 167-83.

Mazur, A. & Boyko, E. (1981). Large-scale ocean research projects: What makes them succeed or fail? *Social Studies of Science, 11,* 425-49.

McKelvey, B. (1999). Complexity Theory in Organization Science; Seizing the Promise of Becoming a Fad. *Emergence: Complexity and Organization, 1*, 5–32.

Mellor, F. (2007). Colliding worlds: Asteroid research and the legitimization of war in space. *Social Studies of Science, 37,* 499-531.

Mirowski, P. & van Horn, R. (2005). The contract research organization and the commercialization of scientific research. *Social Studies of Science, 35,* 503-48.

Mitchell, M. (2009). *Complexity: A guided tour.* New York: Oxford University Press.

Mitleton-Kelly, E. (2003). "Ten principles of complexity & enabling infrastructures" in *Complex Systems & Evolutionary Perspectives of Organisations: The Applications of Complexity Theory to Organisations*.

Montgomery, K. & Oliver, A.L. (2009). Shifts in guidelines for ethical scientific conduct: How public and private organizations create and change norms of research integrity. *Social Studies of Science, 39,* 137-55.

Morel, B. and R. Ramanujam (1999). Through the Looking Glass of Complexity: The Dynamics of Organizations as Adaptive and Evolving Systems. *Organization Science, 3,* 278–93.

National Aeronautics and Space Administration (2006). NASA Unveils Global Exploration Strategy and Lunar Architecture. Retrieved on June 2010 from http://www.nasa.gov/home/hqnews/2006/dec/HQ_06361_ESMD_Lunar_Architecture.html

National Science Foundation Cyberinfrastructure Council (2006), *NSF's Cyberinfrastructure Vision for 21st Century Discovery (Version 5).*

National Science Foundation (2007). NSF 07-28, Cyberinfrastructure Vision for 21st Century Discovery. Retrieved on March 2010 from http://www.nsf.gov/pubs/2007/nsf0728/index.jsp

National Science Foundation DataNet: Curating scientific data, retrieved on February 12, 2010 from http://hdl.handle.net/1853/28513

National Science Foundation Office of Cyberinfrastructure & Directorate for Computer & Information Science & Engineering (2008). *Sustainable digital data preservation and access network partners (DataNet*).

National Science Foundation Office of Cyberinfrastructure retrieved on January 30, 2010 from http://www.nsf.gov/div/index.jsp?div=oci .

National Science Foundation, (2010). Scientists seeking NSF funding will soon to be required to submit data management plans. *Press releases 10-777* retrieved from http://www.nsf.gov/news/news_summ.jsp?cntn_id=116928

National Science Foundation (2011). Program Solicitation – NSF 11-501: Virtual Organizations as Socio-Technical Systems (VOSS). Retrieved from http://www.nsf.gov/pubs/2011/nsf11501/nsf11501.htm

Navarro, A. & Martin, M. (2008). Scientific production and collaboration in epidemiology and public health, 1997-2002. *Scientometrics, 76*, 291-313.

Panzarasa, P., Opsahl, T., & Carley, K. M. (2009). Patterns and dynamics of users' behavior and interaction: Network analysis of an online community. *Journal of the American Society for Information Science and Technology*, *60*(5), 911-932

Pascale, R.T., Millemann, M. & Gioja, L. (2000). *Surfing the edge of chaos*. New York: Three Rivers Press

Paskin, N. (2010). Digital Object Identifier (DOI®) System. Bates, M.J. & Maack, M.N. In *Encyclopedia of Library and Information Sciences, 3rd. Ed.*England: Taylor & Francis.

Patton, M.Q. (2002). *Qualitative research and evaluation methods*. California: Sage Publications.

Pielke, R. (2007). *The honest broker: Making sense of science in policy and politics*. UK: Cambridge University Press.

Powell, A., Piccoli, G., & Ives, B. (2004). Virtual teams: A review of current literature and directions for future research. *The DATA BASE for Advances in Information Systems, 35,* 6-37.

Presser, S. (1980). Collaborations and the quality of research. *Social Studies of Science, 10,* 95-101.

Prigogine, I. & Stengers, I. (1997). *The end of certainty: Time, chaos, and the new laws of nature.* New York: The Free Press.

Rasmussen, N. (2004). The moral economy of the drug company-medical scientist collaboration in interwar America. *Social Studies of Science, 34,* 161-85.

Russell, B. (1961). *History of Western Philosophy*. Boston: Allen & Unwin.

Salem, P. (2009). *The complexity of Human Communication.* New Jersey: Hampton Press.

Sandole, D.J.D. (2006). "Complexity and conflict resolution" in *Complexity in World Politics: Concepts and Methods of a New Paradigms* (N.E. Harrison ed.) New York: State University of New York Press.

Sangam, S.L. (2009), Research collaboration pattern in Indian contributions to chemical sciences. *COLLNET Journal of Scientometrics and Information Science, 3,* 39-45.

Sawyer, R.K. (2005). *Social emergence: Societies as complex systems.* New York: Cambridge University Press.

Schreyögg, G., Sydow, J., & Holtmann, P. (2011). How history matters in organizations: The case of path dependence. *Management & Organizational History, 1,* 81-100.

Scott, W.R. (1981). *Organizations: Rational, natural, and open systems (5th Ed.).* New Jersey: Prentice Hall.

Shrum, W., Genuth, J. & Chompalov, I. (2001). Trust, conflict, and performance in scientific collaborations. *Social Studies of Science, 31,* 681-730.

Shrum, W., Genuth, J. & Chompalov, I. (2007). *Structures of Scientific Collaboration*. Massachusetts: The MIT Press.

Sieber, J. E. (1991). Openness in the social sciences: Sharing data. *Ethics & Behavior, 1,* 69-86.

Simons, R. 1995. Control in an Age of Empowerment. *Harvard Business Review, 2,* 80-8.

Sims, B. (2005). Safe science: Material and social order in laboratory work. *Social Studies of Science, 35,* 333-66.

Sismondo, S. (2009). Ghost in the machine: Publication planning in the medical sciences. *Social Studies of Science, 39,* 171-98.

Smith, J. & Jenks, C. (2006). *Qualitative complexity: Ecology, cognitive processes and the re-emergence of structures in post-humanist social theory*. New York: Routledge.

Sooryamoorthy, R. (2009). Do types of collaboration change citation? Collaboration and citation patterns of South African science publications. *Scientometrics, 81,* 177-93.

Stacey, R.D. (2003). *Complex responsive processes in organizations: Learning and knowledge creation*. London, UK: Routledge.

Steelman, J.R. (1947). *Science and public policy.* The President's Scientific Research Board. Washington, D.C.: U.S. Government Printing Office.

Straus, S. G. 1996. Getting a clue: The effects of communication media and information distribution on participation and performance in computer-mediated and face-to-face groups. *Small Group Research, 27,* 115– 42

Stvilia, B., Twidale, M. B., Smith, L. C., & Gasser, L. (2008). Information quality work organization in Wikipedia. *Journal of the American Society for Information Science and Technology, 59*(6), 983-1001.

Subramanyam, K. (1983). Bibliometric studies of research collaboration: A review. *Journal of Information Science, 6,* 33-8.

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A.U., Wu, L., Read, E., Manoff, M., & Frame, M. (under review) (2011). Data Sharing by scientists: Practices and perceptions. *PLoS One*.

Thietart, R.A. & Forgues, B. (1995). Chaos Theory and Organization. *Organization Science, 6*, 19-31.

Timmermans, S. (2003). A black technician and blue babies. *Social Studies of Science, 33*, 197-229.

United Nations Development Programme (UNDP), (2008) Retrieved on January 2, 2009 from http://hdrstats.undp.org/countries/data_sheets/cty_ds_USA.html

Van Keuren, D.K. (2001). Cold War science in black and white: US intelligence gathering and its scientific cover at the Naval Research Laboratory, 1948-62. *Social Studies of Science, 31,* 207-29.

Vasileiadou, E. (2009). Stabilisation operationalised: Using time series analysis to understand the dynamics of research collaboration. *Journal of Informetrics,1*.

Vertesi, J. (2009). *The social life of spacecraft: Organization, negotiation and instrumentation on the Cassini and Mars Rover missions.* Paper presented at the Society for Social Studies of Science Annual Meeting in Washington, D.C.

Wagner, C. (2002). International linkages: Is collaboration creating a new dynamic for knowledge creation in science? A Research Proposal, University of Amsterdam, Amsterdam School of Communications Research (ASCoR).

Wagner, C. (2008). *The new invisible college: Science for development*. Washington, D.C.: Brookings Institution Press.

Waldrop, M.M. (1992). *Complexity: The emerging science at the edge of order and chaos*. New York: Simon & Schuster.

Webb, C., Lettice, F. & Lemon, M. (2006). Facilitating learning and innovation in organization using complexity science principles. *Emergence: Complexity and Organization, 1,* 30-41

Weinberg, A.M. (1961). Impact of large-scale science on the United States. *Science, 3473,* 161-4.

World Health Organization (2009). *World now at the start of 2009 influenza pandemic.* Retrieved on July 13, 2009 from http://www.who.int/mediacentre/news/statements/2009/h1n1_pandemic_phase6_20090611/en/

Yin, R.K. (1984). *Case study research: Design and methods (4th ed.)*. California: Sage Publications.

Appendix

Appendix A

Interview Guide

Themes	Questions
Emergence	How did you learn about DataONE? / How did you become involved in this group? Why did you become involve with DataONE? How did you find the researchers from other disciplines? Did you know any of the members in DataONE before? Did that acquaintance affect your decision in your involvement in DataONE? How? How it is decided for you to be in the Leadership Team? (added later) What was the recruitment process (added later)
Complexity & Interaction	What motivates you to work together? How do you communicate with other members in DataONE? How do others communicate with you? Have you encountered any problems or barriers regarding communication? (added later) How do you motivate people? (added later) What is your contribution to/role in DataONE? Have you developed new working relationships as a result of DataONE? How much time do you commit for DataONE weekly? Do you expect that to change?

	What does your institution/supervisor/boss thinks about DataONE and your involvement in it? What do you think of the PI and the management style of the project? (added later)
Adaptation	Has your role in the collaboration changed so far? How? Have you observed any changes in the functioning/organization of the collaboration? What are these changes? Why do you think these changes happened? What do you feel about DataONE? How did you decide to have working group structure? (added later)

Appendix B

Survey Instrument

Please tell us a little about yourself:

1. I was born in 19__
2. I am Female__ Male__
3. How long have you been a researcher? _____

Instructions: Thinking about your participation in DataONE, please select the relevant answer.

Section I - Demographics

4. Which one of the following best describes your primary subject discipline? (please select one only)
   - Biology
   - Computer Science
   - Ecology
   - Education
   - Environmental Science
   - Geology
   - Library and Information Science
   - Social Sciences
   - Other (please specify).

5. What is your primary working group in DataONE? (please select one only)
    - Citizen Science and Public Outreach
    - Community Engagement and Education
    - Data Integration and Semantics
    - Data Preservation, Metadata, and Interoperability
    - Distributed Storage
    - Federated Security
    - Scientific Exploration, Visualization, and Analysis
    - Scientific Workflows & Provenance
    - SocioCultural Issues
    - Sustainability and Governance
    - Usability & Assessment.

6. What are your secondary working groups (if any)? (select all that apply)
    - Citizen Science and Public Outreach
    - Community Engagement and Education
    - Data Integration and Semantics
    - Data Preservation, Metadata, and Interoperability
    - Distributed Storage
    - Federated Security
    - Scientific Exploration, Visualization, and Analysis
    - Scientific Workflows & Provenance
    - SocioCultural Issues
    - Sustainability and Governance
    - Usability & Assessment.

7. Why did you join DataONE? (open ended)

_____

_____

_____

_____

8. Are you a working group leader? Yes__ No__
9. How long have you been working in DataONE? ____year(s) ____month(s)

Section II. Related to your work in DataONE, how often do you communicate with

Person contacted is in the → Frequency ↓	More than once a day	Daily	Weekly	Bimonthly	Monthly
10. people in your primary working group?					
11. your primary working group leader?					
12. people in other working groups?					
13. other working group leaders?					
14. PI?					

Section II. Which communication channels do you use to communicate with

Person contacted → Comm. Channel ↓	Telephone	Face to face	Written documents (memos, letters)	Electronic media (email, text)	Virtual media (Skype/MSN, Marratech)	Plone website (shared webspace)	Wikis	Social networking sites
15. people in your primary working group?								
16. your primary working group leader?								

17. people in other working groups?								
18. other working group leaders?								
19. PI?								

Section III. Please rank order (1-most frequent to 4-least frequent) the kind of information you seek related to your role in the DataONE collaboration?

- Legal: Information related to the regulations of commitments of the institutions and the individuals.
- Financial: Information related to financial commitments.
- Scientific: Information related to the content of the information provided by the collaboration.
- Technical: Information related to the cyberinfrastructure components.

20. Information Type	Rank (1-most frequent to 4-least frequent)
Legal	
Financial	
Technical	
Scientific	

Section IV. Which channels do you use to seek information regarding DataONE?

Comm. Channel → Kind of information ↓	Telephone	Face to face	Written documents (memos, letters)	Electronic media (email, text)	Virtual media (Skype/ MSN, Marrtatech)	Plone website (shared webspace)	Wikis	Social networking sites
21. Legal								
22. Technical								
23. Scientific								
24. Other (please specify)								

Appendix C

Coding scheme of interviews for the complexity framework

In this section some quotes from the interviews that are used to support the complex adaptive systems framework are provided.

1. Large number of components

- Not applicable (NA)

2. Variation, diversity, and multidisciplinary structure

- So we had early discussions…this has been over two years ago, but early on into the project it was clear that NSF expected a significant involvement of what we could loosely call library science community in the DataNet partners. At the time when I got involved in this we really did not have a strong library science partner in the organization, in the proposal team.
- I guess the evidence for that is publications, you see that there is a lot of interdisciplinary publishing collaborations, when you look across the people who work in Information Sciences, they definitely reach out to other disciplines but even within their discipline, within the school itself you see that a lot of the colleagues are collaborating on projects and so for me that is a way of building up the school and making sure that there is a momentum there and then I also see where people write grants together and bring in money together so that the research can continue, and people taking this piece and that piece of it and somehow making projects out of it, so those are the sorts of things I was looking at.
- Yeah, they gave me the opportunity to organize this working group on scientific exploration, visualization and analysis. That allowed us to pull together computer scientists from, who are focused on machine learning and higher performance computing and work scientific workflow software and combine them with people who are experts in informatics, data organization and then bring in statistical analysts and quantitative ecologists. So, we have a very diverse group there.
- Well, that is something that we are consciously aware of in the team, and so we always want to get the best person but we also want to make sure we don't exclude anybody, because you know, guys know guys and gals know gals or something like that so there is attention paid to it but it doesn't shape everything we do but it is always a consideration as we are making certain kinds of decisions
- There are a couple things there. One is the, uh, one thing to think about is the disciplines – scientific disciplines. I think geology versus oceanography versus geochemistry. That is one thing. I appreciate that there are difference out there and we can look at what those folks are

doing and see if there are any areas of overlap. From a cyberinfrastructure standpoint, I think there is a lot of research going on out there that is really interesting. What we are trying to do is see what makes sense for us to embrace some of these new cyberinfrastructure developments: who to team up with. Is there some integration of data or is there some visualization tools that are really useful that we have no idea about that we would like to embrace. That is another aspect. The other thing is the idea of the library community. I never really realized what they are up to, to be honest with you. I never knew. Now I am working with you guys.

- I really enjoy people coming at a problem from a variety of backgrounds. I think you get very creative solutions and it is really exciting to be involved in a group like that. There is no one solution. so it is actually an opportunity to try different techniques for getting people to help you meet deadlines

3. Connectivity, interdependence, and interaction

- The criteria was that we wanted to have people who were good communicators who would listen and would really not want to do their own thing so that was a criteria. People who were difficult to work with or whatever, we tried to avoid that. So we built a team of people based on that. It was a real fundamental principal to start.

4. Feedback

- NA

5. Unpredictability and nonlinearity

- NA

6. Far-from equilibrium/edge of chaos

- NA

7. Emergence / Self-organization / Strange attractors

- Yeah, we have a chapter that we wrote in a book that is coming out on data intensive science that focuses on the experience we had with the EVA working group. And we had a brief write-up about us and what we were doing in the working group in August, this past August, in Nature magazine
- So far I have done research based on DataONE but not with DataONE members, I mean we have had a lot of presentations that people have done and everybody has named everybody on those. One person might be presenting but somewhere on there it will say here is the whole

team. So in that way my name showed up with a lot of these folks from things like that but I haven't had a small research project with any one of them. However, one of the grants with science links, that is in collaboration with the people who are with DataONE, they will be providing scientific mentoring that goes to those doctoral students. That is the beginning of it.

- The posters, some of them, if you look at the one out in the hallway has the whole team on there. So I don't even keep a track of them, on my CV, I don't put all of those posters, I probably should but I don't. We have done posters, we have done talks with the new but everybody is new. And in addition to DataONE, there were two DataNets that were funded at the same time, the Data Conservancy is the other one, so there is relationship with people on Data Conservancy as well, also C at Illinois who is on Data Conservancy and does a lot of assessment on that, she and I end up doing talks on panels and stuff, so that is a new relationship.
- The other thing with it is that we planned to build a research agenda around it so it is a huge transformative kind of project for us. It will be 5 years but we have already gotten an IMLS grant based on it at their, IMLS's request, the want us to build on DataONE, and we have NSF grant, one proposal is ready to come out, we will be generating a lot of NSF, IMLS, other proposals that build on DataONE so it is really going to be transformative for the school, we had the science specialty but we really need to expand on that. So this gives us people, we have got A is an adjunct, we have got B who is an adjunct, it gives us students, the science links, it gives us connections and it gives us cache, it will help us build that strength in the college as well as in the school so it is pretty significant for us, for some of the other participants in it maybe not quite but for us it is quite significant.

8. Adaptation and learning

- And as the organization has grown and matured, we can do a little bit more in the way of clarification and specialization. So that is probable the way things have evolved more than anything else. We have gotten better sense of what other people can do and how we best fit with each other.
- The structure changed a little bit originally, we didn't have any second of director written in and that has changed because as the organization morphed and as we got more structured from the founding agency we realized that is something that we needed to add and we changed, we had called him assistant director, we changed that into directors for the two groups, infrastructure and community engagement and over time changed what the responsibilities for those would be a little bit –actually quite a bit,
- I do know that the organization is a little bit more nimble than I thought at first, you probable hear at the meeting talking about the EVA working group, visualization and analysis, that was a working group that wasn't envisioned or kind of put into the plan at the beginning but the leadership team saw the need for it and really forged ahead and created that
- I think it is just an organization that can change as they see the need for change. They started out with a structure, and I don't think the structure has changed a lot but how the people can

move around in that structure it seems fairly easy to do that. In other words if Dr. Michener needs somebody from another working group to help him with something, with a task, you know it is kind of like all he needs to do is ask and It happens but it is not just because he is the big boss though, I think the same is true for everybody, I think a part of it is that, the structure, although it has been set forward, people are still finding their places within that structure and they are finding the places where they can be the most helpful and useful and my impression of the leadership team and the way the organization works, and Dr. Michener's leadership style is that he wants to let people find their own place and be helpful where they think they can so it is kind of a very welcoming environment that way, that makes it nimble, things are able to change as they need to change.

- So my role in it has become bigger than I expected it, I never thought that I would have time to commit to be on a leadership team for example.
- When I saw that this is something I want to be involved in, not just facilitate, originally I wasn't involved but then I said OK, I won't be a CO-I or CO-PI but I will be a working group lead, so as a working group lead, that was my first official role. And helping to get the proposal written and I agreed to be the working group lead and then as it was going on, as I said it was more interesting and the project was evolving and I got to know the people more so I have taken a little more of a lead. I have switched over to be a working group lead with X for the usability and assessment working group. Originally I was not that working group lead but as we developed and as we had our site visit from NSF we realized that that is crucial and that we have to have people who are willing to throw themselves into it and do that so that is when I moved over to be the co-lead of that
- I think there is much more flexibility than was perhaps originally envisioned in terms of having the opportunity for short-term working groups or workshops, having these super groups with working groups collaborating together on projects and also having sub-groups. So, that concept, I do not think was there initially, but has evolved through the working styles of our individuals within the working groups.
- A working group was added that wasn't in the proposal because it became clear that it needed to be added. That is the EVA visualization working group because we needed something that could be shown, we needed something that people could visualize, data isn't sexy but what we can do with data, we needed to get that sexy part, we needed to get the part that looks good, that shows the bird migration and all these lovely things that the visualization team is doing. So that was a piece that we originally did not anticipate and they are already doing great things. The leaders of that are the people who are interested in citizen science and community. So they are scientists but they are also interested in the interface between the science and the public and so that has been added structurally.
- So we are really at the very early stages. So maybe some of the working groups will decide they need to split or merge, some of that might happen, we just now have the executive director in place, he just started. The technical associate director has been in place for a longer time, he has been involved from the beginning, the proposal stage. And the community engagement and education associate director's job is about to go out. So there will be some changes,

- It has, yeah. So, I have always been a member of the leadership team, but, when I first started the process, I co-lead with X the entire community engagement side of the DataONE project and so that lasted for like a little over a year or something like that until Y was hired to take over that position. At that time, Z and I had always been named the community engagement working group leads but we were able to then free up our time to actually concentrate on that working group at that time. It was good to release that responsibility to Amber.

9. Historicity and path-dependence

- The Oak Ridge DAAC has had a long history of collaboration with the Long Term Ecological Research Network (LTER) which has historically been involved with the National Center for Ecological Analysis and Synthesis (NCEAS). So the proposal really started with that group of people as they looked for data centers and groups they want to partner with. They talked with B who is well known among the community for his work on best practices, data preservation and I was brought in with somebody with cyberinfrastructure expertise as well as operational experience in industry.
- I am sure that X and Y and Z has some idea of what we might look like to some degree because they have been working on another proposal called the interopt grant that was a virtual data center which was a part of interopt grant, the proposals. So that was sort of the core little baby acorn version of DataONE so they kind of had something that we can start to think about but when you put the group together things morphed and changed and it kind of grew from there. It was really cool
- Basically, the SEEK project was looking at essentially the problem of data integration to biological sciences and ecological sciences. And, it had fairly ambitious goals, and there was also a lot of research work involved in that project

10. Coevolution

- NA

Appendix D

Coding scheme of interviews for other themes

Motivation to be involved in DataONE

- It was an opportunity to broaden the work that my group does to the National Science Foundation. I have strong opinions about cross agency efforts minimizing duplication of efforts. I get very frustrated when I see something where USGS pays for the same thing that the NSF pays for the same thing that the Department of Energy doing that is identical to what NASA does. Just looking across some of those –even within NSF –I see or different groups do the fundamentally same thing and it is not in their perceived best interests to collaborate. And I find that waste of resources that in the context of the things like climate change and ecology can't effort.
- Personal side of things –it was a change to broad my collaborations. Yet fairly quickly became a means for me to get a more effective tenfold into the University of X. One of the reasons I left industry was I wanted greater involvement with students and have some potential to be in an academic environment but I did not and in fact do not really have the fully credentials to seek tenured position in a university. This gave me a chance get involved with the department and I have already instructed with X, Y and with some other people I helped to teach a class over the course of the summer. This is a way I can get more involved with the university and the students.
- Yeah, primarily. I mean, you know, I have had a long interest in informatics and data interoperability and data organization. We have won several fairly significant National Science Foundation awards for the lab, based on those kinds of things. And we are really excited about supporting DataONE so it can provide a platform for us to do our work.
- It is a very unusual answer, I suspect. On a very personal level and in the very sacred sense of that word, I feel called to do what I do. I have been given a set of gifts by God, and he calls me to use those in a way which is of service to this world. I have a significant concern about a number of aspects of the abuse of climate and ecology and this is a way that I can contribute and be a part of a greater good that is well beyond anything that I could individually do
- You know I am doing all this work, my whole career is looking at scientific communication and scientists and publishing, I really ought to be involved in a project that is working directly with scientists'.
- Um, I thought it was the wave of the future. We have our XY data center but it is really isolated. We realize there are a lot of other activities going on out there and we need ways to link our holdings with other holdings and we need to take advantage of other practices, what folks are doing, citations, and tools and services. We just cannot do it all in isolation. So, we have some skills that we would like to share with others and we want to see what others are doing and see if we can incorporate those practices without having to reinvent the wheel

Bridging Role

- I also see myself as one of the bridges between the cyberinfrastructure and community engagement sides of things because I lean that way. I have done work on public relations, I have done work on developing communication plans, I think people like myself and Q are key bridges across the cyberinfrastructure and community engagement sides of the project.
- I have also had more media training than most other ecologist and probably most other scientist because I am at the head of a research center. So, the media training actually, you know, helps you to, just communicate across disciplines as well as how to communicate to the public. I would credit that to some of the success in communication.
- For a successful project and in a successful organization one needs a combination of people who are deep technical experts within their given area and one also needs somebody who can speak the language of a broad range of experts. That has historically been one of the roles that I have had in projects because that is something that appeals to my personality and matches up with some of my skills. So I can talk to X and Y about the issues in development of the communications plans and the communication strategy, understand some of the issues in educational approaches in engagement, and then turn around and talk to the developers about details of communication protocols and so forth. And I don't understand any one of them to the level that those particular experts do, but I understand enough of what they do that I can translate the language from one to another.

LIS Involvement

- There was a big call from NSF and he was with a group who was thinking of submitting a proposal and they needed a library component and being next to ORNL they thought of UT as providing the library component, the IS component.
- I think, one of the things I think is so exciting about it is the opportunity to work with people in a library background. This is the first time that I have done that. What I see for DataONE is such a wonderful opportunity for libraries in the future. I think we are moving away from books and libraries are going to need a new mission. I think being responsible for data is an excellent mission and they really have a fantastic background for this.

Intimidation

- Well, I was definitely very quiet the first day because I was just taking everything in, learning everybody's names, figuring out who did what because some of the people came totally from the hard sciences, other people came from the mix of hard sciences and computer science. I was the only Librarian in the room, Information Science person, so part of it was there in terms of, gosh; it is a really tough question. It was scary to be the only person, everybody else knew at least one or two people so that was kind of scary because you are the only person but

they really, they made un-scary it very quick, it was like, we really value everybody's opinion. The way the meeting was run, everybody had a chance to say something in terms of the way…it is hard to explain…but it was like, you were not allowed just to sit there and say nothing, I think you experienced that.

- I will just be truthful and say that I thought the group of people that had been gathered together to work on this was overall such high quality people, you know, they have all achieved so much, that it was mildly intimidating, to be involved with that crowd at the very beginning. At the very beginning, I was quite quiet, I guess. Once I started to get to know people in that group, I felt a lot more comfortable. By, I would say, the third meeting or something like that, I suddenly got to know the true them. It was interesting at the very beginning of this experience
- To tell you the truth at first I thought that well I am not sure how helpful I would be on this grant but I certainly welcome the opportunity to learn and to try to be helpful so that's kind of it happened. [Researcher: Why did you think that?] Just because I didn't have any direct experience with data or with the kind of thing you are writing about, scientific collaboration. I am certainly not a science librarian and I really don't know too much about the work that science librarians do. I am also not really involved with the whole digital preservation of the library work which is where I think libraries really fit into DataONE, so I realized early on that I would have a lot to learn but I don't think that is a bad thing, so as long as X and Y had confidence that I could learn what I had to learn I said OK, let's try it.

Communication Problems

- There is a huge split between the community engagement and technical working groups. Technical working groups want everything on the official sites, the ticket system, the plone site, etc. And the community engagement likes to do everything via e-mail. Huge, huge split. We all agreed at the meeting here that we would use the site for the documents and the tickets but the rest of it, it is just too complicated and we are not likely to use it, forget it. But the technical working groups are going to continue to use it. We had a little bit of a rebellion there in a sense that the technical working groups have a site that requires downloading a special browser, it is arcane to a lot of us but for them it is just the way they work, it wasn't a part of the way we did our work. And they said 'you have got to use it, quit sending e-mail, don't do e-mail' and so finally we said 'we will meet half way. We will use some of the official system, we'll use tickets, we will use plone but we will not use the other.' But they are still using it. There was a definite difference and it is cultural in terms of subject disciplines, in terms of what they are used to. It has been interesting, and when I say technical I don't mean 'science technical'. I mean 'computer technical' folks who are used to do – because if you are writing documentation, doing collaborative programming, you have got to have versioning system, you have got to make sure that everybody can get to version 1, version 1.1, version 1.2, it is a , and it is very important to have it in one place all that but the stuff that we are doing is not like that, so we didn't need that so much or we are just not used to it and we didn't want to learn anything new, that might be part of it too.

- Yeah, I think I have a hard time keeping up with the jargon and CI. But, the people on the CI side of the project are very sensitive to that and they truly want to communicate, and can communicate, so it is very easy with these particular individuals to say, " I don't understand," and they can explain it (laughing) in words that I understand. So, there are definitely obstacles and it slows down our conversation but I think that happens anytime you talk across such wide gaps and disciplines. I think we overcome it exceptionally well.
- It can be very challenging finding the common dialog of cross disciplines. That obviously sets up a communication challenge, which DataONE has been very attentive to and has worked to address. I do not think it results in a significant challenge for DataONE but it is certainly one that exists
- Yeah, so, um, I would say that the community engagement side of things we use plone quite a bit as does the executive director and the PI, so the executive team, are quite plone-based. In addition, some of the working groups, sociocultural group have set up their own wiki sites so they can share material and so I make an effort to check that. That is a little bit more challenging to see what has occurred frequently. You know, what has been added that is more recent. You cannot look for changes quite so easily. In terms of the CI, I know that they are using subversion for their main document sharing. That is one of the challenges the organization faces. On one hand, I think it would be valuable for everyone to share the same time of interface or system for sharing material. But, I know that CI do enjoy using subversion and CE do enjoy using plone. So, the potential for one group of other to use the other system might be a hurdle to overcome. So, it is whether that hurdle encourages people, on either side, encouraging people to use something that is not their first response or first nature –may be more of a challenge than just having two systems operational. I think that is something we need to think about as more and more material is produced in moving forward. Because, one of the challenges of having two systems, or repositories of documents we share, is duplication and also non-conformity between the two. Um, so there might be some things that CI has put into plone to share with CE and they have been updated on Subversion but have not been updated on plone for example. So, that is something we really need to be mindful of and to find some sort of resolution for if we are going to maintain two different systems. I
- Oh absolutely. Absolutely. Especially between the cyberinfrastructure side of the house and the community engagement side. And, I think you end up with a solution. It can take longer to get to a solution but I think it is really interesting. For example, the cyberinfrastructure guys, when deciding well what are we going to use for a document repository, they immediately decided, of course, we will use SVN or Subversion, which is a place that if you are a developer you deposit in code because it will version it for you automatically; you can recall back to previous things. What they did not take into account is you have to be a real geek to enjoy that and to be able to actually do it. And, the community engagement side went, 'I don't think so, we are not going to do this'. They gave it a try. They did try but there is absolutely no reason to have to go through all those steps when we could set up a document repository where you can write and drop. So, I thought it was real interesting that the community engagement people

actually learned that there are things for storing code and what they do and why you use them and the cyberinfrastructure guys figured out that maybe not everybody likes the tools they use on a daily basis. So we have now figured out, we use the drag and drop place and the cyberinfrastructure.

Appendix E

Tables

Gender

I am

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	female	23	45.1	46.9	46.9
	male	26	51.0	53.1	100.0
	Total	49	96.1	100.0	
Missing	System	2	3.9		
Total		51	100.0		

Career Age

How long have you been a researcher?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	2	3.9	4.4	4.4
	3	1	2.0	2.2	6.7
	4	1	2.0	2.2	8.9
	5	2	3.9	4.4	13.3
	6	4	7.8	8.9	22.2
	7	1	2.0	2.2	24.4
	8	1	2.0	2.2	26.7
	9	2	3.9	4.4	31.1
	10	4	7.8	8.9	40.0
	12	3	5.9	6.7	46.7
	13	1	2.0	2.2	48.9
	15	3	5.9	6.7	55.6
	16	1	2.0	2.2	57.8
	17	1	2.0	2.2	60.0
	19	1	2.0	2.2	62.2

	20	4	7.8	8.9	71.1
	24	2	3.9	4.4	75.6
	25	3	5.9	6.7	82.2
	30	4	7.8	8.9	91.1
	31	1	2.0	2.2	93.3
	35	2	3.9	4.4	97.8
	40	1	2.0	2.2	100.0
	Total	45	88.2	100.0	
Missing	System	6	11.8		
Total		51	100.0		

Primary subject discipline of DataONE members

Which one of the following best describes your primary subject discipline? (please select one only)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	4	7.8	8.9	8.9
	biology	3	5.9	6.7	15.6
	computer science	7	13.7	15.6	31.1
	ecology	7	13.7	15.6	46.7
	education	2	3.9	4.4	51.1
	environmental science	4	7.8	8.9	60.0
	library & info science	16	31.4	35.6	95.6
	social sciences	2	3.9	4.4	100.0
	Total	45	88.2	100.0	
Missing	System	6	11.8		
Total		51	100.0		

Primary working group frequency table

Working Group	**Frequency**	**Percent**
Citizen Science and Public Outreach	1	2.13

Community Engagement and Education	4	8.51
Data Integration and Semantics	4	8.51
Data Preservation, Metadata, and Interoperability	2	4.26
Distributed Storage	2	4.26
Federated Security	1	2.13
Scientific Exploration, Visualization, and Analysis	3	6.38
Scientific Workflows & Provenance	2	4.26
SocioCultural Issues	11	23.40
Sustainability and Governance	5	10.64
Usability & Assessment	12	25.53
Total	47	100

Secondary working group membership frequency table

Working Group	Frequency	Percent
Citizen Science and Public Outreach	6	13.04
Community Engagement and Education	3	6.52
Data Integration and Semantics	2	4.35
Data Preservation, Metadata, and Interoperability	9	19.57
Distributed Storage	0	0.00
Federated Security	1	2.17
Scientific Exploration, Visualization, and Analysis	4	8.70
Scientific Workflows & Provenance	2	4.35
SocioCultural Issues	8	17.39
Sustainability and Governance	5	10.87
Usability & Assessment	6	13.04
Total	46	100

Involvement in DataONE

How long have you been involved in DataONE?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	2	1	2.0	33.3	33.3

	3	1	2.0	33.3	66.7
	40299	1	2.0	33.3	100.0
	Total	3	5.9	100.0	
Missing	System	48	94.1		
Total		51	100.0		

Frequency of communication

Related to your work in DataONE, how often do you communicate with your working group members?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	more than once a day	2	3.9	4.8	4.8
	daily	2	3.9	4.8	9.5
	weekly	6	11.8	14.3	23.8
	bimonthly	7	13.7	16.7	40.5
	monthly	11	21.6	26.2	66.7
	less than monthly	12	23.5	28.6	95.2
	don't communicate	2	3.9	4.8	100.0
	Total	42	82.4	100.0	
Missing	System	9	17.6		
Total		51	100.0		

Related to your work in DataONE, how often do you communicate with your primary_WG_leader?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	daily	4	7.8	10.0	10.0
	weekly	12	23.5	30.0	40.0
	bimonthly	4	7.8	10.0	50.0
	monthly	8	15.7	20.0	70.0
	less than monthly	11	21.6	27.5	97.5

	don't communicate	1	2.0	2.5	100.0
	Total	40	78.4	100.0	
Missing	System	11	21.6		
Total		51	100.0		

Related to your work in DataONE, how often do you communicate with other_WG_members?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	weekly	8	15.7	18.6	18.6
	bimonthly	4	7.8	9.3	27.9
	monthly	9	17.6	20.9	48.8
	less than monthly	14	27.5	32.6	81.4
	don't communicate	8	15.7	18.6	100.0
	Total	43	84.3	100.0	
Missing	System	8	15.7		
Total		51	100.0		

Related to your work in DataONE, how often do you communicate with other_WG_leaders?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	more than once a day	2	3.9	4.7	4.7
	daily	1	2.0	2.3	7.0
	weekly	5	9.8	11.6	18.6
	bimonthly	2	3.9	4.7	23.3
	monthly	6	11.8	14.0	37.2

	less than monthly	12	23.5	27.9	65.1
	don't communicate	15	29.4	34.9	100.0
	Total	43	84.3	100.0	
Missing	System	8	15.7		
Total		51	100.0		

Related to your work in DataONE, how often do you communicate with the PI?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	more than once a day	1	2.0	2.4	2.4
	daily	3	5.9	7.1	9.5
	weekly	4	7.8	9.5	19.0
	monthly	4	7.8	9.5	28.6
	less than monthly	13	25.5	31.0	59.5
	don't communicate	17	33.3	40.5	100.0
	Total	42	82.4	100.0	
Missing	System	9	17.6		
Total		51	100.0		

Communication channels used

Which communication channels do you use to communicate with?

	phone	**f2f**	**written**	**email**	**virtual**	**plone**	**wiki**	**social**
Primary WG member	12	25	7	37	13	14	12	1
Primary WG leader	10	22	4	36	15	15	8	0
Other WG member	4	17	3	26	10	9	6	2
Other WG leader	5	16	1	20	7	7	4	1

PI	3	13	5	21	8	7	1	0

Vita

Arsev Umur Aydinoglu's research interests are scientific collaborations, complexity theory, emergence, scientists' data practices, science communication, and interdisciplinary studies.

In his dissertation, "Complex Adaptive Systems Theory Applied to Virtual Scientific Collaborations: The Case of DataONE", he investigates the emergence of DataONE (a multidisciplinary, multi-institutional, and multinational virtual data network on earth sciences) and the role of information and communication behaviors in it from a complexity theory perspective using multiple methodologies.

He presented papers at international, national, and regional conferences such as BOBCATSSS, ALISE, Society for Social Studies of Science (4S) and UT's Research Symposium (awarded best research paper in 2010) and published one article in Emergence: Complexity & Organization (please see my CV or click 'publications' link for details).

Arsev Umur Aydinoglu had worked at the World Bank Turkey Office as Public Information Assistant for 4.5 years. He was responsible for the World Bank library and country website and assisted youth and NGO projects funded by the Bank in Turkey.